3.ª edición

# BIOESTADÍSTICA AMIGABLE



Miguel Ángel Martínez González
Almudena Sánchez-Villegas
Estefanía A. Toledo Atucha
Javier Faulín Fajardo

ELSEVIER

CONTENIDO ADICIONAL en línea

# BIOESTADÍSTICA AMIGABLE

# BIOESTADÍSTICA AMIGABLE

## 3.ª edición


**EDITORES**

*Director*
**Miguel Ángel Martínez-González**
Catedrático de Medicina Preventiva y Salud Pública,
Universidad de Navarra

*Codirectores*
**Almudena Sánchez-Villegas**
Profesora Titular de Medicina Preventiva y Salud Pública,
Universidad de Las Palmas de Gran Canaria

**Estefanía Toledo Atucha**
Profesora Contratada Doctora de Medicina Preventiva y
Salud Pública, Universidad de Navarra

**Javier Faulin Fajardo**
Catedrático de Estadística e Investigación Operativa,
Universidad Pública de Navarra





Ámsterdam Barcelona Beijing Boston Filadelfia Londres Madrid
México Milán Múnich Orlando París Roma Sídney Tokio Toronto

ELSEVIER

ELSEVIER


Travessera de Gràcia, 17-21
08021 Barcelona, España



ISBN (versión impresa): 978-84-9022-500-4
ISBN (versión electrónica): 978-84-9022-651-3

Depósito legal (versión impresa): B. 12.965 - 2014
Depósito legal (versión electrónica): B. 12.966 - 2014

Coordinación y producción editorial: *GEA* CONSULTORÍA EDITORIAL, S. L.

**Advertencia**

La medicina es un área en constante evolución. Aunque deben seguirse unas precauciones de seguridad estándar, a medida que aumenten nuestros conocimientos gracias a la investigación básica y clínica habrá que introducir cambios en los tratamientos y en los fármacos. En consecuencia, se recomienda a los lectores que analicen los últimos datos aportados por los fabricantes sobre cada fármaco para comprobar las dosis recomendadas, la vía y duración de la administración y las contraindicaciones. Es responsabilidad ineludible del médico determinar las dosis y el tratamiento más indicados para cada paciente, en función de su experiencia y del conocimiento de cada caso concreto. Ni los editores ni los directores asumen responsabilidad alguna por los daños que pudieran generarse a personas o propiedades como consecuencia del contenido de esta obra.

**El Editor**

# AUTORES

**Francisco Javier Basterra-Gortari**
Profesor Asociado de Medicina Preventiva y Salud Pública, Universidad de Navarra.
Especialista MIR en Endocrinología.

**Maira Bes-Rastrollo**
Profesora Titular de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Jokin de Irala Estévez**
Catedrático de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Pedro A. de la Rosa Fernández-Pacheco**
Residente de Medicina Preventiva y Salud Pública, Servicio Navarro de Salud-Osasunbidea.

**Martín García-López**
Especialista MIR en Cardiología, Clínica Universidad de Navarra.

**Alfredo Gea Sánchez**
Profesor Ayudante Doctor de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Francisco Guillén-Grima**
Catedrático de Medicina Preventiva y Salud Pública, Universidad Pública de Navarra.

**Cristina López del Burgo**
Profesora Contratada Doctora de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Jesús López-Fidalgo**
Catedrático de Estadística e Investigación Operativa, Universidad de Castilla-La Mancha.

**Nerea Martín-Calvo**
Becaria Río Hortega. Medicina Preventiva y Salud Pública, Universidad de Navarra.
Doctora en Medicina. Especialista MIR en Pediatría.

**Jorge María Núñez-Córdoba**
Especialista en Medicina Preventiva y Salud Pública, Doctor en Medicina.
Clínica Universidad de Navarra.

**Miguel Ruiz-Canela**
Profesor Titular de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Carmen Sayón-Orea**
Profesora Ayudante Doctora de Medicina Preventiva y Salud Pública, Universidad de Navarra.

**Juan B. Toledo Atucha**
Investigador posdoctoral en el Centro de Investigación de Enfermedades Neurodegenerativas, Facultad de Medicina, Universidad de Pensilvania, Filadelfia, EE. UU.
Especialista MIR en Neurología.

# PRESENTACIÓN DE LA TERCERA EDICIÓN

*Bioestadística amigable* es un texto que ha ido ganando popularidad durante la última década. Quizás esto se deba a su orientación eminentemente práctica, con un contenido matemático mínimo. Son muchas las facultades que lo han recomendado para el aprendizaje de la Bioestadística. Esto es un motivo de alegría y de gratitud hacia quienes han confiado en nosotros. También esta grata experiencia, junto con otras razones, nos ha supuesto un buen incentivo para acometer una nueva edición, con la idea de poder mejorar el servicio a nuestros lectores. Hemos tenido la fortuna de contar con la editorial Elsevier para esta nueva edición.

En las dos ediciones previas, realizadas magníficamente por Díaz de Santos, la intención del libro fue doble: por una parte, se buscaba enseñar al lector a identificar los procedimientos estadísticos que están indicados para afrontar cada diseño de investigación, y, por otra, se orientó el texto hacia el aprendizaje de la interpretación clínica o práctica de los resultados de un programa informático estándar. Siempre hemos procurado alcanzar una cobertura suficientemente amplia y poder abarcar todos los procedimientos estadísticos que están realmente presentes de modo habitual en la investigación publicada en revistas biomédicas.

Las dos ediciones anteriores, de 2001 y 2006, estaban muy centradas en el programa SPSS. Esto se debía a que SPSS era el *software* más utilizado tanto en hospitales como en investigaciones de laboratorio en el mundo de habla hispana. No obstante, también se incluyeron en la segunda edición aplicaciones con otros programas (Excel, STATA, SAS y Splus o su versión gratuita, R). Pero ya hace tiempo que el *software* STATA se está demostrando superior en muchos aspectos a SPSS, sin dejar de ser igualmente *amigable*. Además, una licencia de STATA es siempre más accesible desde el punto de vista económico que una licencia de SPSS.

Por otra parte, especialmente en el ámbito de la epidemiología, casi se ha abandonado ya SPSS y se usa preferentemente STATA (en España), o bien SAS (en EE. UU. o Europa). Pero SAS es mucho más caro que SPSS y, por supuesto, que STATA. Otro programa emergente y muy completo es R, que tiene la gran ventaja de que es gratuito. Pero R es menos amigable, y está más pensado para profesionales que ya tienen buenos conocimientos de estadística matemática y programación. Todo esto hace que la balanza se incline por dar preferencia a STATA. No obstante la preferencia por STATA, hemos procurado presentar siempre en este libro el modo de resolver cada procedimiento también con SPSS y con R. También se presentan posibles aplicaciones realizadas con Excel, porque pueden ser muy didácticas. En definitiva, no es imprescindible contar con STATA para que el texto cumpla su utilidad.

Pero nuestra experiencia nos dice que STATA es ideal para quien empieza desde cero. Por eso hemos dado preferencia a STATA en esta nueva edición, que escribimos con la pretensión de que sea todavía más *amigable*. También hemos puesto los medios para que esta edición sea más breve que la anterior en aras de acercarla aún más al lector interesado en la aplicación práctica de los resultados estadísticos en las ciencias de la vida.

**Miguel Á. Martínez-González**
*Houston, Texas, Junio 2013*

# ÍNDICE DE CAPÍTULOS

# INTRODUCCIÓN A LOS MÉTODOS DE LA EPIDEMIOLOGÍA Y LA BIOESTADÍSTICA

1

*M. Á. Martínez-González, A. Sánchez-Villegas, J. de Irala*

## 1.1. ESTADÍSTICA, ESTADÍSTICA APLICADA Y BIOESTADÍSTICA

Se suele hablar de «literatura» biomédica, aunque quizá podría dejar de aplicarse ya el término de *literatura* a la bibliografía biomédica. Se constata una realidad: han prevalecido las exigencias del rigor cuantitativo por encima de las pretensiones estéticas propiamente «literarias». Escasean las descripciones meramente cualitativas de un solo caso clínico o del aspecto de un cultivo. Todo en ciencia acaba traducido a una información cuantificable, que se describe y compara mediante medias, porcentajes, histogramas, etc. A esta creciente aparición explícita de conceptos cuantitativos hay que añadirle su incorporación *implícita* en la toma de decisiones. Efectivamente, al interpretar los resultados de un experimento, al aceptar o descartar hipótesis o al realizar juicios etiológicos, diagnósticos o pronósticos, en la práctica clínica se aplican los principios de la probabilidad y de la estadística.

Por todo esto, el siglo XXI será el siglo de la estadística en las ciencias de la vida y el siglo de la medicina basada en pruebas *(evidence-based medicine)*. Su aplicación requiere adquirir soltura y buenas competencias en el manejo de conceptos cuantitativos (**1**). Esto tiene mucho que ver con la bioestadística que se explica en este manual.

La estadística consiste en la recogida, cuantificación, síntesis, análisis e interpretación de la información relevante contenida en unos datos. Puede dividirse en dos grandes campos: estadística matemática y estadística aplicada. La primera es el terreno de los profesionales de las ciencias exactas y puede resultar inaccesible a especialistas en otras áreas.

La *estadística matemática* supone una gran fuerza creativa, ya que desarrolla nuevos procedimientos que se utilizan para resolver problemas en los distintos campos del saber. Requiere un detallado conocimiento de los principios matemáticos y exige el nivel de abstracción y generalización propio de las ciencias matemáticas.

La *estadística aplicada* versa, precisamente, sobre cómo y cuándo utilizar cada procedimiento y cómo interpretar los resultados obtenidos. Estudia, por tanto, la transferencia de los métodos de la estadística matemática a otras disciplinas, como la economía, la publicidad, la sociología o la medicina (**2**).

La *bioestadística* es la rama de la estadística que se ocupa de los problemas planteados dentro de las ciencias de la vida, como la biología o la medicina, entre otras (**3,4**). Médicos, biólogos, enfermeras, nutricionistas o especialistas en salud pública necesitan conocer los principios que guían la aplicación de los métodos estadísticos a los temas propios de cada una de sus respectivas áreas de conocimiento.

## 1.2. BIOESTADÍSTICA DESCRIPTIVA Y BIOESTADÍSTICA ANALÍTICA O INFERENCIAL

La bioestadística se divide en dos grandes apartados: bioestadística descriptiva y bioestadística analítica o inferencial. La bioestadística *descriptiva* simplemente pretende sintetizar y resumir la información contenida en unos datos. Sus misiones son recoger, clasificar, representar y resumir datos. La bioestadística *analítica o inferencial* va más allá, pues pretende demostrar asociaciones o relaciones entre las características observadas. Su misión es hacer inferencias o extraer consecuencias

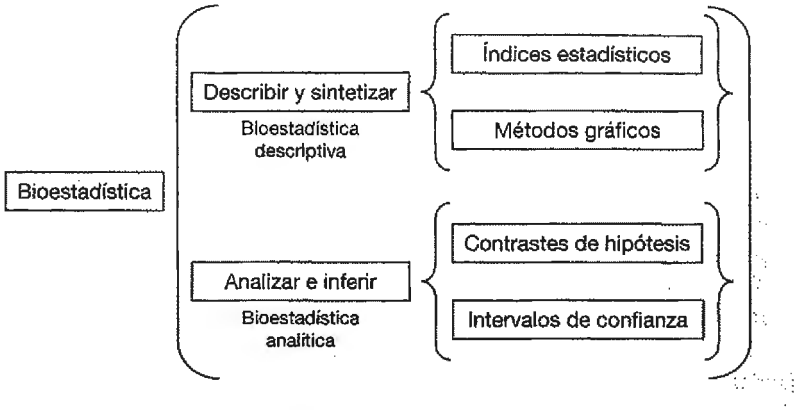



**Figura 1.1** Aspectos estudiados por la bioestadística.

científicas a partir de unos datos. La presencia de estas asociaciones puestas de manifiesto por técnicas estadísticas analíticas servirá de base para contrastar las hipótesis de una investigación frente a los datos recogidos empíricamente.

La bioestadística descriptiva utiliza dos tipos de procedimientos: el cálculo de índices estadísticos, que son números que resumen de modo sencillo la información contenida en datos reales, y el uso de representaciones gráficas. La bioestadística analítica o inferencial usa también dos tipos de procedimientos: la comprobación de hipótesis («contraste de hipótesis») y la estimación de intervalos de confianza. El contraste de hipótesis confronta los resultados encontrados en los datos recogidos con una hipótesis teórica de la que se partía. Se acaba tomando una decisión sobre si los datos apoyan o no esa hipótesis de partida. Un intervalo de confianza es apostar por dar un rango de valores creíbles para un parámetro desconocido. Esta credibilidad se mide en términos probabilísticos.

En la figura 1.1 se resumen los distintos apartados que comprende la bioestadística.

## 1.3. POBLACIONES Y MUESTRAS

La estadística habitualmente estudia solo una *muestra* de individuos. Una muestra es un subgrupo, es decir, una pequeña parte de una población. La población es el conjunto total o «universo» de todos los individuos o elementos que cumplen ciertas características. Los términos «universo» y población pueden intercambiarse. Suele asumirse que la población total es inaccesible desde un punto de vista práctico y hay que conformarse con estudiar solo una muestra. El interés está en poder extraer conclusiones válidas a partir de una muestra. Esto es interesante, porque las conclusiones serán aplicables o generalizables a la población de la cual se extrajo la muestra. Al proceso de extracción de una muestra a partir de una población se le denomina *muestreo*. La interpretación del tratamiento estadístico de unos datos que acaba generalizándose a toda la población se conoce por *inferencia*. Estos conceptos se representan esquemáticamente en la figura 1.2.

## 1.4. ETAPAS DE LA INVESTIGACIÓN CIENTÍFICA: RELACIONES ENTRE LA BIOESTADÍSTICA Y LA EPIDEMIOLOGÍA

En el proceso de investigación científica se precisa una serie de pasos sucesivos. El ciclo que muestra la figura 1.3 pretende sintetizar estas etapas en el abordaje de un determinado problema de investigación desde la medicina basada en pruebas *(evidence-based)* (5). Este ciclo es iterativo,

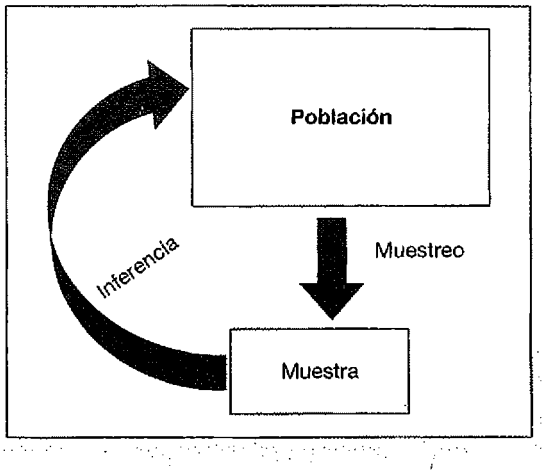



**Figura 1.2** Muestras y poblaciones: procesos de muestreo e inferencia.

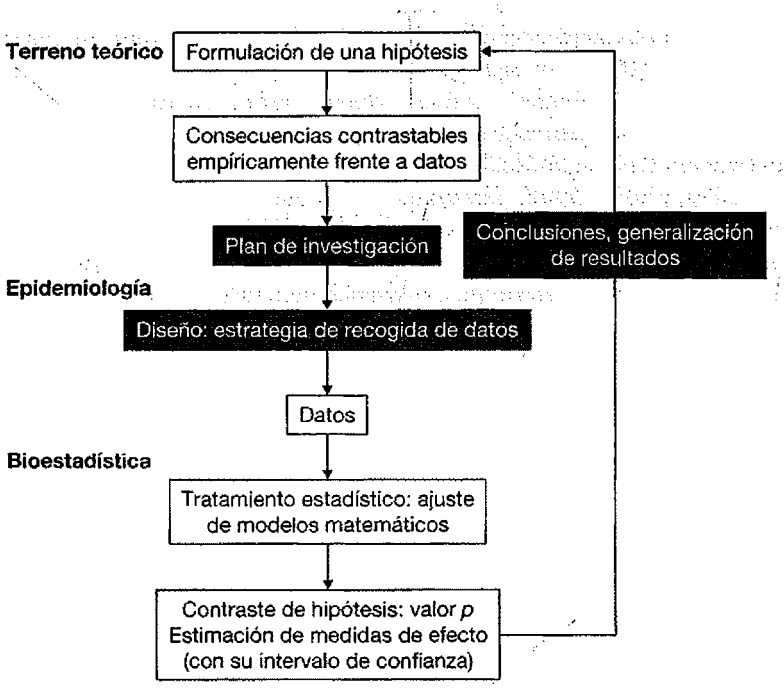



**Figura 1.3** Proceso iterativo de avance del conocimiento científico.

y así va avanzando poco a poco el conocimiento. En este recorrido es importante diferenciar la bioestadística de la epidemiología.

Más que constituir un cuerpo de conocimientos cerrados, la epidemiología es, sobre todo, un método para estudiar una determinada cuestión referente a la salud o la enfermedad. Se trata de un enfoque cuantitativo que puede ser aplicado a una amplia variedad de problemas, desde la

evaluación de la eficacia del tratamiento de una enfermedad a las posibles influencias de la contaminación atmosférica sobre la mortalidad. El método epidemiológico establece el lazo de unión entre el terreno teórico propio de cada área de conocimiento (neurociencia, nutrición, medicina interna, bioquímica, inmunología, genética, etc.) y el terreno matemático característico de los métodos estadísticos. La epidemiología posee un bagaje de técnicas y procedimientos propios, como son, por ejemplo, los diferentes *diseños* que pueden tener los estudios y los modos de prevenir y controlar los *sesgos* o errores sistemáticos (**6-9**). Se pueden cometer sesgos al recoger o analizar datos y, si no se tuviesen en cuenta, llevarían a que las conclusiones acabasen por ser inválidas. El diseño elegido y los diversos detalles que lo componen condicionarán el plan general de la investigación. Se trata de elegir las mejores estrategias de recogida de datos para cada problema que se quiera estudiar. Un correcto enfoque epidemiológico asegura que los datos que se obtienen servirán para establecer conclusiones pertinentes, adecuadas y proporcionadas. La epidemiología garantiza, de esta manera, el nivel científico necesario para el trabajo del profesional sanitario, en sus decisiones y prácticas terapéuticas, y desempeña un papel importante en la calidad global de las funciones ejercidas por los servicios sanitarios (**6,10**). La metodología epidemiológica se ha desarrollado especialmente en las últimas décadas. Gracias al poderoso desarrollo que ha experimentado el tratamiento de los datos con herramientas informáticas, la epidemiología está continuamente cambiando. Va tomando prestadas nuevas técnicas e innovaciones de otras ciencias, fundamentalmente de la bioestadística, pero también de la sociología, la psicología, la economía, etc.

Un ejemplo ayudará a comprender mejor el ciclo seguido para estudiar los problemas científicos. Durante las últimas 2-3 décadas del siglo pasado se había asumido en muchos ambientes científicos una postura algo superficial con respecto a los efectos sobre la salud del cannabis («porro», marihuana). Esto se unía a un aumento de su consumo recreativo entre ciertos sectores de la juventud de Europa, América del Norte y Australia. En general, se tendía a despreciar el efecto perjudicial de su consumo lúdico, y hacia 1990 se pensaba que fumar cannabis no causaba daños sobre la salud a largo plazo (**11**). En algunos ambientes se asumía que fumar cannabis podría incorporarse a las adicciones establecidas y culturalmente aceptables, como el alcohol o la cafeína. El principal ingrediente psicoactivo del cannabis es el delta-9-tetrahidrocannabinol (DTHC), que se identificó y sintetizó por primera vez en 1965.

La primera vuelta al ciclo provino de considerar que clínicamente el uso de cannabis se asociaba a relajación y sensación de bienestar. Se sospechó que el DTHC tendría acciones sobre el sistema nervioso central, debido a sus efectos psicológicos. Así, se valoró la hipótesis de que el cannabis podría dañar al cerebro. Todo lo dicho hasta ahora se mueve en el terreno teórico propio de la neurociencia. Se formula entonces una hipótesis, que consiste en que el cannabis puede perjudicar a los que lo fuman. Esta hipótesis se basa en meras deducciones. El primer intento de responder a esta pregunta científicamente fue fallido. En EE. UU., los *National Institutes of Health* patrocinaron en 1982 un estudio que recogió los datos disponibles y obtuvo una respuesta nula. Tal estudio concluyó que «(...) no hay pruebas de que el cannabis cause daño permanente a la salud... afecte a la estructura cerebral (...) sea adictivo o conduzca al uso de drogas más *duras*» (**12,13**).

Se había recorrido todo el ciclo del contraste de hipótesis. Pero el ciclo es reiterativo y continuamente vuelve al principio. El estudio de 1982 no demostraba la inocuidad, sino que más bien no era capaz de demostrar nada por falta de suficientes datos («...no hay pruebas...»). Se necesitaba mejorar la estrategia de recogida de datos. Una estrategia que ha dado muchas respuestas en investigación biomédica consiste en hacer un estudio longitudinal donde se sigue, de forma paciente y perseverante a lo largo de muchos años, a varios miles de sujetos, de los que se va recogiendo información periódicamente. Así, se puede saber qué relación existe entre sus hábitos y estilos de vida al entrar en el estudio y los fenómenos de salud y enfermedad que les acaban ocurriendo después. Estos estudios longitudinales se llaman *cohortes* y representan una auténtica joya para la investigación (**14**).

En 1987 se publicó el primer estudio longitudinal de cohortes, que encontró que el uso de cannabis predecía el riesgo subsiguiente de desarrollar psicosis. Se formuló la hipótesis de modo deductivo, teniendo en cuenta los mismos hechos que en la primera vuelta dada al ciclo y referida en el párrafo anterior. Ahora, al pasar al plan de investigación se incluyó una muestra de 45.570 reclutas suecos, todos ellos varones. Se valoró a cada recluta cuando tenía 18 años en cuanto a su exposición a drogas y sus diagnósticos y antecedentes psiquiátricos, y después se le siguió durante un período de 15 años, en el cual se contabilizaron los ingresos hospitalarios por esquizofrenia (**15,16**). Aunque la mayoría de los pacientes tratados de esquizofrenia nunca se habían expuesto al cannabis, se encontró que, en el período de seguimiento, que duró hasta que tenían 33 años, aquellos que a los 18 años admitían haber fumado cannabis más de 50 veces eran seis veces más propensos a ser hospitalizados por esquizofrenia que los que nunca lo habían probado. Tras tener en cuenta los factores psiquiátricos, sociales y familiares, así como el uso de otras drogas, el riesgo de desarrollar hospitalización por esquizofrenia era más del doble entre quienes a los 18 años se habían expuesto a cannabis por lo menos 10 veces con respecto a quienes nunca lo habían probado. Esta paciente recogida de datos concluyó que el cannabis parecía causar un daño psiquiátrico grave. No obstante, se debe volver de nuevo hacia el fundamento del contraste de hipótesis para valorar si estos resultados pueden ser válidos, generalizables a mujeres o a varones de otras edades, si cuentan con suficiente plausibilidad biológica, etc. Se entraría así de lleno en el terreno de la epidemiología.

Simultáneamente, se fue sabiendo que el DTHC produce sus efectos psicológicos y musculares mediante la estimulación del receptor cannabinoide 1, que fue identificado en 1988 y clonado en 1990 (**12**). Este receptor se expresa en el sistema nervioso central en el hipocampo, el neocórtex, el cerebelo y los núcleos basales. También se expresa en fibras nerviosas periféricas y en zonas de la médula espinal, lo que puede explicar las propiedades analgésicas de los derivados del cannabis. Existen *endocannabinoides* que son los agonistas naturales del receptor endocannabinoide 1 y de otro que se expresa en menor medida y es conocido como receptor endocannabinoide 2. Los endocannabinoides o cannabinoides naturales son la araquidonoiletanolamida (anandamida) y el 2-araquidonoilglicerol. Estas moléculas actúan habitualmente en nuestro cerebro de manera fisiológica y representan un nivel más alto de regulación fina del papel de los otros neurotransmisores. Se considera que modulan la plasticidad de las conexiones nerviosas (sinapsis) mediadas por glutamato, que se añade a la de neurotransmisores como la serotonina o la dopamina. El papel de los endocannabinoides es modular la liberación de otros neurotransmisores. Funcionan como mecanismos de control y autorregulación, que impiden excesos de liberación de otro neurotransmisor. Se sintetizan localmente y a demanda, de modo que sus dosis estén cuantificadas al detalle para facilitar el correcto funcionamiento del sistema nervioso central. Los endocannabinoides representan mecanismos de gran sutileza. Esto supone un abrupto contraste con la administración brusca de cannabis exógeno, que representa una dramática alteración de la señalización neuronal y de la dinámica de los circuitos cerebrales. Un correlato esperable de todo esto en el plano del comportamiento es que, con el cannabis exógeno, se trastorne el aprendizaje y la memoria. También apoya la hipótesis de que el cannabis puede producir psicosis graves. Todas estas son deducciones, no inducciones. Sirven para generar hipótesis de partida.

Entonces se vuelve a iniciar el circuito del contraste de hipótesis. Sin embargo, el modo actual de razonar en ciencia no se conforma con que exista un mecanismo biológicamente plausible (deducción) para explicar este posible daño, sino que exige que esta hipótesis sea confirmada o refutada frente a datos reales extraídos de la observación (**17**). Para confirmar (o refutar) la hipótesis se debe enfrentar lo establecido teóricamente con unas consecuencias que se puedan verificar en la práctica. Solo así los hechos corroborarán o desmentirán la hipótesis. En esto consiste la *deducción de consecuencias contrastables empíricamente frente a unos datos.* Esta *deducción de consecuencias* lleva a pensar que, si la hipótesis de que el cannabis produce daños psiquiátricos es verdadera,

se encontrarán más casos de psicosis en los expuestos a cannabis que en los no expuestos. A partir de este momento se abandona el terreno teórico y se debe pasar a la metodología epidemiológica. Se debe diseñar una estrategia que permita buscar expuestos y no expuestos, y determinar, tanto en unos como en otros y del modo más económico y válido posible, la frecuencia con la que ocurre la enfermedad psiquiátrica a lo largo del tiempo. Debe procederse cometiendo los mínimos errores en el proceso de selección de participantes en el estudio y en la recogida de información sobre el desarrollo de enfermedad en cada uno de ellos. Especialmente, hay que asegurar la vigilancia exhaustiva y la validez de los diagnósticos, lo cual requiere contestar a muchas preguntas: ¿cuántos sujetos expuestos hacen falta?; ¿cuántos no expuestos?; ¿cuándo se les debe valorar?; ¿cómo?; ¿durante cuánto tiempo hay que seguirles?; ¿con qué periodicidad?; ¿qué otra información se debe recoger de cada uno de ellos?; ¿cómo se cuantifica esta información?; ¿cuántas veces se deben repetir las mediciones?; ¿cuáles son elegibles?; ¿cuáles son los casos y los no casos?; ¿qué debe saber y qué debe ignorar quien recoge la información sobre el diagnóstico de psicosis?; ¿qué datos se les deben dar a los pacientes y a los examinadores antes de que sean examinados?, etc.

La respuesta a todas estas preguntas (y muchas más) constituye el *plan de investigación,* que, como puede comprenderse, incluye muchos detalles, en apariencia pequeños, pero de suma importancia. Este plan pone en marcha una estrategia de recogida de datos, que probablemente requiera una gran movilización de recursos y personas. Así lo hicieron los autores de diversos estudios sobre esta cuestión.

Se fueron realizando estudios de cohortes similares al de los reclutas suecos. Así, en 50.000 varones seguidos desde 1969, se observó que el consumo de por lo menos 50 porros en su vida multiplicaba por 7 el riesgo de padecer una esquizofrenia (**18**). En otro estudio, al valorar a 1.253 personas de Nueva Zelanda seguidas desde su nacimiento hasta cumplir los 26 años, haber consumido cannabis alguna vez en su juventud multiplicaba por casi 4 el riesgo de padecer una esquizofrenia (**19**). En otro estudio realizado en los Países Bajos con 4.045 personas seguidas durante 3 años, se observó que el consumo de porros multiplicaba por 3 el riesgo de presentar algún síntoma psicótico, por 24 el de presentar síntomas psicóticos graves y por 12 el riesgo de necesitar en algún momento atención médica por síntomas psicóticos (**20**).

Los datos fueron analizados estadísticamente ajustando modelos matemáticos y calculando con estos modelos cuántas veces era más frecuente desarrollar psicosis entre los que habían estado más expuestos a cannabis que entre quienes nunca lo habían probado. Estamos ya en el terreno de la bioestadística. Se encontró un riesgo significativamente superior en los expuestos a cannabis, con lo que parecía corroborarse la hipótesis.

Una vez abandonado el terreno de la bioestadística, y antes de volver de nuevo al área teórica para confirmar la hipótesis, es preciso detenerse en el campo de la epidemiología para interpretar críticamente los resultados que salieron del ordenador y acabaron por publicarse (**15,16,18-20**). No se puede prescindir nunca del razonamiento epidemiológico cuando se piensa que se ha obtenido una conclusión a partir del análisis estadístico de unos datos. Prescindir de esta reflexión sosegada y sensata es más arriesgado cuanto más rápidos y fáciles de usar son los ordenadores. Así, es posible dejarse llevar por un peligroso *automatismo* al interpretar lo que el ordenador produce a partir de unos datos, fijándose más en si existen resultados estadísticamente significativos que en realizar una interpretación concienzuda con el sentido crítico que debe caracterizar a cualquier investigación. Cuando se realizan análisis estadísticos, hay que mantenerse siempre mentalmente en contacto con el origen que tuvieron esos datos, es decir, pensar en las debilidades y limitaciones que pueden tener el diseño y los métodos de recogida de esos datos. Hay que valorar posibles errores de los métodos o instrumentos de medición que se utilizaron. No se debe perder el contacto con los objetivos primordiales del estudio y, sobre todo, con la plausibilidad biológica de las hipótesis que se contrastan. Se debe mantener también una conciencia crítica acerca de

los datos o variables que faltan porque no se han recogido. Integrar todos estos aspectos con el conocimiento de los procedimientos que realmente están indicados para cada tipo de variables, sus condiciones de aplicación y los fundamentos de las técnicas bioestadísticas empleadas es lo que capacita para desarrollar una investigación empírica realmente válida y fecunda. Para que sea factible, se requiere como punto de partida mucho conocimiento biológico sobre el problema concreto que se esté investigando.

Teniendo en cuenta todo lo anterior, el método epidemiológico consideraría provisional la conclusión encontrada en los ejemplos precedentes, ya que sería preciso valorar también otros aspectos.

Así, en 2004, un grupo de expertos epidemiólogos sometieron a crítica la posible relación causal del cannabis con los efectos adversos para la salud mental (o el comportamiento social). Para ello realizaron una revisión sistemática (es decir, exhaustiva) de todos los estudios publicados. Revisaron 48 estudios publicados, de los cuales 16 reunían criterios de alta calidad metodológica. Estos autores encontraron problemas para extraer de estos estudios una clara asociación *causa-efecto* (**21**). Desde luego, afirmaron, existían en estos estudios pruebas suficientes de una asociación entre la exposición a cannabis y el daño psíquico. De todos modos, tras realizar su revisión, consideraron que la magnitud y la fuerza de estas pruebas parecían ser inferiores a lo que se venía asumiendo. Diversos problemas psíquicos pueden ser más una causa que una consecuencia del uso de cannabis. Así, la posible existencia de problemas psicológicos no declarados o de carácter subclínico, que podrían haber precedido y precipitado el uso de cannabis, explicaría una asociación, pero la causalidad tendría la dirección inversa. Esto sería teóricamente posible. Podría pensarse que las personas con una tendencia previa a padecer dificultades psicológicas pueden tener también una mayor inclinación a desarrollar patrones problemáticos de uso de drogas. Por otra parte, la exposición a cannabis podría exacerbar una predisposición al daño psíquico, de modo que los efectos adversos del cannabis solo estarían presentes en quienes tuviesen tal predisposición. Además, tanto el uso de cannabis como los problemas psíquicos parecen compartir antecedentes comunes, por ejemplo adversidades padecidas en la infancia o factores relacionados con las amistades o la estructura y el ambiente familiar. Es decir, la asociación entre cannabis y daño psíquico podría explicarse simplemente porque exista algún otro factor que cause ambos efectos, y no porque el cannabis cause el daño psíquico. Todas estas explicaciones alternativas a la causalidad forman parte de la interpretación y discusión de resultados, que es un paso imprescindible en cualquier investigación. Consiste en interpretar *críticamente* los resultados encontrados. Pertenece al terreno de la epidemiología.

Actualmente, en la investigación biomédica se ajustan modelos matemáticos. Tales modelos se usan para equiparar los grupos comparados en cuanto a esas otras características alternativas a la causalidad (antecedentes de enfermedad psíquica, ambiente familiar u otras posibles causas comunes aducidas). A esas variables asociadas tanto a la exposición (cannabis en el ejemplo) como al efecto (psicosis) se les llama *factores de confusión*. Los modelos matemáticos controlan, de algún modo, esa posible confusión, pues son capaces de presentar las asociaciones bajo el supuesto de *igualdad* de los demás factores. De todos modos, los autores de la revisión sistemática de 2004 (**21**) creían que, probablemente, quedaban factores sin controlar, o que estaban mal controlados en los estudios revisados (confusión residual). Estos autores, tras revisar las publicaciones entonces disponibles, consideraban que unos posibles factores denominados conjuntamente «propensión al uso de drogas» podrían todavía explicar, a pesar del ajuste matemático realizado, las asociaciones entre la exposición a cannabis y la incidencia de esquizofrenia al margen de la causalidad. Por otra parte, aducían que el consumo de cannabis se había incrementado sustancialmente en sectores de la juventud en los últimos 30 años. Por ejemplo, el 10% admitía haberlo probado al menos alguna vez en 1969-1970, pero este porcentaje había subido al 50% en 2001 en Gran Bretaña y Suecia. Una relación causal entre cannabis y esquizofrenia se habría acompañado, probablemente, de un incremento notable en las tasas de esquizofrenia (**21**).

La conclusión, tras estas y otras consideraciones, fue que en esos momentos todavía no se encontraban pruebas fuertes de que el consumo de cannabis en sí mismo tuviese consecuencias importantes para la salud psíquica (o social). Los mismos autores reconocían, no obstante, que «este hallazgo no equivale a la conclusión de que el uso de cannabis esté exento de daños en los ámbitos psicosociales. Los mismos problemas de las pruebas y estudios disponibles hacen igualmente indefendible esta proposición. Se necesitan mejores evidencias» (**21**).

Se había dado otra vuelta a todo el ciclo, y el resultado era que había evidencias (aunque de naturaleza débil) sobre la hipótesis inicialmente formulada. La prudencia exige que se encuentre consistencia, es decir, que otros estudios diferentes también apunten en la misma dirección, y solo entonces pueda empezar a pensarse que la asociación propuesta es verdadera. De todos modos, si la metodología utilizada en todos los estudios fuese la misma, no podría excluirse que se debiera a un sesgo inherente a ese diseño. Un sesgo repetido 10 veces sigue siendo un sesgo. Es más convincente una asociación encontrada por estudios que utilizan diferentes diseños, con diferentes tipos de individuos y en diferentes países.

En este estado de cosas, a pesar de las conclusiones débiles de la revisión sistemática de 2004, otros epidemiólogos defendían el *principio de precaución* (**22**). Este principio mantiene que, en caso de amenazas serias o irreversibles para la salud de las personas o los ecosistemas, la existencia de incertidumbre científica no debería invocarse como excusa para posponer las medidas preventivas (**23**). Se propuso, por tanto, ya en 2005, actuar desde la salud pública, con la evidencia entonces disponible, para conseguir limitar unas exposiciones libremente elegidas al uso recreativo del cannabis ante la posibilidad real de que supusiesen un riesgo de causar enfermedad psíquica grave (**22**).

Algún tiempo después se puso en marcha un diseño más poderoso de investigación. Se realizó un metaanálisis *cuantitativo,* algo que no realizaron los autores de la revisión sistemática de 2004. El metaanálisis *(más allá del análisis o análisis de los análisis)* es un conjunto de técnicas utilizadas para cuantificar la información contenida en estudios similares que valoran una misma pregunta de investigación (**24**). El resultado de este metaanálisis pudo proporcionar un fuerte apoyo a la causalidad de la asociación entre el cannabis y la psicosis (**25**). Los autores concluyeron que «la evidencia es consistente con la visión de que el cannabis incrementa el riesgo de enfermedades psicóticas independientemente de los factores de confusión y de los efectos de intoxicaciones transitorias (...), y que ya hay suficientes pruebas para advertir a las personas jóvenes que exponerse a cannabis puede incrementar su riesgo de desarrollar una enfermedad psicótica posteriormente en su vida». La estimación del efecto fue de un incremento relativo del riesgo del 41% (intervalo de confianza del 95% [IC 95%]: 20% a 65%) para quienes probaron cannabis alguna vez y del 109% (IC 95%: 54% a 184%) para quienes lo consumieron más frecuentemente (**25**). Esto supuso una refutación empírica de la hipótesis nula inicial de que el cannabis carecía de efectos adversos para la salud. Se parte siempre de una hipótesis nula y se usan los datos para refutarla (**26**). Otros estudios realizados más recientemente en otros países y con otro tipo de poblaciones han encontrado también, consistentemente, que el cannabis resulta perjudicial psíquicamente, con pruebas de una relación causal. Además, se ha constatado un aumento de los casos de esquizofrenia asociados al consumo de cannabis en personas jóvenes. Actualmente, la relación causa-efecto del cannabis con la psicosis no solo se admite científicamente, sino que se ha considerado perfectamente equiparable a otras relaciones causales de los factores de riesgo que se incluyen de modo habitual en las estimaciones de la carga global de enfermedad (**27-29**).

Esta historia proporciona una lección importante, que consiste en que no se deben precipitar conclusiones científicas antes de completar varias veces el ciclo previamente mencionado, que va desde la hipótesis teórica hasta las conclusiones empíricas. Cada vez que se recorre el ciclo se usa un diseño más riguroso y con mayor capacidad de excluir los sesgos. En la historia del cannabis y la psicosis nunca sería ético usar un diseño experimental donde se expusiese a unos jóvenes a

cannabis y a otros no, con adjudicación al azar. Pero se podría valorar *indirectamente* si se asignase al azar un grupo a un programa de intervención intensivo para evitar el uso de cannabis y a otro grupo a los cuidados de salud convencionales. Tras seguirles varios años se podría verificar si se redujo la incidencia de psicosis con el programa de evitación del cannabis. Si la adjudicación a los grupos fuese al azar (ensayo aleatorizado), entonces se obtendrían (por la ley de los grandes números) grupos exactamente iguales de sujetos, unos con intervención intensiva y otros sin ella, y todas las diferencias entre ellos en cuanto a la ocurrencia futura de psicosis se podrían atribuir con mucha fuerza probatoria de la causalidad solo al programa de evitación de cannabis, puesto que esta sería la única diferencia entre ellos. Esto todavía no se ha hecho. Quizá no sea posible o no se considere ético. Sin embargo, sí se ha realizado en otras áreas de investigación. Así ha sucedido con otra historia, esta vez protectora, que es la relación entre el seguimiento de una dieta mediterránea y la prevención de episodios cardiovasculares (y de otros procesos), en la que se fueron dando sucesivos pasos, todos ellos congruentes (**30-35**). El diseño más riguroso es el que usó la aleatorización (asignación al azar) de 7.447 participantes seguidos luego durante 4,8 años. Así, se demostró en 2013 una reducción relativa del 30% en el riesgo de episodios cardiovasculares graves (**34**). Cuando se ha completado el ciclo varias veces y se usa el mejor diseño posible, dentro de lo que resulta ético o factible, puede hablarse propiamente de una relación causa-efecto demostrada.

La epidemiología procede por acumulación de pruebas *(evidence),* cada vez más convincentes, de que un determinado factor se asocia con un determinado hecho o resultado (**6,10,36,37**). Es decir, una vez que el ciclo dibujado en la figura 1.3 da una y otra vuelta, se van perfilando las hipótesis, se desecha lo que no vale, queda la ciencia consolidada y se va conociendo mejor la verdad de las cosas. Se comienza con un proceso deductivo, que va de arriba hacia abajo. Este proceso consiste en que, a partir de principios generales, se hacen conjeturas sobre consecuencias de la hipótesis que podrían enfrentarse a unos datos. Después se pasa al proceso inductivo: desde los datos empíricos se asciende hacia las conclusiones. De esta combinación de deducción e inducción surge el conocimiento.

La filosofía que subyace a la epidemiología y a la bioestadística le debe mucho a Karl Popper (**26**). Este filósofo estableció que nunca se puede *demostrar* la verdad de una hipótesis. Popper mantenía que la única solución posible es *refutar* la hipótesis contraria y que, por tanto, todo conocimiento científico es efímero y provisional. Sin embargo, no se puede llevar al extremo este modo de ver las cosas. La moderación exige afirmar que hay hechos (aunque sean pocos) que ya están plenamente demostrados en ciencia. Por ejemplo, la sangre circula por las venas, la hemoglobina transporta oxígeno, el ADN contiene nuestro código genético, consumir cannabis es causa de psicosis y la dieta mediterránea reduce el riesgo cardiovascular.

## REFERENCIAS


1. Rao G, Kanter SL. Physician numeracy as the basis for an evidence-based medicine curriculum. Acad Med 2010;85(11):1794-9.
2. Altman DG, Goodman SN. The transfer of technology from statistical journals to the biomedical literature. JAMA 1994;272(2):129-32.
3. http://es.wikipedia.org/wiki/Bioestadística (consultado el 24 de junio de 2013).
4. Martín Andrés A, Luna del Castillo JD. Bioestadística para las ciencias de la salud. 5.ª ed. Madrid: Norma-Capitel Ediciones; 2004.
5. De Leon J. Evidence-based medicine versus personalized medicine. Are they enemies? J Clin Pharmacol 2012;32(2):153-64.

6. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel; 2008.

7. Sackett DL. Bias in Analytic Research. J Chron Dis 1979;32(1-2):51-63.

8. Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health 2004;58(8):635-41.

9. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359(9302):248-52.

10. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

11. Anónimo. Deglamorising cannabis. Lancet 1995;346(8985):1241.

12. Murray RM, Morrison PD, Henquet C, Di Forti M. Cannabis, the mind and society: the hash realities. Nat Review Neurosci 2007;8(11):885-95.

13. Relman A, Marijuana, Health. Report of a Study by a Committee of the Institute of Medicine, Division of Health Sciences Policy. Washington D.C.: National Academy Press; 1982.

14. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet 2002;359:341-5.

15. Andreasson S, Allebeck P, Engstrom A, Rydberg U. Cannabis and schizophrenia: A longitudinal study of Swedish conscripts. Lancet 1987;330(8574):1483-6.

16. McLaren JA, Silins E, Hutchinson D, Mattick RP, Hall W. Assessing evidence for a causal link between cannabis and psychosis: A review of cohort studies. Int J Drug Policy 2010;21(1):10-9.

17. Maclure M. Taxonomic axes of epidemiologic study designs: a refutationist perspective. J Clin Epidemiol 1991;44(10):1045-53.

18. Zammit S, Allebeck P, Andreasson S, Lundberg I, Lewis G. Self-reported cannabis use as a risk factor for schizophrenia in Swedish conscripts of 1969: Historical cohort study. BMJ 2002;325(7374):1199-201.

19. Arseneault L, Cannon M, Poulton R, Murray R, Caspi A, Moffitt TE. Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. BMJ 2002;325(7374):1212-3.

20. van Os J, Bak M, Hanssen M, Bijl RV, de Graaf R, Verdoux H. Cannabis use and psychosis: a longitudinal population-based study. Am J Epidemiol 2002;156(4):319-27.

21. Macleod J, Oakes R, Copello A, Crome I, Egger M, Hickman M, et al. Psychological and social sequelae of cannabis and other illicit drug use by young people: A systematic review of longitudinal, general population studies. Lancet 2004;363(9421):1579-88.

22. De Irala J, Ruiz-Canela M, Martínez-González MA. Causal relationship between cannabis use and psychotic symptoms or depression. Should we wait and see? A public health perspective. Med Sci Monit 2005;11(12):355-8.

23. Martínez-González MA, Guillén-Grima F, Delgado-Rodríguez M. Conceptos de salud pública. En: Martínez-González MA, ed. Conceptos de salud pública y estrategias preventivas: un manual para ciencias de la salud. Barcelona: Elsevier; 2013. p. 9-14.

24. Revisión sistemática y metaanálisis. En: Delgado-Rodríguez M, Sillero Arenas M, Martínez-González MA, editores. Conceptos de salud pública y estrategias preventivas: un manual para ciencias de la salud. Barcelona: Elsevier; 2013. p. 55-62.

25. Moore THM, Zammit S, Lingford-Hughes A, Barnes TRE, Jones PB, Burke M, et al. Cannabis use and risk of psychotic or affective mental health outcomes: A systematic review. Lancet 2007;370(9584):319-28.

26. Glass DJ, Hall N. A brief history of the hypothesis. Cell 2008;134:378-81.

27. Large M, Sharma S, Compton MT, Slade T, Nielssen O. Cannabis use and earlier onset of psychosis. Arch Gen Psychiatry 2011;68(6):555-61.

28. Freedman R. Cannabis, inhibitory neurons, and the progressive course of schizophrenia. Am J Psychiatry 2008;165(4):416-9.

29. Degenhardt L, Hall WD, Lynskey M, McGrath J, McLaren J, Calabria B, et al. Should burden of disease estimates include cannabis use as a risk factor for psychosis? PLoS Med 2009;6(9):e1000133.

30. Martínez-González MA, Fernández-Jarne E, Serrano-Martínez M, Marti A, Martínez JA, Martín-Moreno JM. Mediterranean diet and reduction in the risk of a first acute myocardial infarction: an operational healthy dietary score. Eur J Nutr 2002;41(4):153-60.

31. Martínez-González MA, Estruch R. Mediterranean diet, antioxidants and cancer: the need for randomized trials. Eur J Cancer Prev 2004;13(4):327-35.

32. Martínez-González MA, García-López M, Bes-Rastrollo M, Toledo E, Martínez-Lapiscina E, Delgado-Rodriguez M, et al. Mediterranean diet and the incidence of cardiovascular disease: A Spanish cohort. Nutr Metab Cardiov Dis 2011;21(4):237-44.

33. Martínez-González MA, Corella D, Salas-Salvadó J, Ros E, Covas MI, Fiol M, et al., for the PREDIMED Study Investigators. Cohort Profile: design and methods of the PREDIMED study. Int J Epidemiol 2012;41(2):377-85.

34. Estruch R, Ros E, Salas-Salvadó J, Covas MI, Corella D, Arós F, et al. for the PREDIMED investigators. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 2013;368(14):1279-90.

35. Martínez-González MA, Bes-Rastrollo M. Dietary patterns, Mediterranean diet, and cardiovascular disease. Curr Opin Lipidol 2014;25(1):20-6.

36. Szklo M, Nieto FJ. Epidemiología Intermedia. Madrid: Díaz de Santos; 2003.

37. Rothman KJ. Epidemiology: An Introduction. New York: Oxford University Press; 2002.

# 2 PROCEDIMIENTOS DESCRIPTIVOS

*M. Á. Martínez-González, A. Gea, C. Sayón-Orea*

## 2.1. TIPOS DE VARIABLES

### 2.1.1. Variables y bases de datos

Cualquier investigación suele exigir una fase de *recogida de datos.* Estos datos son la materia prima de la bioestadística. A partir de ellos se calculan otros números, los índices estadísticos, que extraen la información importante contenida en los datos. A las cualidades o cantidades recogidas de cada individuo se les llama *variables,* porque pueden variar de un sujeto a otro, a diferencia de las *constantes,* que se estudian en otras materias, pero no en estadística (**1**).

Resulta básico distinguir los diferentes tipos de variables según las escalas que se usen para medirlas. Diferenciar con claridad los tipos de variables previene muchos errores en la aplicación e interpretación de cualquier procedimiento estadístico (**2**). ¿Qué tipos de datos son los que se suelen recoger?

Supóngase que se desea evaluar la efectividad de un programa dirigido a conseguir que personas fumadoras con diabetes abandonen el tabaco. Se incluyeron 280 fumadores diabéticos, que fueron asignados al azar (aleatoriamente), o bien a una intervención especial de una enfermera entrenada o a un grupo control de cuidados habituales (**3**). Se recogieron —entre otros— datos como los presentados en el cuadro 2.1.

Lo que se presenta en el cuadro 2.1 es una especie de *diccionario* o conjunto de *etiquetas* de una base de datos. Incluye la información pertinente para comprender qué significará, en adelante, cada número en esa base de datos. Hay números que se usan solo como *códigos* (p. ej., 1 = varón, 2 = mujer). Los códigos son los valores 1 y 2; las etiquetas de esos valores son los nombres *varón* y *mujer.* Basta con decirle al ordenador una vez esos nombres para que luego los escriba automáticamente cada vez que en esa columna encuentre el 1 o el 2. Se recomienda vivamente que *todas* las variables de una base de datos se introduzcan en el ordenador en forma de números (y no de letras) mientras eso sea posible. A la derecha, entre paréntesis, se ha indicado el nombre que se va a dar a cada *variable.* Junto con la definición de cada variable, aparecen, además, los *códigos* que puede tomar cada uno de sus posibles valores. El proceso de cambiar los nombres de las categorías por números (1 = varón, 2 = mujer) se denomina *codificación.* Requiere añadir un código a cada etiqueta. En el tratamiento de datos hay dos posibles significados de la palabra *etiqueta* (*label,* en inglés). El primero corresponde a la etiqueta de *cada valor* que puede tomar esa variable (varón es la etiqueta del valor o código 1 para la segunda variable). A este primer aspecto se refiere la codificación. El segundo significado es la etiqueta de la *variable* (sexo, en la etiqueta de la segunda variable). Se suele buscar un nombre breve (mejor si tiene ocho letras o menos) para la etiqueta de cada variable. Al asignar nombres se debe evitar la letra «ñ», las tildes (acentos) o cualquier signo de puntuación distinto del guion bajo (*year_0* sería aceptable, pero tanto *día* como *año0* o *fumador?* darán problemas). Además del nombre breve, se puede poner a toda la variable una etiqueta de mayor longitud.

Los datos se suelen guardar en un fichero que contiene filas y columnas. A este fichero se le llama *base de datos.* Cada fila suele corresponder a un individuo y cada columna, a una variable.

**CUADRO 2.1 EJEMPLO DE VARIABLES CONTENIDAS EN UNA BASE DE DATOS (DICCIONARIO O *ETIQUETAS* DE LA BASE DE DATOS)**

1. Número de identificación («id»)
2. SEXO; etiquetas o códigos («sexo»):
   a. 1 = varón
   b. 2 = mujer
3. *EDAD* (años cumplidos) *(«edad»)*
4. *INICIO del tabaquismo (edad* en *que empezó a fumar) («inicio»)*
5. *Número de cigarrillos fumados al día («numcig»)*
6. *Peso (en kg, aproximado a los 200 g más cercanos) («peso»)*
7. *Talla (en cm) («talla»)*
8. *GRUPO de tratamiento; etiquetas* o códigos *(«grupo»)*:
   a. 1 = Intervención
   b. 2 = Control
9. *Estado civil; etiquetas* o códigos *(«ecivil»)*:
   a. 1 = Soltero
   b. 2 = Casado
   c. 3 = Viudo
   d. 4 = Divorciado/separado
   e. 5 = Otros
10. *FECHA* de inclusión-día *(«dia_0»)*
11. *FECHA* de inclusión-mes *(«mes_0»)*
12. *FECHA* de inclusión-año *(«year_0»)*
13. ESTUDIOS: máximo nivel educativo alcanzado; *etiquetas* o códigos *(«estudios»)*:
    a. 1 = Analfabeto
    b. 2 = Sin estudios
    c. 3 = Estudios primarios incompletos
    d. 4 = Estudios primarios completos
    e. 5 = Bachiller elemental
    f. 6 = Bachiller superior
    g. 7 = Formación profesional
    h. 8 = Estudios universitarios
14. INTERÉS en dejar de fumar; *etiquetas* o códigos *(«interes»)*:
    a. 0 = Ninguno
    b. 1 = Poco
    c. 2 = Regular
    d. 3 = Mucho
15. *FECHA* del final del seguimiento-día *(«dia_5»)*
16. *FECHA* del final del seguimiento-mes *(«mes_5»)*
17. *FECHA* del final del seguimiento-año *(«year_5»)*
18. ABANDONO del tabaco al final del estudio *(«abandono»)*:
    a. 1 = Sí
    b. 2 = No

Algunas veces, esta clase de base de datos se llama de formato ancho *(wide)*, porque tiene más columnas, tantas como variables. En cambio, cuando cada individuo ocupa varias filas (p. ej., porque hay medidas repetidas), el formato se llama largo *(long)*. Al conjunto completo de las variables de todos los individuos se le llama *base de datos*. La tabla 2.1 recoge una base de datos (formato ancho). Esta base de datos se llama canga25.xls y puede descargarse desde http://www.unav.es/departamento/preventiva/recursos_bioestadistica (fig. 2.1).

*Tabla 2.1 Ejemplo de transformación de una variable cuantitativa (número de cigarrillos/día) en una variable cualitativa ordinal (categorización)*

| VARIABLE ANTIGUA = *NUMCIG* | VARIABLE NUEVA = *CIGGRUP* | ETIQUETAS* |
|---|---|---|
| 1-19 | 1 | «De 1 a 19 cig/d» |
| 20-39 | 2 | «De 20 a 39 cig/d» |
| 40-máximo | 3 | «40 o más cig/d» |

**Adviértase que las etiquetas admiten tildes, símbolos y letras diversas.*

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | sexo | edad | inicio | numcig | peso | talla | grupo | ecivil | dia>0 | mes>0 | year>0 | estudios | interes | dia>5 | mes>5 | year>5 | abandono |
| 2 | 1 | 2 | 59 | 20 | 20 | 49 | 150 | 2 | 1 | 19 | 3 | 1998 | 3 | 2 | 25 | 5 | 1998 | 2 |
| 3 | 2 | 1 | 21 | 13 | 20 | 78 | 173 | 1 | 2 | 11 | 2 | 1998 | 7 | 0 | 28 | 5 | 1998 | 2 |
| 4 | 3 | 1 | 54 | 13 | 11 | 83 | 178 | 2 | 1 | 11 | 3 | 1998 | 6 | 2 | 12 | 8 | 1998 | 2 |
| 5 | 4 | 1 | 66 | 15 | 20 | 75 | | 2 | 1 | 13 | 6 | 1998 | 6 | 1 | 22 | 10 | 1998 | 2 |
| 6 | 5 | 1 | 51 | 45 | 60 | 80 | 170 | 1 | 1 | 6 | 2 | 1998 | 7 | 1 | 12 | 6 | 1998 | 2 |
| 7 | 6 | 2 | 27 | 13 | 20 | 57 | 158 | 1 | 2 | 26 | 1 | 1998 | 7 | 0 | 13 | 6 | 1998 | 2 |
| 8 | 7 | 1 | 41 | 16 | 40 | 51 | 157 | 2 | 2 | 13 | 1 | 1998 | 6 | 3 | 15 | 6 | 1998 | 2 |
| 9 | 8 | 2 | 59 | 20 | 15 | 47 | 155 | 1 | 2 | 21 | 1 | 1998 | 5 | 1 | 15 | 6 | 1998 | 1 |
| 10 | 9 | 1 | 31 | 16 | 20 | 61 | 163 | 1 | 2 | 26 | 1 | 1998 | 8 | 1 | 16 | 6 | 1998 | 1 |
| 11 | 10 | 1 | 62 | 14 | 10 | 48 | 155 | 2 | 2 | 11 | 4 | 1998 | 2 | 2 | 16 | 6 | 1998 | 2 |
| 12 | 11 | 1 | 22 | 14 | 7 | 63 | 173 | 2 | 2 | 14 | 1 | 1998 | 6 | 2 | 17 | 6 | 1998 | 2 |
| 13 | 12 | 1 | 51 | 15 | 8 | 78 | | 2 | 1 | 12 | 6 | 1998 | 5 | 2 | 27 | 2 | 1999 | 2 |
| 14 | 13 | 1 | 50 | 21 | 30 | 77 | 164 | 1 | 1 | 15 | 1 | 1999 | 5 | 1 | 18 | 9 | 1999 | 2 |
| 15 | 14 | 1 | 67 | 19 | 30 | 64 | 169 | 2 | 1 | 17 | 3 | 1998 | 4 | 1 | 19 | 6 | 1998 | 2 |
| 16 | 15 | 2 | 56 | 17 | 20 | 100 | 155 | 1 | 4 | 3 | 2 | 1998 | 2 | 1 | 19 | 6 | 1998 | 2 |
| 17 | 16 | 1 | 19 | 13 | 15 | 68 | 170 | 2 | 2 | 19 | 5 | 1998 | 5 | 2 | 22 | 6 | 1998 | 2 |
| 18 | 17 | 2 | 20 | 14 | 15 | 52 | 165 | 2 | 2 | 15 | 4 | 1998 | 5 | 2 | 13 | 7 | 1998 | 2 |
| 19 | 18 | 1 | 69 | | 30 | 71 | 166 | 2 | 1 | 14 | 5 | 1998 | 2 | 2 | 13 | 7 | 1998 | 2 |
| 20 | 19 | 2 | 42 | 20 | 20 | 55 | 167 | 1 | 1 | 16 | 2 | 1998 | 8 | 1 | 13 | 7 | 1998 | 2 |
| 21 | 20 | 1 | 64 | 16 | 10 | 67 | 175 | 1 | 1 | 28 | 2 | 1998 | 5 | 2 | 13 | 7 | 1998 | 2 |
| 22 | 21 | 1 | 72 | 9 | 20 | 90 | 170 | 2 | 2 | 19 | 2 | 1998 | 2 | 1 | 13 | 7 | 1998 | 2 |
| 23 | 22 | 1 | 71 | 22 | 20 | 67 | 165 | 1 | 1 | 3 | 3 | 1998 | 2 | 1 | 15 | 7 | 1998 | 2 |
| 24 | 23 | 1 | 75 | 10 | 20 | 75 | 165 | 1 | 1 | 2 | 3 | 1998 | 1 | 1 | 17 | 7 | 1998 | 2 |
| 25 | 24 | 1 | 47 | 10 | 30 | 90 | 175 | 1 | 1 | 16 | 2 | 1998 | 4 | 2 | 16 | 7 | 1998 | 2 |
| 26 | 25 | 1 | 77 | 10 | 20 | 68 | 170 | 1 | 1 | 3 | 2 | 1998 | 2 | 1 | 16 | 7 | 1998 | 2 |

**Figura 2.1** Aspecto de una base de datos (contenido de la base de datos).

Es evidente que encontrar el número 1 en la variable *«sexo»* no tiene el mismo significado que hallar ese mismo número en la variable *«numcig»*. En el primer caso *(sexo)*, el número es solo un indicador o código que se ha querido asignar en este ejemplo a todos los individuos de sexo masculino. En cambio, en la variable *«numcig»*, el número 1 sería una cantidad real, ya que correspondería exactamente a fumar un cigarrillo al día. La variable *«sexo»* es una variable *cualitativa* o *categórica;* en cambio, la variable *«numcig»* es una variable *cuantitativa.* Las variables *cualitativas* o *categóricas* están medidas en una escala *nominal.* Aunque a sus valores se les asignen códigos numéricos, los números son, en realidad, una *traducción* de sus nombres. Por su parte, la escala de las variables cuantitativas corresponde a la de los valores reales de los números que toman.

Una diferencia fundamenral entre ambas escalas es que, por ejemplo, el número 20 en la columna de la variable *numcig* corresponde a un valor que es exactamente la mitad del de otra casilla que tuviese el número 40, y también vale exactamente el doble que cuando la casilla contenga un 10. En cambio, cuando el número 2 figura en la variable *grupo* no supone que quienes pertenecen al grupo control valgan el doble que los del grupo de intervención, pues a efectos prácticos hubiese dado lo mismo (incluso hubiera sido preferible) codificar el control con 0 y la intervención con 1. En variables correspondientes a tener o no una característica, es muy oportuno codificar con un 1 a quienes la tienen y con un 0 a quienes carecen de ella. Por eso, en bioestadística el uso del 0 suele ser abundante.

Además de *numcig*, otras variables como *edad*, *peso* y *talla* son *cuantitativas*, y, por lo tanto, los datos que aparecen en ellas corresponden realmente a números. En cambio, *sexo*[1], *grupo*, *ecivil* y *abandono* son variables *cualitativas o categóricas*.

A su vez, dentro de las variables cuantitativas o realmente numéricas hay dos posibilidades: las que admiten cualquier valor dentro de un intervalo *(continuas)*, sin más restricción que el límite del aparato de medida, y las que solo pueden tomar números enteros *(discretas)*. El peso y la talla son *variables cuantitativas continuas*, ya que, teóricamente, un individuo puede tener un peso que tome cualquier valor, por ejemplo entre 80 y 81 kg; podría pesar realmente 80,3333693 kg, y lo mismo se aplica para la talla. En cambio, otras variables (p. ej., si se hubiese recogido el número de intentos de dejar de fumar) solo pueden tomar números enteros. Nadie puede realmente haber intentado dejar de fumar 1,33 veces. Estas variables que solo pueden tomar valores de números enteros se conocen por *variables cuantitativas discretas*.

Queda por definir otro tipo de variables que están en una situación intermedia. Se trata, por ejemplo, del interés en dejar de fumar *(interes)*. En este tipo de variables se puede decir que un grado 2 de interés es más intenso que un grado 1, pero nunca puede interpretarse como que tener un código 2 implique *exactamente* el doble de interés que el 1. Este tipo de variables se llaman *ordinales* y su uso es muy frecuente en medicina. Así, el dolor se puede clasificar en ausente/leve/moderado/intenso, y se asignarán respectivamente los códigos 0/1/2/3 a cada categoría. El máximo nivel de estudios alcanzado es otra variable *ordinal*. No lo es, sin embargo, el estado civil, pues no sería estadística ni *políticamente correcto* ordenar o jerarquizar los diferentes estados civiles. La respuesta a un tratamiento podría valorarse mediante una escala ordinal, asignando, por ejemplo, el código –1 a los que empeoran, el 0 a los que quedan igual, el +1 a los que mejoran algo y el +2 a los que mejoran mucho. En algunas determinaciones analíticas se siguen usando —cada vez menos— escalas ordinales en algunos ámbitos; por ejemplo, para evaluar si un paciente tiene glucosa en la orina (glucosuria) podría cuantificarse así: 0/+/++/+++.

### 2.1.2. Variables cualitativas o categóricas nominales

Son variables en las que cada categoría o valor corresponde a una característica o cualidad que la persona posee. Los posibles valores son excluyentes (sexo, asignación a un grupo de tratamiento o a otro, haber dejado o no el tabaco, haber tenido un infarto, etc.). Son *nominales* todas las variables cuyo valor simplemente pueda ser un sí (1) o un no (0), o corresponda a más de dos clases no ordenables jerárquicamente.

Para su medición usamos escalas *nominales*, donde los valores se identifican con palabras. Una escala nominal solo permite clasificar, pero no ordenar o jerarquizar. Únicamente se permiten operaciones de igualdad o desigualdad. Los posibles valores de una escala nominal podrían representarse con letras (A, B, C...). Puede decirse que la clase A es distinta de la B, pero no que sea mayor o menor. Si hay una posible gradación o jerarquización de los valores o categorías (unos son mayores que otros), entonces la escala no es nominal, sino ordinal (v. más adelante).

Las variables cualitativas o categóricas nominales pueden ser:

- *Dicotómicas o binarias* si solo admiten dos categorías: sano/enfermo, expuesto/no expuesto, hombre/mujer, etc.
- *Policotómicas*, con varias categorías: grupo sanguíneo (A/B/0/AB), cierto tipo de tratamiento (A/B/C), estado civil, etc.

1 A pesar de que *género* se haya puesto de moda en algunos ámbitos, en general, suele ser preferible usar simplemente el término *sexo*, en vez de *género*. *Género* es un constructo que busca definir a hombres y mujeres en función de sus características psicosociales, por lo que, en general, sexo (que es un fenómeno biológico, anatómico y fisiológico) parece más apropiado. Si en un estudio realmente se determinó el *género* a partir de la valoración de perfiles psicosociales y del desempeño de *roles*, entonces sí estaría indicado usar *género*, pero se debe definir antes de modo claro y operativo cuál fue la metodología y cuáles los criterios de clasificación utilizados para definir esta variable compleja.

### 2.1.3. Variables cualitativas ordinales

Son aquellas cuyos posibles valores se encuentran jerarquizados y ordenados. El tipo de escala utilizado se denomina *ordinal.* Con estas variables se pueden realizar no solo operaciones de igualdad y desigualdad, sino también operaciones de orden (jerarquizar los diferentes valores). Algunos ejemplos serían el interés en dejar de fumar, el nivel máximo de estudios, el grado de dolor o la intensidad del hábito tabáquico medida en la siguiente escala: nunca fumador/exfumador/fumador activo.

### 2.1.4. Variables cuantitativas

Hablamos de variables cuantitativas cuando los números utilizados para expresarlas equivalen realmente con exactitud a los verdaderos datos. Los datos son realmente numéricos. Hay dos tipos de datos:

- *Discretos:* solo pueden tomar valores numéricos aislados. Sus valores son finitos y coinciden con números enteros. Ejemplos claros son: número de hijos, número de intervenciones previas, número de intentos para dejar de fumar, etc. Permiten operaciones de igualdad-desigualdad y orden *(rango),* pero también operaciones algebraicas.
- *Continuos:* son numéricos y, además, teóricamente, pueden ser iguales a cualquier cantidad intermedia entre dos números enteros. Es decir, idealmente toman valores con un número de decimales que tiende al infinito. Permiten todas las operaciones hasta ahora comentadas y se miden en *escala de razón.* Ejemplos: edad, peso, talla, tensión arterial, etc. Sin embargo, en la práctica, todos los datos que teóricamente son continuos acaban tratándose como discretos, porque los instrumentos de medida son limitados (fig. 2.2).

## 2.2. TRANSFORMACIÓN DE UNA VARIABLE: CATEGORIZACIÓN Y RECODIFICACIÓN

Es posible realizar una transformación de una variable cuantitativa y pasarla a una escala ordinal. Este proceso se suele denominar *categorización* de una variable. Consiste en que una determinada variable que inicialmente poseía características de cuantitativa, porque sus valores estaban expresados en números, se pasa a considerar como cualitativa ordinal, de modo que los valores individuales se colapsan formando unos pocos grupos o categorías que engloban a un rango de los valores anteriores. Por ejemplo, si se quisiera categorizar el hábito tabáquico podrían crearse tres categorías, como muestra la tabla 2.1.

Se obtendrá una nueva variable *«ciggrup»* que solo contiene tres posibles valores (1, 2 o 3). Los códigos de esta nueva variable son «1» para los que fuman menos de 20 cigarrillos al día (cig./día),

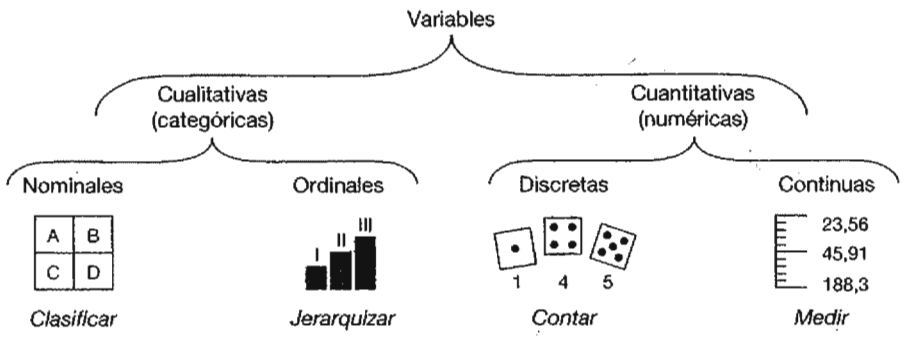



**Figura 2.2** Tipos de variables.

«2» para los que fuman entre 20 y 39 cig./día, y «3» para 40 o más cig./día. Se ha pasado de una escala de razón a una escala ordinal. Este proceso, aunque a veces es práctico, siempre supone perder información. Nunca se debe *recoger* la información en una escala ordinal si se puede hacer en una escala de razón. Siempre es preferible recoger inicialmente las variables con la máxima precisión y detalle posibles (cuanto más *cuantitativas,* mejor) y solo después categorizarlas, si resulta conveniente por el tipo de análisis estadístico que se desee realizar.

### 2.2.1. Transformar y recodificar con STATA

STATA es un programa muy recomendable. Puede adquirirse en: http://www.stata.com/.

Los aspectos básicos de STATA se pueden consultar en: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

Si no se posee experiencia previa con STATA, se recomienda acceder desde la dirección arriba mencionada al primer vídeo: «STATA_0_0: para quien empieza desde cero.»

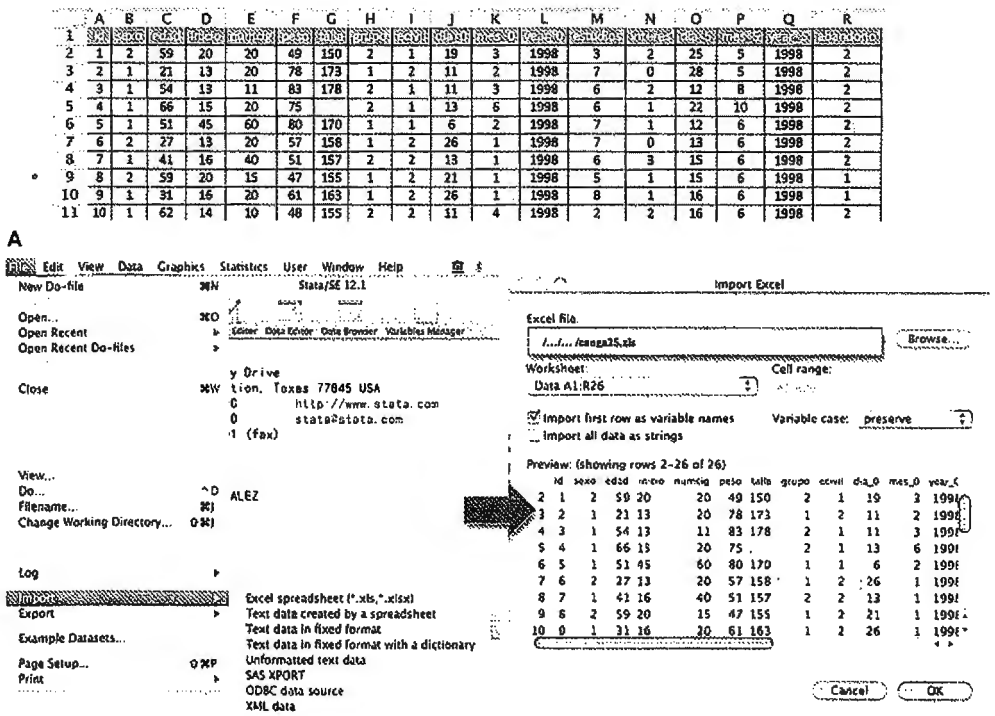
Otra posibilidad es visitar: http://www.stata.com/videos13/stata-interface/.

¿Cómo importar datos de Excel para poder manejarlos en STATA?

Se deben seguir los pasos que indica la figura 2.3.

Resulta muy útil dejar fijado permanentemente para siempre el directorio donde se guardarán los datos. La fijación de directorio permanente de trabajo se hace del modo siguiente:

```
cd C://Documentos/dirname
```



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | |
| 2 | 1 | 2 | 59 | 20 | 20 | 49 | 150 | 2 | 1 | 19 | 3 | 1998 | 3 | 2 | 25 | 5 | 1998 | 2 |
| 3 | 2 | 1 | 21 | 13 | 20 | 78 | 173 | 1 | 2 | 11 | 2 | 1998 | 7 | 0 | 28 | 5 | 1998 | 2 |
| 4 | 3 | 1 | 54 | 13 | 11 | 83 | 178 | 2 | 1 | 11 | 3 | 1998 | 6 | 2 | 12 | 8 | 1998 | 2 |
| 5 | 4 | 1 | 66 | 15 | 20 | 75 | | 2 | 1 | 13 | 6 | 1998 | 6 | 1 | 22 | 10 | 1998 | 2 |
| 6 | 5 | 1 | 51 | 45 | 60 | 80 | 170 | 1 | 1 | 6 | 2 | 1998 | 7 | 1 | 12 | 6 | 1998 | 2 |
| 7 | 6 | 2 | 27 | 13 | 20 | 57 | 158 | 1 | 2 | 26 | 1 | 1998 | 7 | 0 | 13 | 6 | 1998 | 2 |
| 8 | 7 | 1 | 41 | 16 | 40 | 51 | 157 | 2 | 2 | 13 | 1 | 1998 | 6 | 3 | 15 | 6 | 1998 | 2 |
| 9 | 8 | 2 | 59 | 20 | 15 | 47 | 155 | 1 | 2 | 21 | 1 | 1998 | 5 | 1 | 15 | 6 | 1998 | 1 |
| 10 | 9 | 1 | 31 | 16 | 20 | 61 | 163 | 1 | 2 | 26 | 1 | 1998 | 8 | 1 | 16 | 6 | 1998 | 1 |
| 11 | 10 | 1 | 62 | 14 | 10 | 48 | 155 | 2 | 2 | 11 | 4 | 1998 | 2 | 2 | 16 | 6 | 1998 | 2 |



**Figura 2.3** Cómo importar dentro de STATA una base de datos antes creada en Excel. **A.** Base de datos en Excel: la primera fila contiene nombres. No hay etiquetas. Guardar esta base de datos en Excel (nombre: canga25.xls). **B.** Se abre STATA y se selecciona: **File → Import → Excel spreadsheet (*.xls;*.xlsx). C.** Se selecciona **Browse** y se elige el fichero que se quiere abrir. Se debe hacer clic en la opción *Import first row as variable names.*

En ***dirname*** se escribe el nombre de la carpeta (preexistente) que el usuario desee usar[2].
Al final, para guardar los datos, se debe usar la instrucción:

```
save dataname
```

En ***dataname*** se escribe el nombre que el usuario desee utilizar. En futuras ocasiones, si se desea volver a usar estos datos, bastará indicar:

```
use dataname
```

En STATA la base de datos no está a la vista continuamente, como pasa en otros programas. Si se desea verla y no se va a cambiar nada, se escribe:

```
browse
```

Si se desea verla y además cambiar algo, se escribe:

```
edit
```

Ambas opciones (solo ver, o ver y cambiar) están accesibles también en los iconos de la parte superior. La posibilidad de ver los datos sin que se puedan cambiar es interesante para prevenir errores accidentales (p. ej., al tocar el teclado inadvertidamente). Se han subrayado las dos primeras letras de ambas órdenes (**browse** y **edit**) porque basta con escribir esas dos letras para que STATA ejecute la orden. Así se agiliza el trabajo con STATA. Siempre que se presente aquí una instrucción de STATA, aparecerá parcialmente subrayada la parte abreviable. Cuando no se subraya nada, es que la instrucción no es abreviable.

Un modo sencillo de recodificar consiste en usar la orden **recode:**

```
recode numcig 1/19=1 20/39=2 40/max=3, generate(ciggrup)
```

Podría haberse hecho también así:

```
egen cig_gr3=cut(numcig), at(1 19 39 61)
```

Se ha cambiado el nombre de la variable destino para evitar errores. Téngase en cuenta que, al usar **egen** (extensión de generar) seguido de **cut** y **at,** la opción *at* establece tanto los puntos de corte (valor mínimo de cada categoría) como la etiqueta que tendrá cada categoría. Una precaución necesaria al usar **egen... cut, at(...)** es que el último punto de corte debe ser *superior al máximo* valor observado. Si se desea pedir la descripción de una variable (que incluirá los valores máximo y mínimo) se puede usar la instrucción **summarize:**

```
summarize numcig
```

```
. summarize numcig
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| numcig | 25 | 21.24 | 11.23714 | 7 | 60 |

Así se sabe que el máximo es 60; por eso, el último punto de corte que se puso para **egen... cut, at(...)** fue 61. Hubiese dado lo mismo poner 70.

La instrucción **table** devuelve cuántos sujetos hay en cada categoría:

```
table cig_gr3
```

2 Se usará **negrita** para las instrucciones de STATA y **negrita** + *cursiva* cuando son *nombres de variables* o carpetas que el usuario ha inventado. Las líneas con cada instrucción para STATA serán párrafos independientes. Si se desea continuar escribiendo una instrucción en la línea siguiente, se debe añadir un espacio en blanco seguido de tres barras inclinadas: ///
El espacio previo es importante. Usar /// permite seguir escribiendo la instrucción en la línea siguiente.
Algunas instrucciones incluyen una coma, lo que sirve para especificar opciones de esa instrucción que se escribirán después de la coma.

| cig_gr3 | Freq. |
|---|---|
| 1 | 8 |
| 19 | 15 |
| 39 | 2 |

La tabla anterior corresponde a la distribución de frecuencias. Es una tabla con una sola variable. Los números 1, 19 y 39 indican dónde empieza cada categoría, ya que STATA les ha puesto como etiqueta el valor inferior de la propia categoría. Los números 8, 15 y 2 indican cuántos sujetos hay en cada categoría. Para comprobar que STATA recodificó correctamente debe escribirse:

```
tabulate numcig ciggrup
```

La pantalla de resultados presentará entonces la siguiente salida:

```
. tabulate numcig ciggrup
```

| | RECODE of numcig (numcig) | | | |
|---|---|---|---|---|
| numcig | 1 | 2 | 3 | Total |
| 7 | 1 | 0 | 0 | 1 |
| 8 | 1 | 0 | 0 | 1 |
| 10 | 2 | 0 | 0 | 2 |
| 11 | 1 | 0 | 0 | 1 |
| 15 | 3 | 0 | 0 | 3 |
| 20 | 0 | 11 | 0 | 11 |
| 30 | 0 | 4 | 0 | 4 |
| 40 | 0 | 0 | 1 | 1 |
| 60 | 0 | 0 | 1 | 1 |
| Total | 8 | 15 | 2 | 25 |

Ahora no se trata de una, sino de dos variables. Este modo de presentar los datos se llama *tabla de contingencia,* donde las columnas corresponden a una variable y las filas a otra. Dentro de la tabla, las casillas, definidas por su fila y su columna, contendrán el número de observaciones, que presentan el valor de la fila en una variable y el de la columna en la otra. En la jerga estadística, esta acción se refiere como *cruzar* dos variables. Por ejemplo, hay 11 personas que fumaban 20 cig./día y están en la categoría 2 de ***ciggrup.*** También se puede obtener esta tabla usando los menús de STATA (fig. 2.4):

**Data → Create or change data → Other variable-transformation commands → Recode categorical variable**

Se acaba por preferir las órdenes a los menús, especialmente en STATA. Al trabajar con menús, a veces es conveniente finalizarlos pulsando *Submit* en vez de *OK;* así no se cierra la última ventana y se pueden hacer cambios sin tener que repetir todo el recorrido de cuadros de diálogo. Si se conoce el nombre de una orden y se desea abrir directamente el cuadro de diálogo *(dialog box),* basta con escribir **db** delante del nombre de la orden. Por ejemplo, para abrir la ventana del menú de **summarize,** se escribirá:

```
db summarize
```

Una gran ventaja de STATA reside en las ayudas. Para obtener ayudas basta con escribir **help** delante de cualquier orden o bien abrir el menú *Help.* La primera opción al abrir este menú es

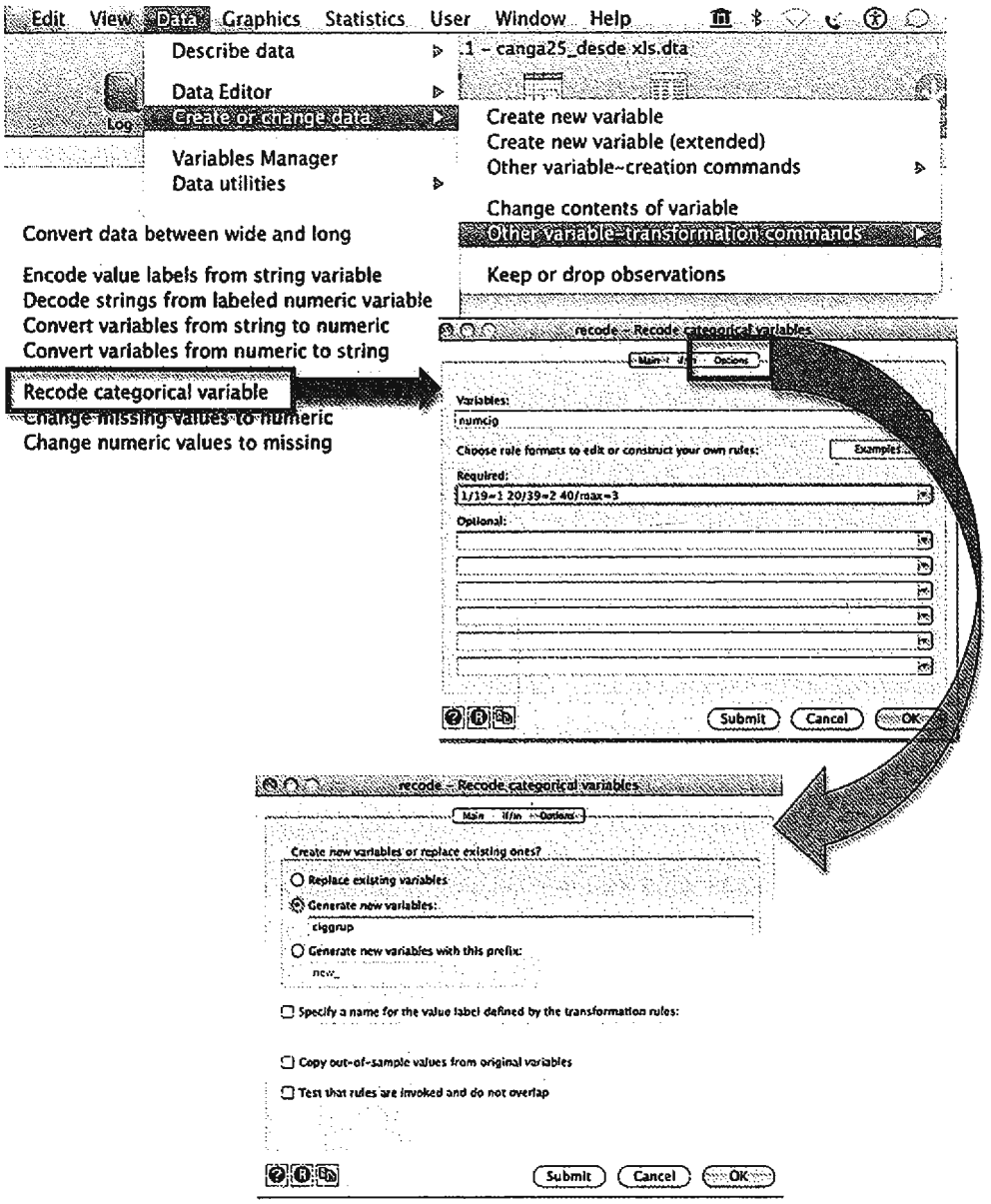



**Figura 2.4** Cómo recodificar con STATA. 1. Seleccionar los siguientes menús en este orden: **Data → Create or…→ Other variable-transf… → Recode categorical…** 2. Seleccionar la variable origen de la que se desea partir para hacer la recodificación ***(numcig)***. 3. Escribir las reglas de recodificación en la ventana central del menú *Main*. 4. Seleccionar el menú *Options* y darle nombre a la variable destino ***(ciggrup)***. 5. Hacer clic en *Submit*.

*Search*, que permite una búsqueda libre. La segunda opción es *PDF Documentation*, que conduce a una amplia serie de manuales detallados de STATA en formato PDF que están contenidos en todo ordenador que tenga instalado STATA:

**Help → PDF Documentation**

Se recomienda usar con frecuencia todas estas ayudas.

Otra ventaja de STATA es que, cuando se ejecuta una orden usando menús, siempre aparece después la misma orden escrita como instrucción en la ventana de resultados. Esto facilita aprender las instrucciones de memoria o copiarlas para usos posteriores. La mejor forma de guardar las instrucciones de STATA para futuros usos es abrir un *Do-file,* que es muy parecido a un procesador de textos. Se puede crear un *Do-file* con el icono correspondiente o desde la barra de menús:

**File → New Do-file**

Una vez copiadas allí las instrucciones pertinentes (cada instrucción en una línea), se pueden ejecutar marcándolas con el ratón y pulsando el icono *Do,* o bien con la combinación *Ctrl* + *D* (en Mac: *Comand* + *Shift* + *D*).

Se recomienda ver el siguiente vídeo: «STATA_0_0_do files: lo básico de texros con órdenes».

### 2.2.2. ¿Cómo recategorizar en SPSS para Windows?

SPSS es un programa con licencias más caras y más transitorias que las de STATA, pero se ha usado mucho, especialmente en los ámbitos biomédicos. Puede adquirirse en: http://www-01.ibm.com/software/es/analytics/spss/.

En SPSS se ven los datos continuamente y pueden modificarse de forma directa, como sucede en Excel. El usuario de SPSS tiende a trabajar con menús, lo cual a la larga no es muy recomendable. SPSS contiene una opción —como suelen tenerla otros programas— de *Archivo → Abrir datos.* Esta opción permitirá abrir archivos escritos en Excel, simplemente seleccionando .xls en la parte inferior dentro de la opción «Archivos de tipo». SPSS también puede abrir así bases de datos previamente escritas en STATA, y permite que un archivo que se ha trabajado en SPSS se acabe archivando en formato STATA dentro de la opción de SPSS llamada *Guardar como* (similar a cualquier programa).

Si se desea recodificar en SPSS usando menús, hay que elegir la opción *Transformar* en la barra superior y luego *Recodificar en distintas variables,* como indica la figura 2.5.

A continuación aparecerá otro menú en el que se pregunta qué variable se desea recodificar. Se selecciona con el ratón *numcig* y luego el botón en forma de flecha que está en medio. A continuación se escribe el nombre que se quiera dar a la variable de resultado o destino *(ciggrup)* en la casilla de la derecha que está en blanco y luego se pulsa en *Cambiar.* Después, todo consiste en abrir el cuadro de diálogo *Valores antiguos y nuevos...* e ir dando las órdenes pertinentes para cada nueva categoría seguida de *Añadir.* Se finaliza con *Continuar* y luego *Aceptar.* Si se opta por *Pegar* en vez de aceptar, se abrirá una ventana de sintaxis que es análoga a la del *Do-file* de STATA. El contenido es:

```
RECODE numcig
(Lowest thru 19=1)
(20 thru 39=2)
(40 thru Highest=3)
INTO ciggrup.
EXECUTE.
```

En SPSS cada orden puede ocupar varias líneas, pero debe acabar siempre con un punto. SPSS no es sensible a mayúsculas y minúsculas, es indiferente usar unas u otras; en cambio, STATA las considera letras distintas (se recomienda usar solo minúsculas en STATA). Para ejecutar una orden en SPSS, se debe marcar la orden con el ratón y oprimir después *Ctrl* + *R.*

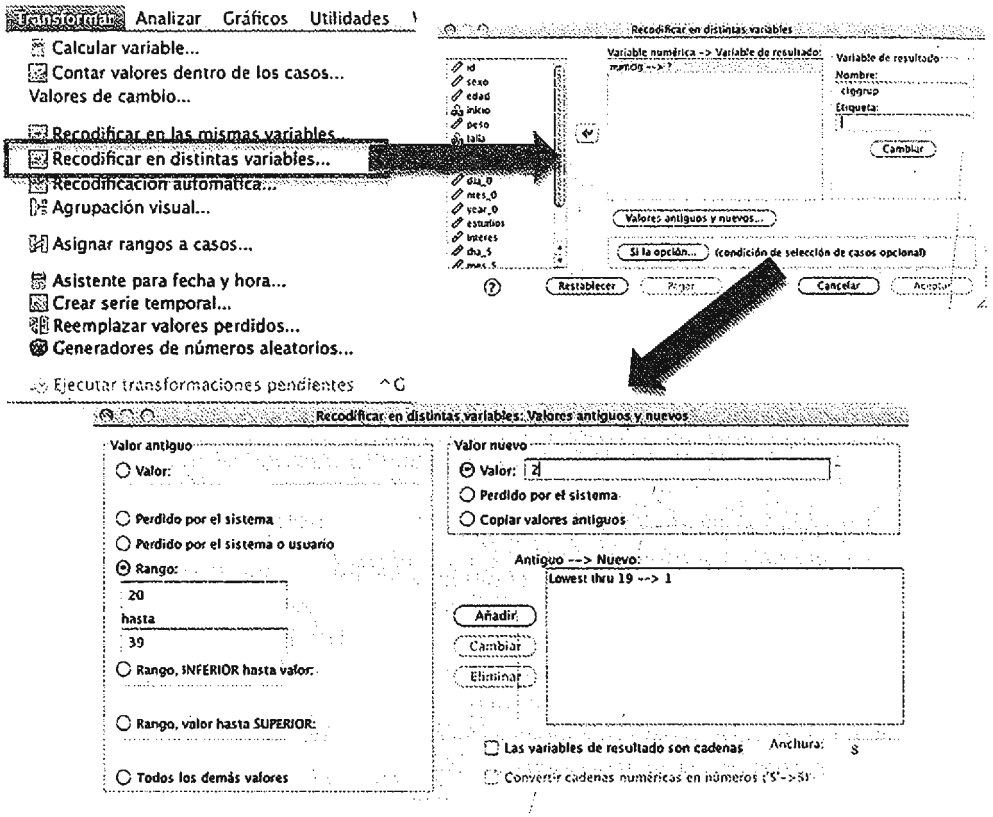



**Figura 2.5** Cómo recodificar con SPSS.

Para obtener una tabla de contingencia en SPSS que cruce la variable antigua y la nueva, se debe seguir la ruta siguiente en los cuadros de diálogo:

**Analizar → Estadísticos descriptivos... → Tablas de contingencia**

Se abre una nueva ventana y se elige así la variable que se desea que ocupe las filas y la que ocupará las columnas. Esto mismo con sintaxis (usando *Pegar* en vez de *Aceptar*) se hará del modo siguiente:

```
CROSSTABS
  /TABLES=numcig BY ciggrup
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
```

Aunque se obtienen idénticos resultados al escribir simple y directamente lo siguiente en la ventana de sintaxis de SPSS:

```
CRO numcig BY ciggrup.
```

Se obtendría entonces la tabla 2.2.

***Tabla 2.2** Tabla de contingencia en SPSS que cruza una variable cuantitativa (número de cigarrillos/día) con su recodificación como variable ordinal (categorización)*

TABLA DE CONTINGENCIA NUMCIG * RECODE OF NUMCIG (NUMCIG)

| | | RECUENTO | | | |
|---|---|---|---|---|---|
| | | RECODE OF NUMCIG (NUMCIG) | | | TOTAL |
| | | 1 | 2 | 3 | |
| numcig | 7 | 1 | 0 | 0 | 1 |
| | 8 | 1 | 0 | 0 | 1 |
| | 10 | 2 | 0 | 0 | 2 |
| | 11 | 1 | 0 | 0 | 1 |
| | 15 | 3 | 0 | 0 | 3 |
| | 20 | 0 | 11 | 0 | 11 |
| | 30 | 0 | 4 | 0 | 4 |
| | 40 | 0 | 0 | 1 | 1 |
| | 60 | 0 | 0 | 1 | 1 |
| Total | | 8 | 15 | 2 | 25 |

### 2.2.3. Leer ficheros y recodificar variables con R/Splus

A diferencia de STATA y SPSS, R es un programa *gratuito*. Está en constante evolución[3]. Hay varios modos de leer una base de datos en R/Splus. Por ejemplo, se sugiere dar los siguientes pasos:

1. Introducir los datos en una hoja de cálculo de Excel: cada columna corresponderá a una variable. La primera fila de cada columna debe llevar el nombre de la variable. Los decimales deben estar separados por puntos y no por comas. Tampoco se pueden usar tildes (acentos) ni la letra ñ ni otros signos: solo se escribirán números y nombres breves de variables. Además, hay que asegurarse de que no haya casillas en blanco.
2. Guardar los datos desde Excel como texto MS-DOS (formato .txt, sin hacer caso de posibles avisos de Excel). También se puede descargar la base de datos directamente desde esta dirección: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

   Quedará guardado un archivo (canga25R.txt) en el disco duro, por ejemplo en la siguiente dirección:

   ```
   c://r/canga25R.txt
   ```

1. Abrir R/Splus y usar el menú para definir como directorio de trabajo aquel en el que esté contenido el fichero que se quiere abrir. Esto puede hacerse desde el menú como:

   **Misc → Change working directory**

   o bien usando *Ctrl + D*.

   Resulta más útil dejar fijado permanentemente el directorio donde se guardarán los datos. La fijación de directorio permanente de trabajo puede hacerse desde el menú general de R:

   **R → Preferences → General (startup) → Initial working directory**

2. Dar la siguiente orden:

   ```
   > d <- read.table("canga25R.txt", header=T)
   ```

   Si después se escribe simplemente

   ```
   > d
   ```

   el programa devolverá los resultados que aparecen en la figura 2.6.

3 Se puede descargar gratuitamente (y legalmente) R, que es casi equivalente a Splus, en las siguientes direcciones: http://www.r-project.org y http://cran.es.r-project.org.

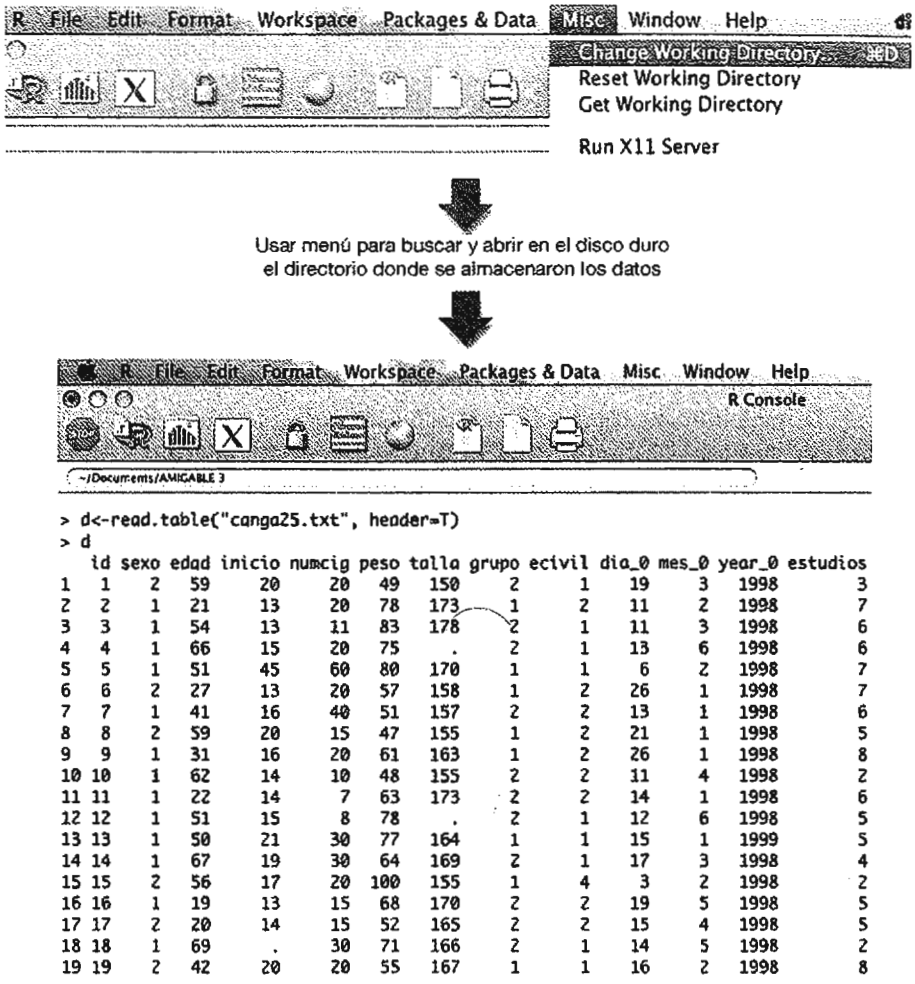



**Figura 2.6** Apertura de una base de datos en R.

Para trabajar con variables de esa base de datos, a la que se ha llamado en su conjunto *d*, se podrá utilizar cada variable añadiendo delante el nombre de la base de datos *(d)* seguido del símbolo $. Así se pide, por ejemplo, la media de la edad:

```
> mean(d$edad)
```

y el programa devolverá:

```
[1] 50.92
```

Otra posibilidad consiste en dar un paso previo *(attach)* que indique a R que se desea trabajar solo con esa base de datos. Ya no harán falta dólares:

```
> attach(d)
> mean(edad)
[1] 50.92
```

R/Splus, al igual que STATA, interpreta de modo distinto las mayúsculas y las minúsculas. Si se diese la orden **mean(*d$EDAD*)**, el programa daría error. En R se recodificarían así los cigarrillos fumados al día, con los mismos criterios usados antes en STATA y SPSS:

```
> ciggrup<-cut(numcig, c(1,19,39,60))
> table(ciggrup)
ciggrup
(1,19] (19,39] (39,60]
   8      15       2
> table(numcig,ciggrup)
```

La orden **table** sirve para pedir un recuento de los valores de cada categoría. La segunda vez que se usa [**table(numcig,ciggrup)**] proporcionará la tabla de contingencia. La primera categoría va desde 1 a 19 cigarrillos, ya que el corchete «]» significa que incluye el valor 19. Esta primera categoría contiene 8 personas. La segunda va desde >19, pues eso es lo que indica el paréntesis «)» y llega hasta 39 inclusive. Contiene 15 personas. La tercera va desde >39 hasta el máximo (cig./día = 60). Para obtener el valor inferior o superior de una variable se debe pedir:

```
> min(d$numcig)
[1] 7
> max(d$numcig)
[1] 60
```

- Tampoco en R/Splus la base de datos está a la vista. Si se desea verla, se escribe:

```
> edit(d)
```

Para poner etiquetas a los valores de una variable categórica en R se debe indicar primero que se trata de un **factor.** De lo contrario, R la tratará como cuantitativa. Después se puede dar nombre a cada categoría escribiendo las etiquetas según el orden numérico de la codificación utilizada (1 = varón; 2 = mujer). Se haría así:

```
> sexo<-factor(sexo,labels=c("varón","mujer"))
```

## 2.3. CONSEJOS PRÁCTICOS SOBRE CATEGORIZACIÓN DE VARIABLES CUANTITATIVAS

### 2.3.1. Consejos prácticos con STATA

Con frecuencia se necesita establecer categorías con variables cuantitativas. Con mucha frecuencia se acaban creando cinco categorías ordenadas de menor a mayor con igual número de sujetos en cada una, que se llaman quintiles. A veces se crean cuatro categorías (cuartiles). Son, por tanto, variables cualitativas *ordinales.* Esto se puede hacer así en STATA:

```
xtile peso5=peso, nq(5)
tabstat peso, by(peso5) stat(n min max)
```

Con lo que se obtiene:

```
Summary for variables: peso
     by categories of: peso5 (5 quantiles of peso)
```

| peso5 | N | min | max |
|---|---|---|---|
| 1 | 5 | 47 | 52 |
| 2 | 5 | 55 | 64 |
| 3 | 5 | 67 | 71 |
| 4 | 5 | 75 | 78 |
| 5 | 5 | 80 | 100 |
| Total | 25 | 47 | 100 |

La primera orden **xtile** sirve para crear estas variables categóricas ordinales, llamadas en general *cuantiles*. Tendrán el nombre que se desee (***peso5*** en este ejemplo) y se derivan de una variable original cuantitativa (***peso***). Tras la coma de las opciones, y precedido de la opción *nq*, se indicará cuántas categorías de igual tamaño se desean (cinco en el ejemplo; cuatro si fuesen cuartiles).

La segunda orden (**tabstat**) es útil y versátil. Se sugiere ver **help tabstat**, especialmente con los ejemplos que vienen al final de esa ayuda[4]. En el ejemplo arriba mencionado sirve para describir el ***peso*** (variable cuantitativa) dentro de cada quintil (***peso5***, variable ordinal). Se han pedido tres índices estadísticos, el número de sujetos (n), el valor mínimo (**min**) y el máximo (**max**). Se comprueba que se han construido cinco categorías, cada una de las cuales contiene exactamente cinco personas. El primer quintil va de 47 a 52 años; el segundo, de 55 a 64, etc. Los cuantiles se basan en medidas de posición. Se verán con mayor detalle más adelante.

Los quintiles tienen una gran ventaja y es que se consigue repartir el total de los sujetos en grupos de igual tamaño, lo cual minimiza el riesgo de acabar teniendo grupos muy pequeños y, por tanto, poco informativos. Por otra parte, cinco grupos son suficientes para hacerse una idea de si hay o no una tendencia en los datos. Por ejemplo, se comparó el volumen cerebral en cinco categorías ordenadas de consumo de alcohol y se vio una clara tendencia dosis-respuesta: cuanto más alcohol, menor volumen cerebral (4). Pero no se usaron quintiles, sino categorías fácilmente comparables con otros estudios. Una cautela que debe tenerse en cuenta antes de usar quintiles es que, a veces, *no* interesa utilizarlos, ya que: a) los puntos de corte podrían variar entre nuestro estudio y el de otros investigadores, y esto haría poco comparables los resultados, y b) a veces los cuantiles no expresan las categorías científicamente relevantes, porque podría haber un efecto umbral o un efecto saturación. Debe combinarse la ventaja de crear grupos iguales propia de los quintiles con el conocimiento experto del tema de investigación para seleccionar los puntos de corte que puedan ser más claves *per se* en el problema concreto que se estudia (5).

Por ejemplo, podría ser preferible obtener cuatro grupos de peso con unos puntos de corte más comparables y que se recuerden mejor porque sean múltiplos de 10. Se haría así en STATA:

```
egen peso_x10=cut(peso), at(47 60 70 80 101)
tabstat peso, by(peso_x10) stat(n min max)
```

4 Muchas de las ayudas de STATA recurren a una base de datos llamada ***auto.dta*** que viene instalada como parte del programa. Esta base de datos se puede descargar automáticamente escribiendo **sysuse** ***auto.dta***. Una vez hecho esto, se pueden ir probando los ejemplos que vienen al final de las ayudas para cada instrucción. Esto facilitará entender mejor cómo funciona cada orden.

| peso_x10 | N | min | max |
|---|---|---|---|
| 47 | 7 | 47 | 57 |
| 60 | 7 | 61 | 68 |
| 70 | 6 | 71 | 78 |
| 80 | 5 | 80 | 100 |
| Total | 25 | 47 | 100 |

Obsérvese el detalle de usar 101 en vez de 100 como último límite (el máximo peso observado era 100). Se obtienen así cuatro categorías con puntos de corte <60, 60 a <70, 70 a <80 y ≥80. Podría haberse hecho también así:

```
generate pesox10=1 if peso<60
(18 missing values generated)
replace pesox10=2 if peso>=60 & peso<70
(7 real changes made)
replace pesox10=3 if peso>=70 & peso<80
(6 real changes made)
replace pesox10=4 if peso>=80 & peso<.
(5 real changes made)
```

Al usar instrucciones lógicas para definir condiciones, tanto en STATA como en SPSS el signo & corresponde a la intersección («AND» = que se cumplan *ambas* condiciones). En cambio, el signo | corresponde a la unión («OR»: basta con que se cumpla algunas de las condiciones). El punto (.) que se ha usado en la última orden se refiere a valores *missing* (datos que faltan). Es importante tener en cuenta que STATA considera un valor faltante (.) como si fuese *superior* a todos los valores observados. Por lo tanto, cuando se dé una orden que quiera referirse a todas las observaciones superiores a un valor, se debe añadir lo que se ha hecho aquí (& *varname*< .) al final.

### 2.3.2. Ejecución con SPSS

En SPSS se pueden crear quintiles con la siguiente expresión:

```
RANK VAR=peso /NTILES(5).
```

La nueva variable que SPSS creará se llama ***npeso*** y tendrá los valores (etiquetas) 1 a 5, que corresponden a los quintiles. Para cambiarle el nombre:

```
RENAME VAR npeso=peso5.
```

Si se desea dar puntos de corte en múltiplos exactos de 10, podría usarse **RECODE**, o se podría hacer también así:

```
COMPUTE pesox10=peso<60.
IF peso>=60 & peso<70 pesox10=2.
IF peso>=70 & peso<80 pesox10=3.
IF peso>=80 pesox10=4.
EXE.
```

### 2.3.3. Ejecución con R/Splus

En R, una vez dado el paso **attach(d)**, se pueden crear quintiles con la siguiente secuencia de órdenes:

```
> p5<-quantile(peso,probs=c(0,20,40,60,80,100)/100)
> peso5<-cut(peso,breaks=p5,labels=c(seq(1:5)),include.lowest=T)
> table(peso5)
peso5
1 2 3 4 5
5 5 5 5 5
```

Si se prefieren puntos de corte exactos en vez de quintiles, se hará así:

```
> pesox10<-cut(peso,breaks=c(0,60,70,80,101),
+ labels=c("<60","60-<70","70-80","80+"),include.lowest=T)
```

En el programa R, cuando se acaba una línea, pero no se ha terminado de escribir la instrucción, aparece automáticamente el signo + en la siguiente línea. Esto le indica al usuario que debe completar su orden.

## 2.4. REPRESENTACIONES GRÁFICAS

Una imagen vale más que mil palabras. Las gráficas son importantes en epidemiología y en estadística. Se pueden usar con un fin descriptivo para transmitir una información muy rica con un solo golpe de vista. Dominar los métodos gráficos capacita para proporcionar información de manera condensada e inteligible. Una gráfica pertinente y bien pensada logra una proximidad e inmediatez únicas. Se describe así un gran volumen de datos y se evita al lector la molesta sensación de hallarse ante una desagradable masa informe de números. *Gráficas adecuadas y simples* mejoran la visión de conjunto y previenen errores. Los consumidores de información biomédica andan escasos de tiempo y valoran la brevedad que se puede lograr con una buena gráfica.

### 2.4.1. Gráfico de sectores

Es un gráfico sencillo. En inglés se conoce como *pie chart*. En español se le llama también *tarta* o *pastel*. Tiene pocas indicaciones: solo para variables *categóricas nominales*. Como dicta el sentido común, el área asignada a cada categoría será directamente proporcional a su frecuencia. Se asigna a cada categoría el porcentaje de los 360 grados de circunferencia que corresponden a la frecuencia relativa de esa categoría. En el ejemplo *(canga25)* hay 19 varones que son el 76% del total (n = 25; 19/25 = 0,76). Se asignarían 0,76 × 360 = 273,6 grados a la categoría «Varón» y el resto 360 − 273,6 = 86,4 a la categoría «Mujer». Simple regla de tres. No habría que hacer cálculos, ya los hace automáticamente el ordenador. A STATA se le pide así:

```
graph pie, over(sexo)
```

y se obtiene la figura 2.7.

Para mayores detalles puede consultarse:

```
help (graph pie)
```

En SPSS se obtendrá este gráfico de sectores con:

```
GRAPH /PIE=COUNT BY sexo.
```

El gráfico de sectores muchas veces *no* es el más recomendable. Solo está indicado si la escala es estrictamente *nominal*.

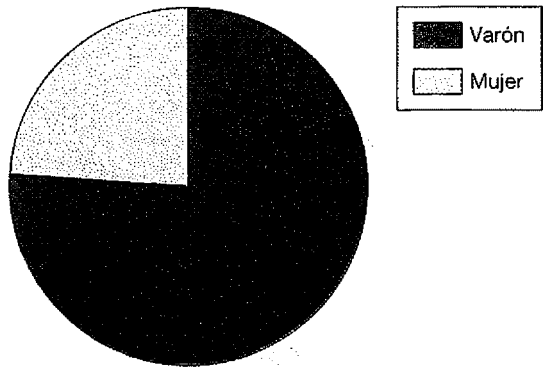



**Figura 2.7** Gráfico de sectores con STATA: **graph pie, over*(sexo)*.** En STATA, por omisión, el gráfico empieza en las 12 de un reloj y sigue el sentido de las agujas del reloj según el orden de numeración de las categorías (primero 1 = varón, luego 2 = mujer). Si se desea cambiarlo y que vaya *en contra* de las agujas del reloj, se usará la opción **noclockwise** al final, tras la coma de la opción.

Puede resultar útil recurrir a Excel para realizar con rapidez este u otros gráficos de una sola variable. Mientras que en STATA o SPSS cada persona suele ocupar una fila, en Excel basta con escribir en una columna solo dos números, el de varones y el de mujeres, como se representa en la figura 2.8. Es decir, es suficiente con poner los totales de cada categoría.

Una vez que están así introducidos los datos en Excel, se selecciona con el ratón el bloque en que está la información introducida, se elige *Gráficos* en el menú superior y se selecciona el gráfico que se desee. *Salvo bien pensadas excepciones, se deben evitar los efectos 3D en todos los gráficos*, ya que tales efectos tridimensionales pueden dificultar que se cuantifique visualmente la información con exactitud (6).

### 2.4.2. Diagrama de barras

— ¿Se podría haber utilizado el gráfico de sectores para describir el número de cigarrillos/día en tres categorías (p. ej., <20, 20-39, ≥40)?

— *De ningún modo.*

Usar un gráfico de sectores para una variable *ordinal* es *erróneo*, pues se pierde el orden de las categorías. Hay una jerarquía en esta variable, que el gráfico de sectores no puede recoger. Por tanto, está indicado otro gráfico más específico que permita ordenar las categorías de menos a más. Es el gráfico indicado para variables ordinales y se llama *diagrama de barras*. Está formado por barras o rectángulos cuya altura es proporcional al número de observaciones en cada categoría. Los rectángulos están separados entre sí y no hay ninguna agrupación de categorías. Cada valor diferente es representado por una barra distinta. Solo se consideran los valores que realmente se hayan observado en la muestra; por lo tanto, el eje horizontal no tiene por qué tener valores consecutivos.

Antes de construir en STATA un diagrama de barras, conviene etiquetar las variables y sus valores con la siguiente secuencia de órdenes:

```
label var ciggrup "cig/día"
label define ciggrup 1 "<20" 2 "20-39" 3 "40+"
lab val ciggrup ciggrup
```

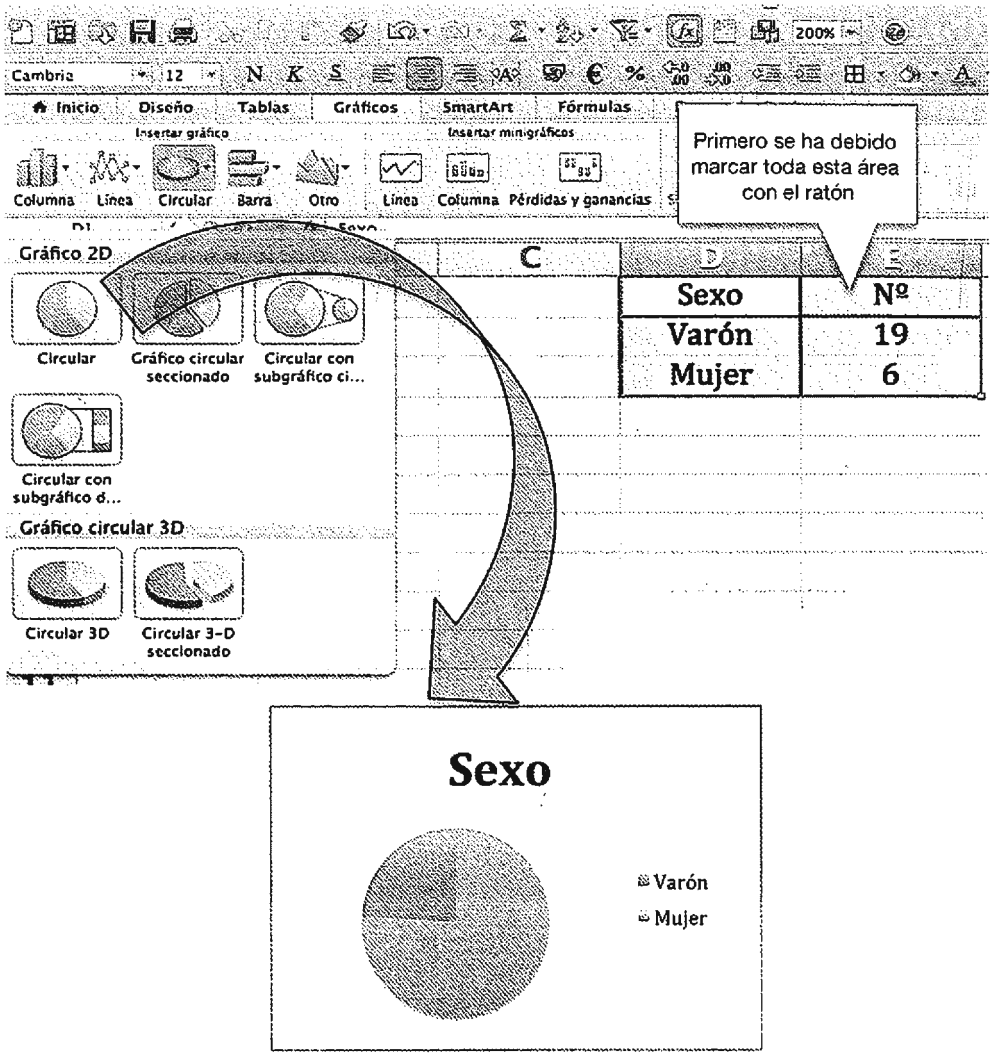



**Figura 2.8** Gráfico de sectores con Excel. El gráfico de sectores solo debe usarse si la escala es estrictamente *nominal*. Como regla general, se recomienda evitar los efectos 3D en los gráficos.

Se construirá después el diagrama de barras en STATA con:

```
histogram ciggrup, discrete frequency gap(30) addlabel ///
xlabel(1(1)3, valuelabel)
```

Se obtendrá la figura 2.9.

En la dirección http://www.unav.es/departamento/preventiva/recursos_bioestadistica se pueden encontrar los dos vídeos siguientes, que amplían los conceptos relacionados con etiquetas y con diagrama de barras en STATA:

- STATA_0_3: etiquetas.
- STATA_2_3: BARRAS.

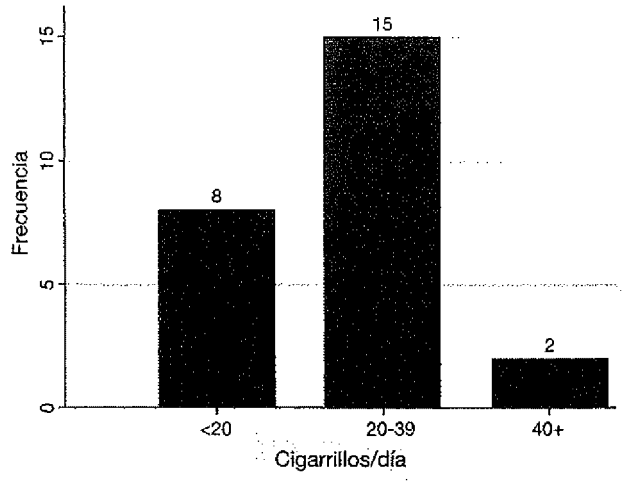



**Figura 2.9** Gráfico de barras con STATA. **histogram *ciggrup*, discrete frequency /// gap(30) addlabel xlabel(1(1)3, valuelabel)** El gráfico de barras es el indicado para variables ordinales.

En la tabla 2.3 se resume el modo de etiquetar valores y variables en STATA y SPSS. En la tabla 2.4 se sintetizan opciones útiles para gráficos en STATA.

Un diagrama de barras sería manipulable tendenciosamente para conseguir una impresión poco objetiva en el observador. Se debe tener cuidado con esto. Por ejemplo, sucesivos balances anuales con beneficios levemente crecientes pueden ser presentados haciendo casi coincidir el primero con la base de la gráfica; así, el crecimiento dará la impresión de ser mayor de lo que es. En realidad se está omitiendo gran parte de la gráfica, que debería empezar en el 0 y daría entonces su verdadera

*Tabla 2.3 Procedimientos para poner etiquetas en STATA, SPSS y R*

| | | |
|---|---|---|
| **STATA** | | |
| **label variable** | Etiquetar una variable | **la var ecivil "Estado civil"** |
| **label define** | Crear etiquetas para los valores o categorías de variables cualitativas (y que así luego estén disponibles) | **la de *ec* 1 *casado* 2 /// *soltero* 3 /// *"separ. o divorc"* /// 4 *viudo*** |
| **label value** | Asignar a cada categoría de una variable etiquetas previamente definidas | **la val ecivil ec** |
| **SPSS** | | |
| **VARIABLES LABEL** | Etiquetar una variable | **VAR LAB ecivil "Estado civil".** |
| **VALUES LABEL** | Asignar etiquetas a cada valor o categoría (sin necesidad de definirlas previamente) | **VAL LAB *ecivil*** *1 "Casado"* *2 "Soltero"* *3 "Separado o divorciado"* *4 "Viudo"* |
| **R/Splus** | | |
| **factor(varname, labels= c("...", "..."))** | Informar de que una variable es un factor (es decir, CATEGÓRICA) y a la vez asignar etiquetas (solo para los valores *realmente* existentes) según su orden numérico | **ecivil < -factor(ecivil, + labels = c("casado", + "soltero" +,"viudo"))** |

*Tabla 2.4 Opciones comunes de STATA para configurar gráficos diversos*

| | ACCIÓN | EJEMPLO |
|---|---|---|
| **Áreas** | | |
| **color** | Color de relleno | **graph pie, over(*sexo*) /// pie(1, col(blue)) /// pie(2, col(red))** |
| **fcolor** | Color de relleno | **hist *edad*, freq addl /// start(15) w(10) /// fc (purple) xlab(20(10)70)** |
| **fintensity** | Intensidad del color (debe añadirse inten50, etc., desde 0 a 100) | **hist *edad*, freq addl /// start(15) w(10) /// fc (green) fin(inten30) /// lc(black) xlab(20(10)70)** |
| **Líneas** | | |
| **lpattern** | Tipo de línea: continua, de puntos... | **tw (scatter *peso talla*) /// (lfit *peso talla*, lp(dash))** |
| **lcolor** | Color de línea | **tw (sc *peso talla*) /// (lfit *peso talla*, lc(blue))** |
| **lwidth** | Anchura de línea | **tw (sc *peso talla*) /// (lfit *peso talla*, lw(thick))** |
| **Etiquetas** | | |
| **mlabel** | Selecciona la variable que se usará como etiqueta | **scatter *peso talla* if /// *peso* > 75, mlabel(sexo)** |
| **addlabels** | Añade la etiqueta con su altura a las barras | **hist *edad*, freq addl /// start(15) w(10)** |
| **addlabopts** | Necesario para poder cambiar las opciones de etiquetas | |
| **mlabsize** | Tamaño de la etiqueta | **hist *edad*, freq addl /// addlabop(mlabs(large))** |
| **mlabcolor** | Color de la etiqueta | **hist *edad*, freq addl /// addlabop(mlabc(olive))** |
| **mlabposition** | Posición: la hora correspondiente a la esfera de un reloj | **hist *ecivil*, freq discrete gap(40) addl /// addlabop(mlabp(9))** |
| **Marcadores** | | |
| **msymbol** | Tipo de marcador | **sc *peso talla*, m(diamond)** |
| **mcolor** | Color de marcador | **sc *peso talla*, mc(brown)** |
| **msize** | Tamaño | **sc *peso talla*, msiz(tiny)** |
| Se puede consultar en STATA: **help colorstyle** **help linestyle** **help symbolstyle** | | |

magnitud a las diferencias observables. Se debe ser cauto al analizar estas gráficas, fijándose, sobre todo, en la escala de medida y en los puntos de origen.

En SPSS se construye el diagrama de barras siguiendo los pasos que indica la figura 2.10.

### 2.4.3. Histogramas

A diferencia del diagrama de barras, en el histograma los rectángulos están ahora juntos y, como la variable se considera continua (admite teóricamente todos los decimales), el punto medio del intervalo es el que da título a cada rectángulo, y ese punto medio admite, teóricamente, decimales. En el eje *x* se representan todas las categorías, también las que no existen (que quedarán vacías). El histograma contempla todos los posibles valores dentro de un rango (aunque no ocurran en la muestra) y los valores próximos se reúnen en una misma categoría. El diagrama de barras está pensado, sobre todo, para variables ordinales, mientras que el histograma está concebido para variables que siguen una escala numérica de razón (cuantitativas, idealmente continuas).

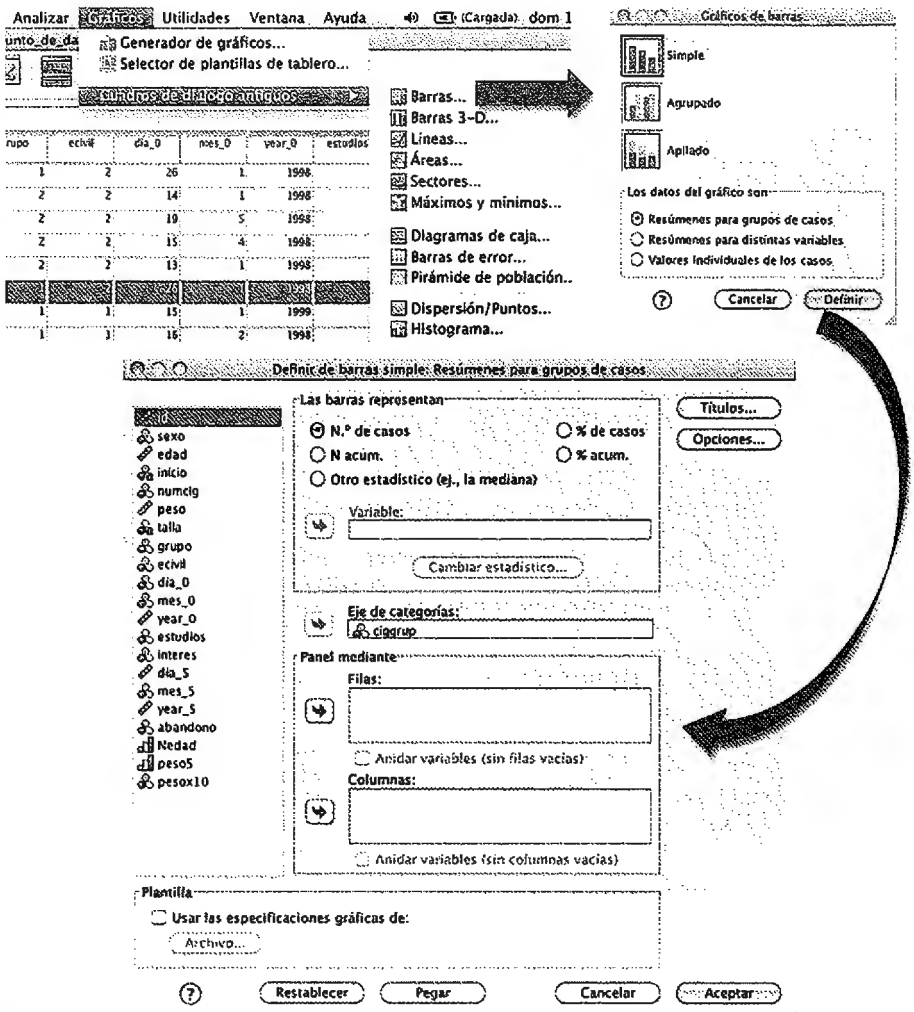



**Figura 2.10** Gráfico de barras con SPSS.

Hay que pensar bien dónde se empieza un histograma y cuáles se desea que aparezcan rotulados como puntos medios de cada intervalo. Si se decide escoger, por ejemplo, como primer punto medio de intervalo el valor 10,0 y la anchura del intervalo también en 10, se calculan sus extremos sumando y restando 10/2 = 5,0 al punto medio. Así, el primer intervalo (punto medio = 10,0) abarcará desde 5 hasta *exactamente antes de 15*. Por eso, si se desea que el primer rótulo que aparezca en el eje horizontal sea 10,0 y que las categorías vayan de 10 en 10 unidades, el valor más bajo que se incluirá será 5 (punto de comienzo). Un error frecuente es indicar 10 en vez de 5 como punto de comienzo porque se desea que el primer rectángulo tenga el 10,0 como título. El ordenador no hace lo que el usuario desea, sino lo que el usuario le pide, y entonces aparecerá 15,0 como punto medio del primer intervalo. Otro error es que, si se define 5 como punto de comienzo y se fija una anchura de 10 para cada rectángulo, el usuario podría esperar que quien tiene un valor exactamente de 15,0 esté en el primer intervalo (que va de 5 a 15). Pero no será así, esa persona caerá en el segundo intervalo, que es el que empieza exactamente en 15,0.

### 2.4.3.1. Histograma con STATA

Para explicar un histograma se volverá a usar el ejemplo anterior con la variable original, sin agrupar *(**numcig**)*, del número de cigarrillos/día. Se obtendría el histograma representado en la figura 2.11.

En general, se debe indicar lo siguiente a STATA cuando se pide un histograma:

- Valor en el que comenzará el primer rectángulo (**start**) o punto de comienzo.
- Anchura que se desea que tenga cada rectángulo (**width**).
- Si se desea que el eje de la *y* indique número de casos (**frequency**) o porcentajes (**percent**) en cada rectángulo.
- Los rótulos que se desea que aparezcan en los ejes. Por ejemplo: **xtitle(10(10)70)** significa empezar en 10 e ir de 10 en 10 (20, 30...) hasta 70.
- Si se desea sobrescribir dentro de la gráfica el número exacto de datos en cada rectángulo, se deberá añadir (**addlabel**).

Además, opcionalmente, se pueden especificar los colores de las líneas y del relleno, los tamaños de la letra, la posición, los títulos, etc.

Un modo simple de obtener un histograma similar al de la figura 2.11, pero esta vez basado en porcentajes en vez de número de sujetos, sería:

```
histogram numcig, width(10) start(5) percent ///
addlabel xlabel(10(10)70) ylabel(0(20)100)
```

Se puede obtener más información con los menús desplegables:

**Graphics → Histogram**

Se recomienda también ver el breve vídeo «STATA_2_1: HISTOGRAMAS» en: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

En STATA, una vez que se ha obtenido un gráfico, se pueden cambiar los colores, tamaños de fuente, etc. Esta acción se realiza con el gestor de gráficos, pulsando el icono correspondiente

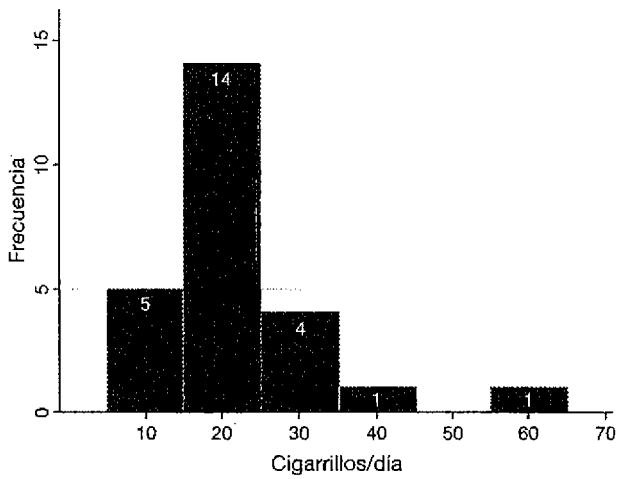



**Figura 2.11** Histograma con STATA. **histogram *numcig*, width(10) start(5) frequency /// fcolor(stone) lcolor(black) /// addlabel addlabopts(mlabsize(medium) /// mlabcolor(maroon) mlabposition(6)) /// ytitle(*Frecuencia*) xtitle(*Cigarrillos*/d) xlabel(10(10)70)**

situado en la parte superior de la ventana que se abre al presentar el gráfico. Es conveniente guardar en sintaxis *(Do-file)* las especificaciones escritas que se juzguen más idóneas, para reproducirlas así exactamente en el futuro.

#### 2.4.3.2. Histograma con SPSS

En SPSS se pueden crear histogramas con la siguiente expresión:

```
GRAPH /HISTOGRAM=numcig.
```

Una vez que se ha hecho lo anterior, basta con hacer doble clic en el gráfico obtenido para cambiar las especificaciones con los respectivos menús.

#### 2.4.3.3 Histograma con R

Se pedirá así:

```
> hist(numcig, col=3)
```

La opción **col**= selecciona el color de los rectángulos (2 = rojo, 3 = verde, etc.).

#### 2.4.3.4. Histograma con otros programas

Se recomienda, sobre todo, STATA o SPSS (y nunca Excel) para hacerlos.

### 2.4.4. Tallo y hojas *(stem and leaf)*

Para combinar la representación gráfica con la información directa proporcionada por las cifras se usa el gráfico de tallo y hojas *(stem and leaf)*. La ventaja es que el rectángulo está relleno de los propios valores numéricos, pero se evita la repetición de los primeros dígitos de cada cifra. Se puede elegir su amplitud, aunque siempre es preferible que las amplitudes sean de 5 o de 10 unidades. Véase la tabla 2.5.

De un vistazo aparece el histograma, sin más que girar la figura mentalmente 90 grados hacia la izquierda.

R/Splus proporciona el gráfico de tallo y hojas escribiendo el comando siguiente:

```
> stem(edad, scale=2)
```

### 2.4.5. Polígono de frecuencias acumuladas

Esta representación considera las frecuencias acumuladas. Se trata de ir representando cuántos sujetos en la muestra presentan, por lo menos, un determinado valor de la variable, es decir, cuántos

*Tabla 2.5 Tallo y hojas de la edad realizado con STATA (izquierda) y SPSS (derecha)*

| STATA | SPSS |
|---|---|
| **Stem-and-leaf plot for edad (edad del paciente)** | **edad Stem-and-Leaf Plot** |
| 1* \| 9 | **Frequency Stem & Leaf** |
| 2* \| 0127 | **1,00 1. 9** |
| 3* \| 1 | **4,00 2. 0127** |
| 4* \| 127 | **1,00 3. 1** |
| 5* \| 0114699 | **3,00 4. 127** |
| 6* \| 24679 | **7,00 5. 0114699** |
| 7* \| 1257 | **5,00 6. 24679** |
| (Hay una persona con 19 años, otra con 20, otra con 21, etc., hasta la mayor, que tiene 77 años. Girándolo mentalmente 90° hacia la izquierda se *ve* el histograma) | **4,00 7. 1257** |
| **stem edad** | **Stem width: 10** |
| | **Each leaf: 1 case(s)** |
| | **EXAMINE edad** |
| | **/PLOT STEMLEAF** |
| | **/STATISTICS none.** |

tienen ese valor *o uno menor.* Por ejemplo, se trataría de contar cuántos sujetos tienen esa edad o una inferior, que es fácilmente estimable a partir de una tabla de frecuencias acumuladas como la presentada en la figura 2.12.

La primera columna recoge los valores observados. La segunda muestra cuántas veces se observa cada valor (frecuencia). Puede apreciarse que los valores 51 y 59 son los únicos que se repiten. La tercera indica el porcentaje de sujetos que tienen cada valor. La siguiente (porcentaje acumulado) es la que nos interesa, porque indica el porcentaje de sujetos que presentan ese valor o uno inferior.

Para hacer el polígono de frecuencias acumuladas, en cada intervalo se incluirá el porcentaje acumulado hasta ese valor, es decir, el porcentaje que suponen todas las observaciones de ese intervalo y los intervalos inferiores a él. Esto se puede representar gráficamente usando los valores que toma la variable en el eje horizontal (abscisas = años de edad) y los porcentajes acumulados de los que tienen esa edad o una inferior en el eje de ordenadas. Así se obtiene la figura 2.13.

La interpretación es bastante directa. La línea dibujada marca, en el eje de ordenadas, el porcentaje de la muestra que tiene al menos la edad que aparece en el de abscisas. Por ejemplo, hay un 40% de sujetos que tienen 50 años o menos. Para obtener esta gráfica hay que dar dos pasos. Primero se pide a STATA (**tabulate**) o SPSS (**FREQ**) una descripción de la variable (distribución de frecuencias). Después se usan los datos acumulados para realizar la gráfica con las opciones propias de un gráfico de dispersión (v. más adelante).

| edad | Freq. | Percent | Cum. |
|---|---|---|---|
| 19 | 1 | 4.00 | 4.00 |
| 20 | 1 | 4.00 | 8.00 |
| 21 | 1 | 4.00 | 12.00 |
| 22 | 1 | 4.00 | 16.00 |
| 27 | 1 | 4.00 | 20.00 |
| 31 | 1 | 4.00 | 24.00 |
| 41 | 1 | 4.00 | 28.00 |
| 42 | 1 | 4.00 | 32.00 |
| 47 | 1 | 4.00 | 36.00 |
| 50 | 1 | 4.00 | 40.00 |
| 51 | 2 | 8.00 | 48.00 |
| 54 | 1 | 4.00 | 52.00 |
| 56 | 1 | 4.00 | 56.00 |
| 59 | 2 | 8.00 | 64.00 |
| 62 | 1 | 4.00 | 68.00 |
| 64 | 1 | 4.00 | 72.00 |
| 66 | 1 | 4.00 | 76.00 |
| 67 | 1 | 4.00 | 80.00 |
| 69 | 1 | 4.00 | 84.00 |
| 71 | 1 | 4.00 | 88.00 |
| 72 | 1 | 4.00 | 92.00 |
| 75 | 1 | 4.00 | 96.00 |
| 77 | 1 | 4.00 | 100.00 |
| Total | 25 | 100.00 | |

**Figura 2.12** STATA: distribución de frecuencias de la variable edad.

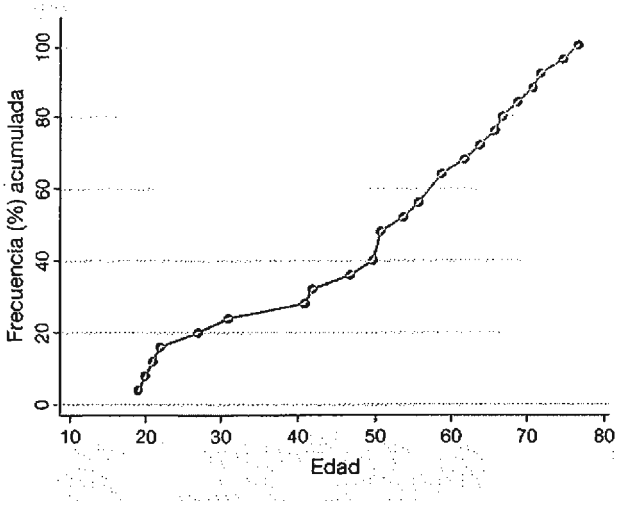



**Figura 2.13** Polígono de frecuencias (porcentajes) acumuladas correspondientes a la tabla 2.5.

### 2.4.6. Gráfico de caja *(box plot, box and whisker plot)*

Es una representación de uso creciente por sus interesantes propiedades descriptivas. Consiste en un rectángulo, que es la *caja,* y unas prolongaciones verticales, que son los bigotes o *whiskers.* Los límites de la caja son los percentiles 25 y 75. La línea que se encuentra en el centro de la caja es la *mediana* (percentil 50). Los bigotes van desde los percentiles 25 y 75 hasta los valores *adyacentes* mínimo y máximo. Pero pueden existir puntos periféricos, más allá del mínimo y máximo adyacentes, que superarían a los bigotes. Lo más habitual es presentarlo en vertical, como sucede en la figura 2.14, que describe dos variables, pero se puede presentar también en horizontal (fig. 2.15). En la figura 2.16 se presenta en vertical.

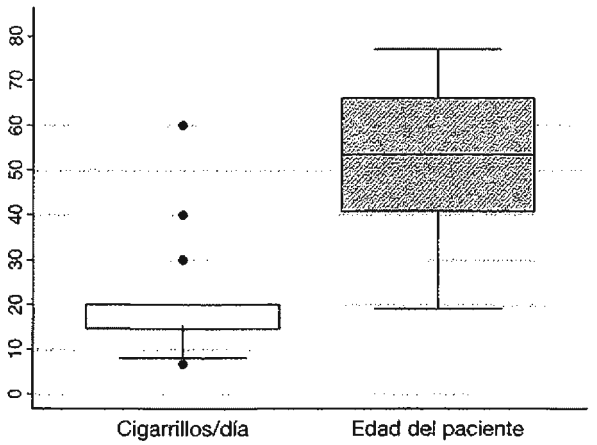



**Figura 2.14** Gráficos de cajas para dos variables distintas. Hay periféricos *(outliers)* solo en la primera. **graph box *numcig edad*, legend(off) showyvars /// box(1, fc(gs4) lc(red)) /// box(2, fc(green) lc(blue)) intensity(10) /// marker(1, mc(gs4) msiz(small)) ylab(0(10)80)**

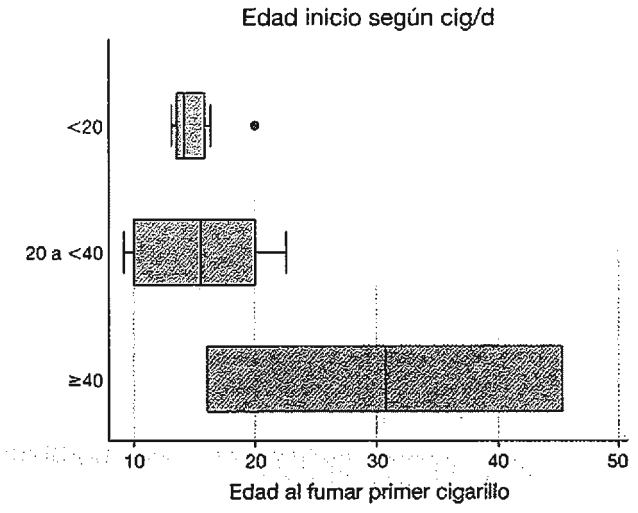
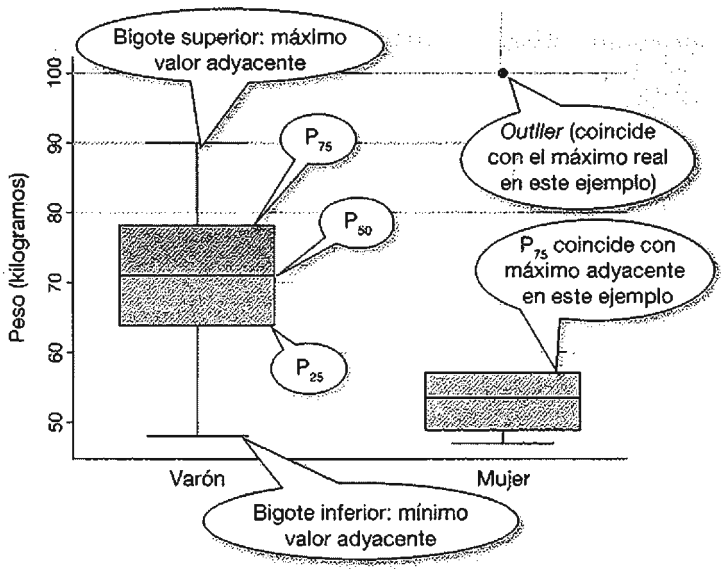



**Figura 2.15** Gráficos de cajas en horizontal. **gr hbox *inicio*, over(*ciggrup*) title("Edad inicio según cig/d")**




| Valores de los pesos | |
|---|---|
| Hombres | Mujeres |
| 48 | 47 |
| 51 | 49 |
| 61 | 52 |
| 63 | 55 |
| 64 | 57 |
| 67 | 100 |
| 67 | |
| 68 | |
| 68 | |
| 71 | |
| 75 | |
| 75 | |
| 77 | |
| 78 | |
| 78 | |
| 80 | |
| 83 | |
| 90 | |
| 90 | |

**Figura 2.16** Gráficos de cajas para una sola variable (peso) según categorías de otra variable (sexo). **gr box *peso*, over(sexo) inten(20) box(1, fcolor(blue))**

En la figura 2.15 se interpreta un gráfico de cajas en horizontal para la variable edad de inicio en el tabaco en función de las categorías de cigarrillos-día. En la figura 2.16 se describe el peso separadamente para hombres y mujeres.

Las llamadas indican que los límites inferior y superior de la caja corresponden a los percentiles 25 ($P_{25}$, que es el límite que deja por debajo al 25% de los sujetos más delgados) y 75 ($P_{75}$, deja debajo al 75% más delgado). También se llaman cuartiles. El primer cuartil es el 25%, con pesos

inferiores, y el cuarto cuartil es el 25%, con pesos superiores. Es decir, el cuarto cuartil está formado por todos los sujetos cuyos pesos están por encima de $P_{75}$. La mediana, llamada $P_{50}$ en la figura 2.16, es el valor que deja la mitad de los sujetos debajo (los más delgados) y la mitad encima (los más pesados). Las explicaciones de estos conceptos se verán más adelante cuando se hable de medidas de posición. De momento, puede comprobarse que la altura de la caja va desde 64 a 78 kg en varones, y desde 49 a 57 kg en mujeres. A esta distancia se le llama rango intercuartílico (RIC), y sus límites corresponden a los percentiles 25 y 75. Entre esos dos límites (con frecuencia, pero no siempre) estará el 50% de los sujetos.

La *línea horizontal* que está dentro de la caja es la *mediana o percentil 50 ($P_{50}$)*. La mediana es el valor que deja a la mitad de los individuos por encima y a la otra mitad por debajo. En el ejemplo, la mediana vale 71 kg en varones y 53,5 kg en mujeres.

El *bigote superior* es una prolongación de la caja que termina en el valor absoluto que sea igual o inferior al percentil 75 más 1,5 veces el RIC; a ese valor, que muchas veces (pero no siempre) será el máximo observado, se le llama valor *adyacente* superior. En la figura 2.16 se considerará que un valor máximo sigue siendo adyacente siempre que no supere, por ejemplo, en *mujeres*, el límite de 69 kg.

$$RIC = P_{75} - P_{25}$$
$$RIC = 57 - 49 = 8$$
$$8 \times 1,5 \text{ veces} = 12$$

Máximo valor *posible adyacente* para el bigote superior: 57 + 12 = 69

Como no hay nadie que pese exactamente 69 kg, el valor adyacente superior será la persona de peso inmediatamente inferior, que pesa 57 kg, y se pone ahí el bigote superior.

El *bigote inferior* termina en el valor absoluto que sea igual o superior al percentil 25 menos 1,5 veces el RIC. En mujeres, este límite será:

$$P_{25} - (1,5\,RIC) = 49 - 12 = 37$$

Como no hay ninguna mujer que pese 37 kg, se elige a la inmediatamente superior, que pesa 47 kg, y se sitúa ahí el bigote inferior.

El error más frecuente que se comete al construir o interpretar un diagrama de cajas es confundir el límite teórico de un bigote (en el ejemplo, 69 y 37) con el valor adyacente realmente presente en los datos que más se acerca a él *desde el centro* de la distribución (en el ejemplo, 57 y 47). Salvo que exista en los datos, tal límite teórico *no* se representará nunca en el diagrama de cajas.

Los valores *periféricos o outliers* son aquellos que quedan más allá de los bigotes. Se consideran observaciones «raras» *(outliers)*. La palabra *outlier* no tiene sentido peyorativo. Por ejemplo, se ha usado para denominar a personas que lograron éxitos extraordinarios (7). En nuestro ejemplo hay una persona «*outlier*» en el peso entre las mujeres. Estos valores periféricos deben ser tratados con cuidado, porque su presencia puede alterar engañosamente las medidas numéricas que se suelen calcular para resumir o analizar los datos.

### 2.4.6.1. Cajas con STATA

¿Cómo hacer un diagrama de cajas con STATA? Una observación atenta de los pies de las figuras 2.14 a 2.16, en combinación con la tabla 2.4, permitirá adquirir bastante destreza para realizar estos gráficos.

También puede consultarse:

```
help box
```

y el siguiente vídeo: STATA_2_4: CAJAS (box-plot) en www.unav.es/departamento/preventiva/recursos_bioestadistica.

El modo en que STATA calcula los límites para los bigotes (valores adyacentes) es ligeramente diferente de lo que se ha explicado arriba, aunque se parece mucho y casi siempre coincidirá con lo aquí explicado.

#### 2.4.6.2. Cajas con SPSS

¿Cómo hacer un diagrama de cajas con SPSS?

Se encuentra en la opción:

**Gráficos → Cuadros de diálogo antiguos → Diagramas de cajas**

Realmente SPSS está programado, en esta opción, para comparar la distribución de distintas variables, situando verticalmente, una al lado de la otra, las cajas de diversas variables. Por ejemplo, para obtener la figura 2.14 se daría la siguiente orden:

```
EXAM numcig edad
 /COMPARE VARIABLE
 /PLOT=BOXPLOT
 /STATISTICS=NONE.
```

Pero si se introduce solo una variable, también la representará en solitario y no es necesario ya incluir la instrucción **/COMPARE**.

Para obtener la figura 2.16 se hará lo siguiente:

```
EXAM peso
 /STAT NONE /PLOT=BOXPLOT /PANEL COLVAR=sexo.
```

Cuando un valor periférico es muy lejano, en vez de representarlo como un punto, SPSS lo dibuja como un asterisco (valor *extremo*). Se considera simplemente como un *outlier* o valor periférico si excede en más 1,5 veces el rango intercuartílico al percentil 75 (como se ha visto, y lo representa como un punto); en cambio, se considera un valor *extremo* si supera al percentil 75 en *tres veces* el rango intercuartílico.

Si se hace doble clic sobre el gráfico, se pueden introducir en él modificaciones.

#### 2.4.6.3. Cajas con R

Si se guardaron los datos con el nombre ***d*** la última vez, para obtener una gráfica similar a la figura 2.16 bastará hacer:

```
> attach(d)
> sexo<-factor(sexo,labels=c("varón","mujer"))
> boxplot(peso~sexo, col="green")
```

El diagrama de cajas es útil en la fase de depuración de una base de datos, antes de iniciar el análisis, cuando se desea comprobar la calidad de la recogida de datos estadísticos detallados. Esta depuración es un paso imprescindible y de suma importancia. Se aconseja vivamente hacer, al menos, un diagrama de cajas de cada variable para detectar aquellos *outliers* (siempre existen en alguna variable) que probablemente se deban a errores en la recogida o anotación de los datos. No se debe proceder al análisis estadístico hasta averiguar a qué se deben estos posibles errores y corregirlos.

### 2.4.7. Gráficos de dispersión

La descripción de la relación entre dos variables numéricas se hace mediante un gráfico de dispersión *(scatter plot)*, también llamado *nube de puntos*. Si se desea, por ejemplo, representar la talla con respecto al peso, se deberá construir un gráfico como el de la figura 2.17.

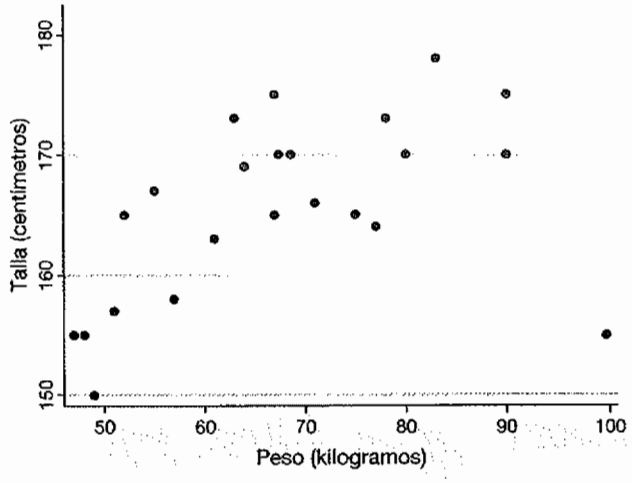



**Figura 2.17** Gráfico de dispersión para relacionar dos variables numéricas. **twoway scatter *talla peso***

Si existiesen dos o más individuos con valores idénticos de ambas variables, se podría mover mínimamente alguno de los dos datos para que aparezcan ligeramente separados los distintos puntos que realmente existen, pues de lo contrario se solaparán y el ordenador los representará como si fuesen solo uno[5].

### 2.4.7.1. Dispersión en STATA *(twoway scatter)*

STATA tiene muchas posibilidades que exceden los objetivos de este capítulo. La orden inicial es:

```
twoway
```

Debe ir seguida del tipo de marcador que se desee. Para nube de puntos:

```
twoway scatter yvar xvar
```

Para una línea que una todos los puntos unos con otros —se usa poco y requiere que la base de datos esté ordenada por la variable *xvar* (**sort** *xvar*)—:

```
twoway line yvar xvar
```

Para ambas (se usa poco):

```
twoway (scatter yvar xvar) (line yvar xvar)
```

5 En el ejemplo, esto se ha hecho en STATA con las siguientes órdenes:

```
clonevar w=peso
sort peso
list peso tallain 13/14
replace w=68.6 in 13
replace w=67.4 in 14
list peso w tallain 13/14
twoway (scatter talla w)
```

Para una nube de puntos y la recta que mejor resume los puntos (se usa mucho):

```
twoway (scatter yvar xvar) (lfit yvar xvar)
```

Para una nube de puntos acompañada de la curva que mejor los resume:

```
twoway (scatter yvar xvar) (lpoly yvar xvar)
```

Se pueden usar varias (y otras) opciones a la vez e introducir cambios de colores, de grosor de línea y otras modificaciones, según indica la tabla 2.4.

#### 2.4.7.2. Dispersión en SPSS

El gráfico se puede seleccionar en el menú:

**Gráficos → Cuadros de diálogo antiguos → Dispersión/Puntos**

Con sintaxis, se pide así:

```
GRAPH /SCATTER peso WITH talla.
```

Una vez que aparece el gráfico, si se hace doble clic sobre él se pueden introducir modificaciones. Se puede pedir, por ejemplo, que ajuste una recta que resuma la relación lineal existente entre las dos variables. Una vez que se tiene el gráfico ampliado en pantalla y se ha pulsado en uno de los puntos, se elige la opción:

**Gráficos → Elementos → Línea de ajuste total**

También se le puede pedir que ponga una línea en la media de Y, una curva (regresión cuadrática o cúbica) o una línea con múltiples inflexiones que corresponde al procedimiento LOESS de regresión no paramétrica o suavizada *(smoothing)*. Para este último procedimiento (LOESS), el ordenador ofrece la opción de contar con el 50% de los puntos observados, pero esto se puede cambiar. Seleccionar el 70% supondrá que se emplean solo los puntos más próximos (el 70%) del total en cada trámo del trayecto que recorre la línea. Además, las influencias de cada uno de los otros puntos se ponderan localmente (se les da menos peso cuanto más lejanos estén del punto correspondiente) para definir por dónde pasará la línea en ese punto concreto. Cuanto menor sea el porcentaje de puntos que influya en la definición de la línea, más picos tendrá esta. Si se pide que el modelo tenga en cuenta un gran porcentaje de puntos (digamos el 80% o el 90%), la línea se presentará como suavizada *(smoothing)*.

#### 2.4.7.3. Dispersión en R

En primer lugar, se le debe pedir que utilice la base de datos ***d;*** en segundo lugar, que la variable talla tiene que ser tratada como numérica; después, que haga un gráfico de dispersión con los puntos en rojo, y, por último, que represente la recta resumen en azul (sin cerrar la ventana del gráfico de dispersión).

```
> attach(d)
> talla<-as.numeric(talla)
> plot(peso,talla, col="red")
> abline(lm(talla~peso), col="blue")
```

## 2.5. MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central estiman cuál es el valor más típico o representativo de una muestra. Son el elemento indispensable de cualquier estadística descriptiva.

### 2.5.1. Media aritmética

La primera y principal medida de tendencia central es la media aritmética. Cuando se habla del «promedio» o de «la media» sin más especificaciones, siempre se trata de la media aritmética. Es

la suma de todos los valores ($x_i$) dividida por el número de observaciones (n). La media de la población se expresa como $\mu$ y la media de una muestra, como $\bar{x}$ (se lee «x barra»). Sin embargo, es mejor no usar esta notación, sino simplemente escribir «media» cuando se presenten resultados en una publicación científica o en una comunicación a un congreso, y reservar la «x barra» solo para las ecuaciones. Para calcular la media aritmética, se suman[6] todos los valores y se dividen por el número de individuos.

$$\bar{x} = \frac{\sum(x_i)}{n}$$

Si se desea calcular la media de tres valores, 1, 2 y 3, la media será:

$$\text{Suma} = \sum(x_i) = 1+2+3 = 6$$

$$\text{Media aritmética} = \bar{x} = \frac{6}{3} = 2$$

Es la medida de tendencia central más usada. En su cálculo intervienen todos los valores. Su inconveniente es que se deja influir mucho por los valores extremos, especialmente si la muestra no es de gran tamaño.

### 2.5.2. Media geométrica

$$\text{Media geométrica} = \sqrt[n]{x_1 x_2 \ldots x_n} = \sqrt[n]{\prod_1^n x_i}$$

El símbolo que aparece dentro de la raíz ($\prod$, una letra griega pi mayúscula) es el multiplicatorio y significa que hay que multiplicar uno por otro todos los valores de la variable. La raíz no es una raíz cuadrada, sino una raíz n-ésima, siendo n el tamaño de muestra. Una raíz cuadrada es la raíz 2, una raíz cúbica es la raíz 3. Eso es lo que significa la n en el superíndice junto al símbolo de raíz. La media geométrica de los valores 1, 2 y 3 sería la raíz cúbica del producto de $1 \times 2 \times 3$.

$$\text{Media geométrica} = \sqrt[3]{1\times 2\times 3} = \sqrt[3]{6} = 1{,}817$$

La media geométrica suele usarse poco habitualmente, pero a veces resulta útil, por ejemplo, en microbiología, ya que las variables que se manejan suelen crecer exponencialmente. Además, la media geométrica es más robusta que la media aritmética, ya que se desvirtúa menos si existen valores muy extremos.

Se puede calcular también la media geométrica de otro modo:

1. Se toman logaritmos neperianos de los datos de la variable.
2. Se calcula la media de esos logaritmos.
3. Se eleva el número e a la cantidad calculada en el paso anterior.

Es decir:

1. $\ln(1)+\ln(2)+\ln(3) = 0+0{,}693+1{,}099 = 1{,}792$
2. $\frac{1{,}792}{3} = 0{,}597$
3. $e^{(0{,}597)} = 1{,}817$

6 El símbolo Σ significa sumar todos los valores de una variable, se lee «Sumatorio».

### 2.5.3. Media armónica

Se calcula dividiendo el número de observaciones por la suma del inverso de cada valor.

$$\text{Media armónica} = \frac{n}{\sum\left(\frac{1}{x_i}\right)} = \frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}} = \frac{3}{1{,}833} = 1{,}636$$

También se utiliza poco, pero tiene aplicaciones en farmacología. Se dice que tanto la media geométrica como la media armónica son estimadores de tendencia central más *robustos* que la media aritmética: esto significa que se dejan influir menos por valores raros o extremos.

### 2.5.4. Media ponderada

Se utiliza mucho. Por ejemplo, supongamos que un alumno ha obtenido las siguientes calificaciones en la asignatura «Bioestadística» de primero de medicina:

- Pruebas de clase: 8.
- Prácticas: 10.
- Examen final: 4.

Si a las prácticas y a las pruebas de clase se les da un peso del 25%, y al examen final del 50%, ¿cuál será la media ponderada? Si llamamos $w_i$ a los pesos:

$$\text{Media ponderada} = \frac{\sum(w_i x_i)}{\sum(w_i)} = \frac{w_1x_1 + w_2x_2 + w_3x_3}{w_1 + w_2 + w_3}$$

$$\text{Media ponderada} = \frac{(25\times 8)+(25\times 10)+(50\times 4)}{25+25+50} = \frac{650}{100} = 6{,}5$$

La presión media en la raíz de la aorta, una medida frecuente en clínica, es estimada de modo bastante aproximado a partir de las mediciones de la presión en sístole (contracción del corazón, 120 mmHg) y diástole (relajación, 80 mmHg). Debido a su desigual duración, a veces se le asigna el doble de peso a la diástole que a la sístole. El resultado es, por tanto, una presión media de 93,3 mmHg:

$$\text{Media ponderada} = \frac{\sum(w_i x_i)}{\sum(w_i)} = \frac{(1\times 120)+(2\times 80)}{1+2} = \frac{280}{3} = 93{,}3 \text{ mmHg}$$

### 2.5.5. Mediana

La mediana es otra medida de tendencia central, más fácil de calcular que las dos anteriores. Puede definirse como el valor central del conjunto ordenado de observaciones; es decir, el valor que deja la mitad de las observaciones por debajo y la mitad por encima. Por ejemplo, si las edades de cinco personas (ordenadas de menor a mayor) son 32, 35, 45, 67 y 79 años, la mediana será 45, que es el valor que deja dos observaciones por debajo y dos por encima. El 50% de los individuos tendrán una edad inferior a la mediana y en el otro 50% su edad superará a la mediana. Como se verá más adelante, la mediana coincide con el percentil 50. Cuando el número de datos es par, para hallar la mediana se calcula la media entre los dos datos centrales. En el ejemplo «canga25», las edades ordenadas de las mujeres eran:

20 27 42 56 59 59

La mediana sería (42 + 56)/2 = 49.

Se hace la media entre 42 y 56, y esa es la mediana. La mediana es una medida de tendencia central que es *robusta*. Esto significa, por ejemplo, que si la mujer de mayor edad tuviese 100 años, la mediana seguiría siendo 49. En cambio, la media aritmética subiría de 43,8 a 50,7. Al mismo tiempo, la mediana es más fácil de calcular que la media geométrica o la media armónica. Por tanto, la mediana es *la medida de tendencia central que se usará cuando en muestras pequeñas haya alguna observación extrema* («outlier») *o cuando existan datos truncados* o *«censurados»* (v. apartado 2.15). Se dice que la mediana es *robusta* porque no se deja influir mucho por valores extremos. La mediana es, en muchos aspectos, más robusta que la media geométrica o la media armónica. Sin embargo, tiene un inconveniente, y es que no se usan todos los valores observados para calcularla, sino solo el valor central o los dos valores centrales.

### 2.5.6. Moda

La moda tiene poco interés. Es el valor más frecuente, por lo que es muy sencilla de calcular. En el ejemplo «canga25», el valor más frecuente de edad no es único (v. fig. 2.12), por lo que se dice que hay dos modas, o que la distribución es *bimodal*. Las dos modas son 51 y 59. De cada uno de estos valores hay dos observaciones. La moda es una medida de tendencia central poco rigurosa *(la moda es frívola)* y casi nunca tiene utilidad práctica para describir datos continuos.

## 2.6. MEDIDAS DE DISPERSIÓN

### 2.6.1. Varianza

Para resumir unos datos no basta con decir cuál es su centro, sino que también hay que indicar en qué medida están juntos o separados de ese valor central. A esta característica se le llama *dispersión*. Cuanto más separados estén unos datos del valor central, más dispersos serán. La dispersión expresa el grado de *variabilidad* de unas observaciones.

- A continuación se presentan las edades (en años) de dos grupos de nueve personas: tienen la misma media ($media_1 = media_2 = 49$), pero se trata de dos grupos de personas muy diferentes. La variabilidad cambia mucho de un grupo a otro. En el primer caso, la media se aproxima al valor de cualquier sujeto. En cambio, en el segundo ejemplo, con mucha dispersión, la media sería poco representativa del valor de cada sujeto.

| | |
|---|---|
| Poca dispersión (A) | 47 48 48 49 49 49 49 50 52 |
| Mucha dispersión (B) | 3 11 22 34 49 66 73 84 99 |

Las situaciones A y B son muy diferentes. Por tanto, para resumir la información que hay en un conjunto de datos no basta con decir cuál es su media (u otra medida de tendencia central). Es preciso indicar también su variabilidad o dispersión. Cuanto más separados estén los valores de la media, mayor será su dispersión. La varianza es una medida de dispersión. La idea que hay detrás del concepto de varianza es hacer un promedio de las desviaciones de cada valor con respecto a la media ($x_i - \bar{x}$), pero la suma de estas cantidades siempre resultará cero, porque hay unas positivas y otras negativas, que se anulan exactamente. La solución consiste en elevar estas diferencias al cuadrado.

Así, la varianza de una *muestra* tiene la siguiente expresión:

$$\text{Varianza muestral } (s^2) = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

En esta expresión hay que tener presente que:

1. Al numerador de esta expresión se le conoce como *suma de cuadrados:*

$$\text{Suma de cuadrados} = \sum (x_i - \bar{x})^2$$

2. En vez de usar la media poblacional ($\mu$), se usa la media muestral ($\bar{x}$).
3. En vez de usar $n$ como denominador, se usa $n-1$. A este denominador $n-1$ se le llama *grados de libertad*[7].

$$\text{Grados de libertad} = n-1$$

En el ejemplo anterior, la varianza de la primera muestra sería:

$$\text{Varianza muestral (A)} = s^2 = \frac{(47-49)^2+(48-49)^2+\ldots+(52-49)^2}{8} = 2$$

En cambio, para la segunda muestra, la varianza sería más de 500 veces superior:

$$\text{Varianza muestral (B)} = s^2 = \frac{(3-49)^2+(11-49)^2+\ldots+(99-49)^2}{8} = 1.138$$

En todos los textos de estadística, además de la varianza muestral, se menciona la varianza *poblacional*. Esta es incalculable en una muestra y normalmente corresponde tan solo a un concepto teórico, ya que habitualmente es imposible acceder al total de la población de donde se extrajo una muestra.

Hay otras opciones, menos usadas, para estimar la dispersión[8].

Las fórmulas matemáticas de la varianza muestral y poblacional son distintas, tal como se presenta en la tabla 2.6.

Entre la varianza muestral y la poblacional hay varias diferencias que tienen importancia conceptualmente, pero no mucha para el cálculo práctico, ya que la que se deberá calcular es la varianza muestral. En la varianza poblacional:

---

7 Puede parecer paradójico algo tan elevado y poético como el concepto de libertad tenga que ver con algo tan prosaico como n – 1. Pero, al menos intuitivamente, se comprenderá que la varianza muestral tiene un grado menos de libertad que el tamaño de la muestra (n), porque expresa la dispersión en torno a una media muestral que, a su vez, es variable. Esto significa que, si se sabe la media y se van conociendo los valores de cada dato uno a uno, cuando se llegue al penúltimo valor necesariamente se sabrá cuál será el último, porque es el que se necesita para que obtener la media. Este último valor está condicionado, no es libre. Si se supiesen los ocho primeros datos del primer ejemplo (47 48 48 49 49 49 49 50) y que la media es 49, no haría falta decirnos el último dato (52), pues se puede deducir de los restantes ocho y la media. Por eso, los grados de libertad son uno menos que los datos, porque la media debe mantenerse constante.

8 Otra alternativa es la desviación absoluta media (DAM), que prescinde del signo de las diferencias:

$$\text{DAM} = \frac{\sum|x_i - \bar{x}|}{n}$$

Pero su uso es muy infrecuente. En la segunda muestra, la DAM podría calcularse así:

$$\text{DAM} = (|3-49|+|11-49|+|22-49|+|34-49|+|47-49|+|66-49|+|73-49|+|84-49|+|99-49|)/9$$
$$\text{DAM} = (46+38+\ldots+50)/9 = 28$$

Una tercera alternativa, que puede tener más uso en algunos supuestos, es la desviación absoluta mediana:

$$\text{DAMd} = \text{mediana}\,(|x_i - \text{mediana}|)$$

La mediana es 47 para esos nueve valores. En un primer paso se calculan las diferencias absolutas:

dif. abs.: $|3-47| = 44$; $|11-47| = 36$; $|22-47| = 25$; $|34-47| = 13$; $|47-47| = 0$; $|66-47| = 19$; $|73-47| = 26$; $|84-47| = 37$; $|99-47| = 52$

Que, una vez ordenadas, permiten calcular fácilmente que la mediana de estas diferencias será 26:

0, 13, 19, 25, 26, 36, 37, 44, 52

Por tanto: DAMd = **26**

***Tabla 2.6 Fórmulas de la varianza y la desviación estándar (se usará la* muestral*)***

| Varianza muestral | Varianza poblacional |
|---|---|
| $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ | $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$ |
| **Desviación estándar muestral** | **Desviación estándar poblacional** |
| $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ | $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ |

- Se ha sustituido $s^2$ por $\sigma^2$.
- La media que se usa es la poblacional ($\mu$) y no la muestral.
- Los grados de libertad son N en vez de n – 1.

Acompañar la media de una variable de su varianza permitiría al lector hacerse idea más completa de cómo son esos datos. Sin embargo, *la varianza no es muy adecuada para describir directamente cuál es la variabilidad de unos datos, ya que se encuentra expresada en otras unidades que los datos originales; la varianza está en unidades al cuadrado, y esto dificulta su interpretación directa.* Por este motivo se debe buscar otro índice estadístico de dispersión que esté en las mismas unidades que la media: ese índice es la desviación estándar.

## 2.6.2. Desviación típica o desviación estándar

Para calcular la varianza se elevaban las desviaciones respecto a la media al cuadrado para evitar que se anulasen unas a otras, ya que unas son negativas y otras positivas. La desventaja es que el resultado acaba medido en unidades distintas a las de la media por la elevación al cuadrado. Para eliminar este defecto, extraemos la raíz cuadrada de la varianza. Al resultado de esta raíz cuadrada se le llama desviación típica o desviación estándar (s si es muestral, $\sigma$ si es poblacional).

La desviación estándar en cada uno de los dos casos anteriores será:

$$\text{Poca dispersión (A)}: \quad s = \sqrt{2} = 1{,}414$$

$$\text{Mucha dispersión (B)}: \quad s = \sqrt{1.138} = 33{,}73$$

Si se suma o resta una cantidad constante a todos los datos, la media se incrementará o disminuirá en esa cantidad, pero la desviación estándar no cambiará.

Se puede calcular la desviación estándar (DE) de los siguientes tres números:

$$\text{Valores}: 1,\ 2,\ 3$$

Se comprobará que DE = 1. Ese valor (DE = 1) estima la distancia típica entre cada uno de los demás valores y la media.

A diferencia de la varianza, la desviación estándar sí tiene las mismas unidades de medida que los datos originales y puede, por tanto, ser más fácilmente comprendida y presentarse como descripción de la variabilidad de unos datos en un trabajo de investigación. Se puede presentar como resumen de los datos la expresión: media ± desviación estándar (pero teniendo cuidado de indicarlo). Quizás sea mejor presentar la media y entre paréntesis la desviación estándar, indicando de qué se trata. En concreto:

$$\text{Edad}: 49 \pm 1{,}4\ (\textbf{media} \pm \textbf{desviación estándar})$$

o bien, simplemente:

$$\text{Edad}: 49\ (\text{DE}: 1{,}4)$$

Cuando una variable sigue una distribución que se parece a un modelo conocido como *distribución normal* o campana de Gauss (v. apartado 3.10), puede asumirse que el 95% de los valores se situarán aproximadamente en el espacio comprendido por la media ± 2 veces la desviación estándar. Así, si en un congreso alguien presenta una comunicación y dice que sus pacientes tenían una edad media de 50 años y una desviación estándar de 5 años, el auditorio puede imaginar que la gran mayoría de ellos tenían entre 40 y 60 años. Se estará asumiendo entonces una distribución *normal* de los datos. Pero esto muchas veces no será así, porque la distribución de los datos puede diferir mucho de la campana de Gauss. Cuanto más se distancie la muestra de la campana de Gauss teórica, *peor* será esta aproximación de sumarle y restarle 2 desviaciones estándar a la media para saber entre qué dos valores estará el 95% de las observaciones. A veces, tal aproximación será *pésima*. Esto sucede, sobre todo, porque, con cierta frecuencia, el histograma es asimétrico (una cola, generalmente la derecha, es más larga que otra).

Esta descripción de una variable (media y DE) permite así demostrar cierta sutileza como lectores críticos de artículos científicos. Surgen sospechas sobre la simetría de la distribución de la variable cuando la desviación estándar es muy grande con respecto a la media: esto suele significar (en variables que no pueden tomar valores negativos) que la cola de la derecha del histograma es muy larga (**8**).

Si, por ejemplo, en un estudio se lee que la media de consumo de alcohol entre 1.837 varones era de 14,6 g/día y su desviación estándar era de 18,9 g/día (**9**), se podrá hacer la operación antes mencionada (media ± 2 DE):

$$14{,}6-(2\times 18{,}9)=-23{,}2$$
$$14{,}6+(2\times 18{,}9)=+52{,}2$$

Se obtendría el rango teórico de valores para el 95% central de los participantes. En este ejemplo se aprecia que tal rango no es posible, pues no pueden darse valores negativos para el consumo de alcohol. Esto sucede porque la distribución del consumo de alcohol es asimétrica, con una cola derecha más larga que la izquierda. Unos pocos valores de consumo de alcohol muy altos y, por tanto, muy separados de la media por arriba son los responsables de que la desviación estándar sea tan grande con respecto a la media.

Siempre que una variable no pueda tener valores negativos y se observe que la desviación estándar es mayor que la mitad de la media, se podrá intuir que su histograma será asimétrico, con una cola más larga por la derecha.

### 2.6.3. Coeficiente de variación

El coeficiente de variación es la razón o cociente entre la desviación típica y el valor de la media aritmética.

$$\text{Coef. de variación} = \frac{\text{desviación típica}}{\text{media}} = \frac{s}{\bar{x}}\times 100$$

En el ejemplo del alcohol antes mencionado, la media = 14,6 g/día y la desviación estándar = 18,9 g/día. Por tanto, el coeficiente de variación valdrá:

$$\text{Coef. de variación} = \frac{s}{\bar{x}}\times 100 = \frac{18{,}9}{14{,}6}\times 100 = 1{,}295\times 100 = 129{,}5\%$$

El coeficiente de variación suele expresarse como tanto por ciento, pues estima qué porcentaje de la media supone la desviación típica. El coeficiente de variación representa la desviación estándar medida en unidades de magnitud relativas a la media. Aunque se exprese como porcentaje, puede tomar valores por encima del 100% (como en el ejemplo del alcohol).

El coeficiente de variación (y no la desviación estándar) es la medida de dispersión indicada para comparar la variabilidad de distintos parámetros cuando están medidos en unidades diferentes. La desviación estándar depende de las unidades de medida de la variable. El coeficiente de variación, en cambio, no se ve afectado por las unidades de medida.

### 2.6.4. Error estándar de la media

El «error estándar» o error típico es un concepto clave en estadística. No se debe confundir error estándar con desviación estándar. La desviación estándar —como se ha visto— mide el grado de dispersión de los *individuos* que forman la muestra. En cambio, el error estándar de la media medirá el grado de dispersión de las *medias* de todas las posibles muestras que pudieran extraerse de la población.

En el mismo ejemplo anterior de 1.837 varones (**9**), la media y la desviación estándar de la edad valían 66 y 6,6 años, respectivamente. Esta desviación estándar (6,6) estima la distancia a la que se encuentra un individuo (de los que hay en la muestra) con respecto al valor medio de la edad del grupo (66 años). Del mismo modo, la desviación estándar (DE) de los números 1, 2 y 3 será DE = 1, y ese valor (DE = 1) estimará la distancia típica de cada uno de los demás a la media.

En cambio, el error estándar estima la desviación típica de *un grupo imaginario* de valores formado por las *medias* de las posibles muestras de tamaño que se pueden obtener de la población de la que procedió esa muestra. Una de esas muestras imaginarias es la que se está estudiando. El error estándar[9], por tanto, es el error estándar *de la media*. Se calcula así[10]:

$$\text{Error estándar} = \frac{s}{\sqrt{n}}$$

Para la edad, el error estándar de la media sería en el ejemplo de los tres datos:

$$\text{Error estándar} = \frac{\text{desviación estándar}}{\sqrt{n}} = \frac{1}{\sqrt{3}} = 0{,}58$$

En el ejemplo del alcohol en 1.837 varones, el error estándar de la media será:

$$\text{Error estándar} = \frac{18{,}9}{\sqrt{1.837}} = 0{,}44$$

Para interpretar el error estándar de la media hay que partir del principio de que la muestra ha sido extraída de una población de mucho mayor tamaño. Esta es solo una de las posibles muestras que se pueden extraer. En nuestra muestra hemos calculado una media, pero podíamos haber elegido otras muestras distintas procedentes de la misma población y habrían resultado otras medias. ¿Qué grado de dispersión tendrán las medias de todas las posibles muestras de 1.837 varones que se podrían extraer de esa población de la que se ha extraído esta muestra concreta? Eso es lo que estima el error estándar de la media.

### 2.6.5. Otras medidas de dispersión

El *rango o amplitud* es otra medida de dispersión, que simplemente consiste en restar los valores mayor y menor que se observen. En los tres datos (1, 2, 3), el rango sería:

$$\text{Rango} = 3 - 1 = 2$$

---

9 Esta es la expresión del error estándar de la media. Hay muchos errores estándar aparte del de la media. Los iremos viendo en sucesivos apartados. En general, se trata de la medida de dispersión del *estimador muestral* que se utilice en cada caso.

10 Esta expresión procede de $\sqrt{\frac{s^2}{n}}$, es decir, la varianza se divide por el tamaño de muestra. Para volver a las unidades de medida de la variable, se extrae la raíz cuadrada.

Cuando el histograma sea fuertemente asimétrico (se aleja mucho de la forma de campana), suele ser preferible, como medida de tendencia central, usar la mediana en vez de la media y, como medida de dispersión, utilizar el rango, o simplemente presentar dos percentiles (v. más adelante), como el $P_{25}$ y el $P_{75}$, o el $P_{10}$ y el $P_{90}$.

## 2.7. MEDIDAS DE FORMA: ASIMETRÍA Y CURTOSIS

### 2.7.1. Asimetría

Las distribuciones pueden ser simétricas o asimétricas. Se dice que son simétricas cuando las dos colas de su histograma (derecha e izquierda) tienen la misma longitud. Esto es más fácil de visualizar que de explicar[11]. Los tres histogramas que recoge la figura 2.18 corresponden a tres posibles situaciones en cuanto a la asimetría; en cada situación es posible calcular un *coeficiente de asimetría,* que puede tomar valores negativos o positivos.

La expresión matemática del coeficiente de asimetría es complicada y habitualmente se recurrirá al ordenador para calcularla. Cuando hay asimetría positiva, la cola de la derecha es más prolongada y su coeficiente de asimetría será positivo. En caso de asimetría negativa, la cola de la izquierda será más larga y el coeficiente, negativo. *Lo ideal para muchos procedimientos estadísticos es que la asimetría no sea grande y el coeficiente de asimetría esté lo más próximo posible a 0.*

En una variable que no puede tomar valores negativos, solo con conocer la media y la desviación estándar, ya podría decirse que tendrá siempre asimetría positiva cuando su desviación estándar sea superior al 50% de la media (es decir, si su coeficiente de variación es superior al 50%).

### 2.7.2. Curtosis o apuntamiento

El apuntamiento o curtosis mide el grado en el que un histograma resulta picudo o aplastado (fig. 2.19). Lo ideal es que el valor de la curtosis sea intermedio (próximo al valor nulo, mesocúrtico o normocúrtico). En casi todos los programas de estadística, el valor nulo de la curtosis es 0. Sin embargo, STATA suma tres unidades al calcular el coeficiente de curtosis y entonces el valor nulo es 3. Cuando se cumple esta condición y la asimetría es casi inexistente, se podrá considerar la distribución de los datos como normal. Como se verá, este tipo de distribución facilita enormemente el trabajo.

## 2.8. MEDIDAS DE POSICIÓN: CUANTILES, PERCENTILES

Los cuantiles son medidas de posición. Indican qué puesto ocupa un determinado valor de una variable en el conjunto ordenado de los datos de esa variable. Este puesto o posición se expresa como la proporción o porcentaje de los datos que queda por debajo de ese valor. A esta cantidad se le llama percentil. Así, que un niño esté en el percentil 80 del peso para su edad quiere decir que el 80% de los niños de su edad pesan menos que él. Si un alumno está en el percentil 100 de las notas de la clase, es que es el que mejor nota tiene de toda la clase.

Para calcular los percentiles se ordenan todas las observaciones de la distribución de menor a mayor y se busca aquel valor que deja un determinado porcentaje de las observaciones por debajo de él. Ya se ha visto que la mediana es el percentil 50 ($P_{50}$) porque deja por debajo al 50% de los sujetos. El percentil 5 es el que deja al 5% debajo de él, el percentil 90, al 90% de los individuos de la muestra, y así sucesivamente.

11 En casi todos los libros de estadística, estos conceptos se presentan con las curvas teóricas de distribuciones con mayor o menor grado de asimetría. Se presentan estos histogramas para aproximarnos más a la realidad práctica de que siempre se trabaja con muestras.

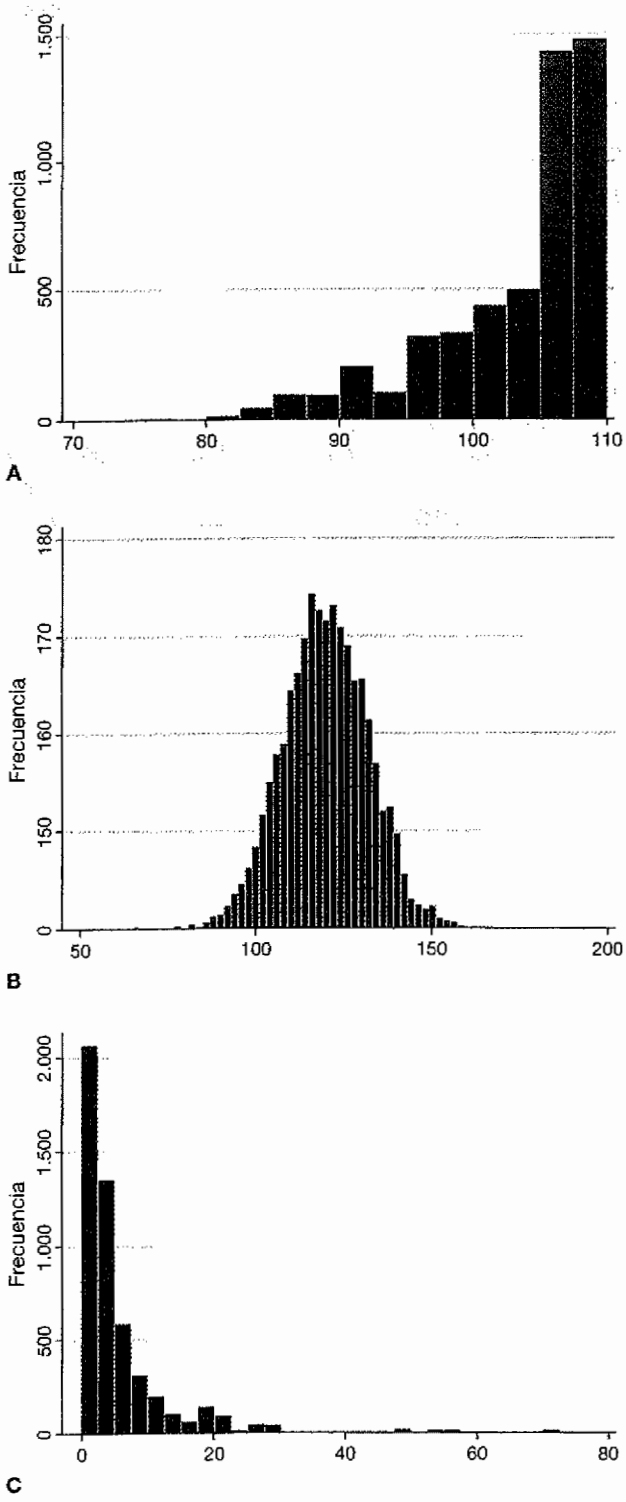



**Figura 2.18** Medidas de forma: coeficientes de asimetría. **A.** Asimetría negativa $< 0$. **B.** Simetría perfecta $= 0$. **C.** Asimetría positiva $> 0$.

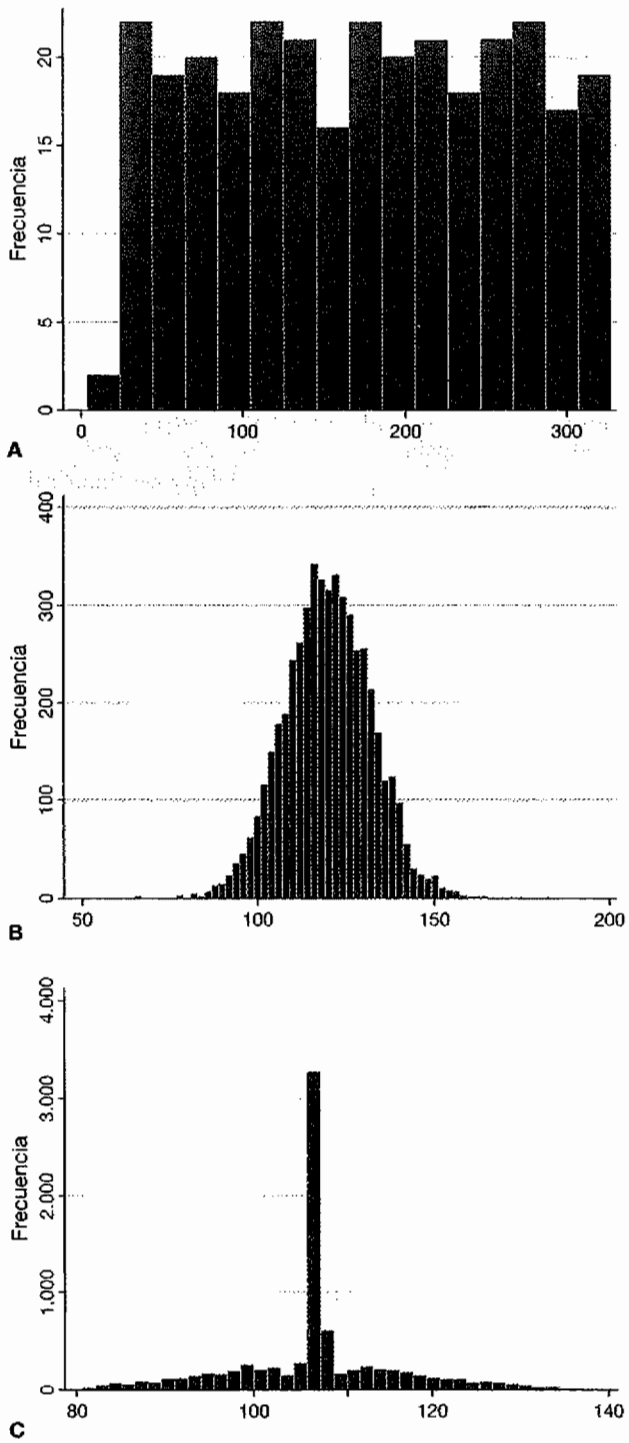



**Figura 2.19** Medidas de forma: coeficientes de curtosis. **A.** Curtosis negativa, <3 (en STATA), <0 (en otros), platicúrtica. **B.** Mesocúrtica (normocúrtica): curtosis = 3 (STATA), curtosis = 0 (otros). **C.** Curtosis positiva, >3 (STATA), >0 (otros), leptocúrtica.

Al hablar de los diagramas de caja ya se habían mencionado los percentiles 25 ($P_{25}$) y 75 ($P_{75}$). La mediana y estos dos percentiles ($P_{25}$ y $P_{75}$) son tres puntos de corte que dividen la muestra ordenada en cuatro partes iguales. Estos tres puntos de corte se llaman cuartiles. El rango intercuartílico (RIC) es la distancia entre el primer y el tercer cuartil ($RIC = P_{75} - P_{25}$).

También se habla de *terciles*, que son aquellos dos valores que dividen la muestra en tres grupos de igual tamaño. El primer tercil (o tercil 1) sería equivalente al percentil 33,33 y el segundo tercil, al percentil 66,67. Hay cuatro *quintiles* correspondientes a dar puntos de corte en los percentiles 20, 40, 60 y 80. También podría hablarse de *deciles*. Existen nueve puntos de corte (del percentil 10 al percentil 90) para definir 10 deciles.

No obstante, son términos equívocos y en la literatura científica es muy común el uso de, por ejemplo, quintil para hacer referencia tanto a los cuatro puntos de corte ($P_{20}$, $P_{40}$, $P_{60}$ y $P_{80}$) como a los cinco grupos de observaciones que quedan delimitados por estos cuatro cortes. De esta manera, el grupo de observaciones que queda por debajo del $P_{20}$ se denominaría el primer quintil, entre $P_{20}$ y $P_{40}$ el segundo quintil, etc. A su vez, al grupo situado por encima de $P_{80}$ se le llamará el quinto quintil. Conviene prestar atención para identificar en qué caso nos encontramos.

Para explicar cómo calcular un percentil se usará un ejemplo sencillo. Se dispone de las edades ordenadas de menor a mayor de ocho sujetos:

28 31 33 33 34 38 40 42

Se aplica una interpolación. Si se desea calcular, por ejemplo, el *percentil 25*, se debe calcular la siguiente expresión, donde $i$ es el percentil expresado en tanto por uno:

$$\text{Puesto} = i(n+1)$$

$$\text{Puesto} = 0{,}25 \times (8+1) = 2{,}25.^{\circ}$$

El puesto que le correspondería al percentil 25 es el número de orden 2,25.°

Para hallar el percentil 25 ($P_{25}$) se buscará, por tanto, el valor que ocupa el puesto 2,25.° en el conjunto ordenado de datos. El puesto 2.° está ocupado por el valor 31. El siguiente valor (el 3.er puesto) es 33. Interpolando resulta:

$$P_{25} = 31 + [0{,}25 \times (33 - 31)] = 31 + (0{,}25 \times 2) = 31{,}5$$

El percentil 25 valdrá por tanto *31,5*. Puede comprobarse que $P_{75}$ = 39,5.

El fundamento de este procedimiento es el siguiente: el decimal del número de puesto sirve de «factor de peso» para interpolar una fracción de la diferencia entre el puesto previo y el posterior. De este modo, el valor del percentil será más cercano a aquel de los dos valores que lo flanquean que se acerque más a su posición. El resultado del puesto o número de orden (2,25.° para el percentil 25) indica que el percentil 25 está a un 25% de la distancia que hay entre el puesto 2.° (valor = 31) y el 3.° (valor = 33). Se calcula cuál es el 25% de la distancia entre 31 y 33, y se suma esa distancia a 31. Por eso se dice que el cálculo se basa en la interpolación. No es el único modo de calcular percentiles. Hay otras aproximaciones. Por ejemplo, cuando se usa STATA para hacer gráficos de caja, a veces se obtiene otro resultado, porque STATA buscará los valores que se hayan *observado realmente* y estén más próximos al percentil teórico cuando se dibuja el gráfico de caja. No hay que preocuparse por esto. Habitualmente se hará con ordenador y se debe aceptar el gráfico resultante. Cuando el tamaño de muestra es grande, estas diferencias no se suelen notar.

## 2.9. PONDERACIÓN, MEDIAS PONDERADAS

La media ponderada es un método que ya se ha explicado y usa un sistema de ponderación consistente en dar a unas observaciones más peso o importancia que a otras. La ponderación se puede usar con muchas finalidades en bioestadística; por ejemplo, cuando se tienen razones fundadas para

pensar que hay un tipo de observaciones que están infrarrepresentadas en los datos disponibles. En esa situación se puede dar más peso a esas pocas observaciones para que representen mejor el conjunto de todos los valores faltantes. Pero esta aproximación tiene sus indicaciones y debe aplicarse solo en ciertas condiciones y con las técnicas apropiadas.

La ponderación puede utilizarse como un método eficiente de introducir datos en el ordenador y así evitar tener que repetir muchas veces el mismo dato.

### 2.9.1. Ponderación en STATA

La siguiente secuencia de órdenes escrita en un *Do-file* en STATA, si se ejecuta, consigue crear una base de datos con más de 500 observaciones:

```
input ///
sexo fuma caso n
1    0    0    200
1    0    1    32
1    1    0    50
1    1    1    62
2    0    0    220
2    0    1    12
2    1    0    38
2    1    1    28
end
expand n
ta fuma caso if sexo==1, row
ta fuma caso if sexo==2, row
```

La orden *expand* ***n*** advierte a STATA de que se desea ponderar por ***n***. Es decir, cada fila se repetirá tantas veces como vale ***n***. Aquí ha aparecido un hecho muy importante que ha de tenerse siempre en cuenta en STATA: *se deben escribir dos signos de igualdad (==),* y no uno solo, cuando lo que se desea indicar a STATA es una *condición.* Siempre que el igual sea *condicional,* debe escribirse por duplicado.

En las dos últimas órdenes se están pidiendo las tablas que cruzan ***fuma*** con ***caso,*** con la condición de que el sexo valga 1 (en la penúltima instrucción) y de que el sexo valga 2 (en la última instrucción).

En varias instrucciones de STATA existe una opción de añadir pesos, que pueden representar las frecuencias con que se observó cada valor **[fweight]**, o bien la probabilidad con que se quiere ponderar cada observación **[pweight]**, porque se trata de un muestreo. La primera no admite decimales, la segunda sí.

### 2.9.2. Ponderación en SPSS

Para ponderar en SPSS se debe seleccionar la opción que muestra la figura 2.20. Por omisión, el programa no pondera los casos. Si se desea ponderarlos, se deberá incluir en el recuadro correspondiente (v. fig. 2.20) la variable (***n*** en el ejemplo) que contiene los pesos para cada observación.

En cuanto a sintaxis, se escribe así:

```
WEIGHT BY n.
```

### 2.9.3. Ponderación en R

Primero introducimos datos originales (v. apartado 2.5.4), después asignamos los pesos a cada nota, luego generamos la nota ponderada y, por último, calculamos la media ponderada.

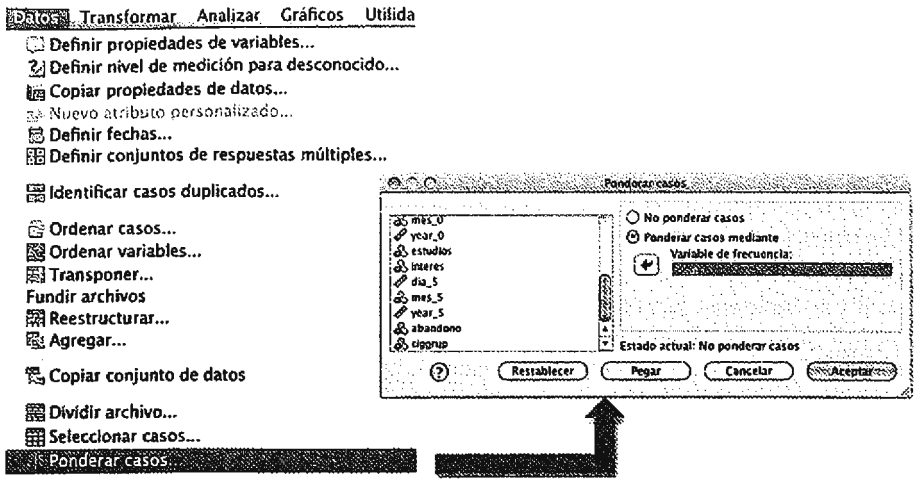



**Figura 2.20** Ponderar casos en SPSS.

```
> notas <- c(8,10,4)
> w <- c(25,25,50)
> nota_w <- rep(notas,w)
> mean(nota_w)
```

## 2.10. VALORES EXTREMOS *(OUTLIERS)* Y CONSEJOS SOBRE SU USO

Cuando se encuentre un *«outlier»*, siempre debería considerarse su origen. ¿Es legítimo un dato puntual cuyo valor es inusualmente grande o pequeño? ¿Es un valor mal registrado? ¿Es el resultado de un error o accidente en la experimentación? En los dos últimos casos, que corresponden a errores y que suelen ser los más frecuentes, pueden tomarse dos opciones:

1. La opción más correcta es averiguar concienzudamente el origen del error (si se tomó mal el dato, se apuntó erróneamente, hubo un fallo tipográfico al introducirlo en el ordenador, etc.) y corregirlo antes de seguir adelante con el análisis estadístico.
2. Si hay motivos fundados para sospechar que se trata de un error, pero resulta imposible corregirlo, debería borrarse ese dato del conjunto y completarse el análisis usando solo los datos restantes o bien aplicando procedimientos de imputación (v. apartado 19.5.3) para reemplazarlo.

Si el *outlier* no se debe a ningún error, sino que corresponde a un valor realmente *raro,* se sugiere que se dé a conocer la presencia del *«outlier»* y que los análisis estadísticos se realicen con y sin él. De esta forma, el experto en la materia puede tomar la decisión de incluir o no el *«outlier»* en futuros análisis. Si se decidiese incluirlo, se representarían en el diagrama de cajas como puntos, más allá de los bigotes.

## 2.11. PREFERENCIA DE NÚMEROS

Un caballo de batalla de la epidemiología es conseguir que las mediciones sean lo más precisas posibles. Sin embargo, eso choca con la tendencia natural del ser humano a redondear sus apreciaciones acerca de los números.

Se habla de preferencias de números o preferencias de dígitos cuando en la recogida de datos se impone el prejuicio —generalmente inconsciente— de recoger unos números que no son los

reales, sino los que prefiere el observador. Habitualmente ocurre con las cifras finales. Así, un observador que está tallando a un grupo de personas puede tener la tendencia a aproximar las alturas de cada sujeto al número par más cercano, y puede suceder que todas las tallas recogidas acaben en 0, 2, 4, 6 u 8. Un médico o una enfermera que toma la tensión a sus pacientes suele anotar que la tensión arterial diastólica es 80, 85 o 90 mmHg, pero puede que casi nunca anote una cifra de 82 mmHg o de 89 mmHg. Eso denota que esa persona no ha sido suficientemente instruida o no pone cuidado al recoger los datos.

Hay otras personas que tienen una especial predilección hacia un número en particular (el 7, el 9, o el que sea) y, sin darse cuenta, apuntan más veces ese número que otro cuando deben redondear una cifra.

Normalmente, estas preferencias por las cifras finales de los datos suelen tener cierta influencia en su tratamiento estadístico, ya que se pierde información. No obstante, esta mala influencia no es muy perjudicial. Este fenómeno se puede aprovechar con una utilidad insospechada: será posible verificar la *calidad* de los datos recogidos, ya que la preferencia de dígitos puede indicar indirectamente con qué cuidado y precisión se realizó la recogida de datos. Cuanto más se aproximen los valores unos a otros, mejor será normalmente la calidad de los datos. Si solo se encontrasen cifras acabadas en 0 o en 5 en una base de datos de tensión arterial, podría sospecharse que se puso poco cuidado en recoger adecuadamente esa variable.

## 2.12. ÍNDICES ESTADÍSTICOS DESCRIPTIVOS CON STATA

Con STATA podrían explorarse muchas posibilidades descriptivas. La más sencilla es la instrucción **summarize,** que, aplicada a la edad (base de datos ***canga25***), produciría lo siguiente:

```
summarize edad
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| edad | 25 | 50.92 | 18.54032 | 19 | 77 |

STATA proporciona así una descripción básica de una variable numérica con el número de observaciones (25), la media (50,92), la desviación estándar (18,54), el valor mínimo (19) y el máximo (77). Si se desease una información más amplia, se puede pedir la opción **detail,** que, como todas las opciones de STATA, debe ir después de una coma.

```
Su edad, detail
```

edad

| | Percentiles | Smallest | | |
|---|---|---|---|---|
| 1% | 19 | 19 | | |
| 5% | 20 | 20 | | |
| 10% | 21 | 21 | Obs | 25 |
| 25% | 41 | 22 | Sum of Wgt. | 25 |
| 50% | 54 | | Mean | 50.92 |
| | | Largest | Std. Dev. | 18.54032 |
| 75% | 66 | 71 | | |
| 90% | 72 | 72 | Variance | 343.7433 |
| 95% | 75 | 75 | Skewness | -.4598694 |
| 99% | 77 | 77 | Kurtosis | 1.980636 |

No está de más volver a insistir en que el valor nulo para la curtosis en STATA no es el 0, sino el 3. Por lo tanto, en este ejemplo, la curtosis es *negativa* (platicúrtica). La asimetría también es negativa. STATA incorpora automáticamente varios percentiles (1, 5, 10, 25, 50, 75, 90, 95 y 99) cuando se pide la opción **detail.** También ofrece los cuatro valores individuales menores y los cuatro mayores. Con respecto al tamaño de muestra, da una doble información: *Obs* presenta el número de filas (número de sujetos que existen sin aplicar ponderación); en cambio, *Sum of Wgt* es la suma de los pesos que se han aplicado para expandir (o reducir) la base de datos según esos pesos (v. apartado 2.9).

Otras dos órdenes interesantes en STATA son **tabstat** para variables numéricas y la ya vista **tabulate** para variables categóricas. Un ejemplo instructivo con **tabstat** sería:

```
tabstat peso, by(sexo) stat(n mean sd p50 min max sum)
```

```
Summary for variables: peso
     by categories of: sexo (sexo)
```

| sexo | N | mean | sd | p50 | min | max | sum |
|---|---|---|---|---|---|---|---|
| 1 | 19 | 71.26316 | 11.31293 | 71 | 48 | 90 | 1354 |
| 2 | 6 | 60 | 19.93991 | 53.5 | 47 | 100 | 360 |
| Total | 25 | 68.56 | 14.24512 | 68 | 47 | 100 | 1714 |

donde *N* es la frecuencia, *mean* es la media, *sd* la desviación estándar, *p50* la mediana, *min* y *max* los valores inferior y superior, y *sum* es la suma de todos los valores.

La orden **means** se usa para obtener la media aritmética, geométrica y armónica (junto con sus intervalos de confianza; v. capítulo 4):

```
means peso
```

| Variable | Type | Obs | Mean | [95% Conf. | Interval] |
|---|---|---|---|---|---|
| peso | Arithmetic | 25 | 68.56 | 62.67991 | 74.44009 |
| | Geometric | 25 | 67.13718 | 61.55891 | 73.22094 |
| | Harmonic | 25 | 65.72272 | 60.42489 | 72.03881 |

La orden **centile** calcula los percentiles. STATA, además, obtiene sus intervalos de confianza. Por ejemplo, para pedir los percentiles 25, 50 y 75 del peso de los varones se escribiría:

```
centile peso if sexo==1, centile(25 50 75)
```

| Variable | Obs | Percentile | Centile | — Binom. Interp. — [95% Conf. | Interval] |
|---|---|---|---|---|---|
| peso | 19 | 25 | 64 | 50.327 | 68.56658 |
| | | 50 | 71 | 66.08235 | 78 |
| | | 75 | 78 | 74.24456 | 90 |

## 2.13. PROCEDIMIENTOS DESCRIPTIVOS CON EXCEL

En el programa Excel, si se selecciona:

**Insertar → Función...**

aparecerá un menú (fig. 2.21) que ofrece múltiples posibilidades de solicitar índices estadísticos.

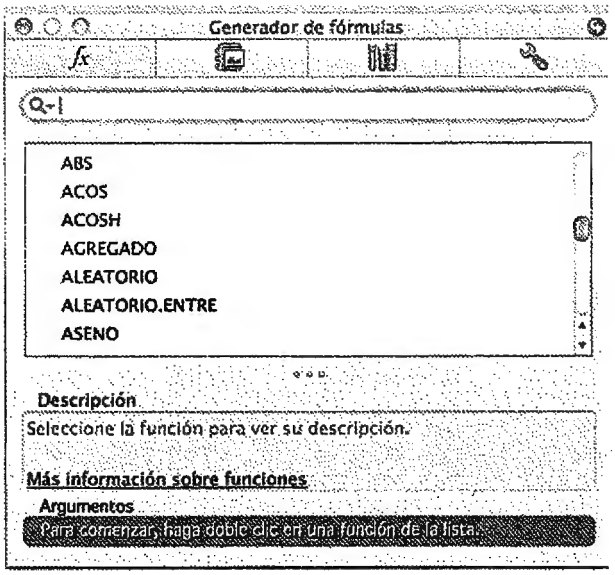



**Figura 2.21** Menú para seleccionar funciones en Excel. Aparecerá cuando se selecciona: **Insertar → Función...**

Cada una de estas funciones viene adecuadamente explicada en las múltiples ayudas y ventanas que ofrece este programa. Para que una función se ejecute se debe escribir su nombre en una casilla, pero siempre debe precederse del signo igual (=). Luego, se debe dar una indicación entre paréntesis de cuáles son las casillas en que están situados los datos. Por ejemplo, **=PROMEDIO (A1:A9)** significa que se pide la media aritmética de los nueve datos que ocupan las casillas A1, A2, A3, A4, A5, A6, A7, A8 y A9. Las funciones de mayor utilidad están recogidas en la tabla 2.7.

## 2.14. PROCEDIMIENTOS DESCRIPTIVOS CON OTROS PROGRAMAS

### 2.14.1. Funciones descriptivas en R

Con R/Splus, las instrucciones son parecidas a las de Excel, pero hay que tener en cuenta que hay que *hablarle* en inglés y que en vez de referirse a casillas se deben usar los nombres de las variables.



***Tabla 2.7 Funciones de mayor utilidad en Excel aplicadas a los nueve primeros dígitos***

| SE INTRODUCE | EXCEL DEVUELVE | VALOR |
|---|---|---|
| =PROMEDIO(A1:A9) | Media aritmética | 5 |
| =MEDIA.GEOM(A1:A9) | Media geométrica* | 4,147 |
| =MEDIA.ARMO(A1:A9) | Media armónica* | 3,181 |
| =MEDIANA(A1:A9) | Mediana | 5 |
| =MODA(A1:A9) | Moda** | #N/A |
| =VAR(A1:A9) | Varianza | 7,5 |
| =DESVEST(A1:A9) | Desviación estándar | 2,739 |
| =MIN(A1:A9) | Mínimo | 1 |
| =MAX(A1:A9) | Máximo | 9 |

*Se deben haber introducido antes los nueve valores (1, 2, 3, 4, 5, 6, 7, 8 y 9) en las casillas A1:A9.*
**No funcionará si hay algún valor negativo o igual a 0.*
***No funcionará si ningún valor se repite; en caso de una distribución multimodal, presentará el valor inferior.*

*Tabla 2.8 Funciones descriptivas con R sobre la variable días que contiene los nueve primeros dígitos*

| SE INTRODUCE | R/SPLUS DEVUELVE | VALOR |
|---|---|---|
| **> length(dias)** | Tamaño de muestra (n) | 9 |
| **> mean(dias)** | Media aritmética | 5 |
| **> median(dias)** | Mediana | 5 |
| **> y < -log(dias)** | Media geométrica | 4.147 |
| **> geom.mean < -exp(mean(y))** | | |
| **> geom.mean** | | |
| **> y <-(1/dias)** | Media armónica | 3.181 |
| **> n <-length(dias)** | | |
| **> harm.mean < -n/sum(y)** | | |
| **> harm.mean** | | |
| **> var(dias)** | Varianza | 7.5 |
| **> var(dias)^0.5** | Desviación estándar | 2.739 |
| **> min(dias)** | Mínimo | 1 |
| **> max(dias)** | Máximo | 9 |
| **> quantile(dias,c(0.25,0.5,0.75))** | Percentiles 25, 50 y 75 | 3 5 7 |

La tabla 2.8 recoge las principales funciones descriptivas disponibles en R, con el ejemplo que corresponde a estos datos:

```
> dias<-c(1,2,3,4,5,6,7,8,9)
```

R/Splus permite añadir funciones definidas por el usuario con la orden **function.** El programa guardará esa nueva función para otras ocasiones. Por ejemplo, para crear una función que calcule directamente medias geométricas, primero se escribe:

```
> media.geom<-function(x){exp(mean(log(x)))}
```

Luego, cada vez que se escriba la nueva función seguida del nombre de una variable entre paréntesis, por ejemplo:

```
> media.geom(dias)
```

el programa devolverá su media geométrica:

```
[1] 4.147166
```

Otra función interesante que se puede dejar programada es el error estándar de la media (eem):

```
> eem<-function(x){(var(x)/length(x))^0.5}
```

### 2.14.2. Funciones descriptivas en SPSS

Casi todas las medidas de tendencia central en SPSS están situadas en:

**Analizar → Estadísticos Descriptivos**

La opción más usada es:

**Analizar → Estadísticos Descriptivos → Frecuencias...**

Esta opción ofrece un primer menú, donde se selecciona la variable de interés: por ejemplo, ***edad.*** Si luego se pulsa el botón:

**Estadísticos...**

aparecerá la figura 2.22 en el momento en que se habían seleccionado (cuando se hizo la captura de pantalla) las tres opciones de medidas de tendencia central (media, mediana y moda).

Después se seleccionaron otras (cuartiles, asimetría y curtosis, etc.). Finalmente se pulsa:

**Continuar → Aceptar**

o bien:

**Continuar → Pegar**

(Esta opción «Pegar» es la adecuada si lo que se desea es seguir trabajando con sintaxis.)

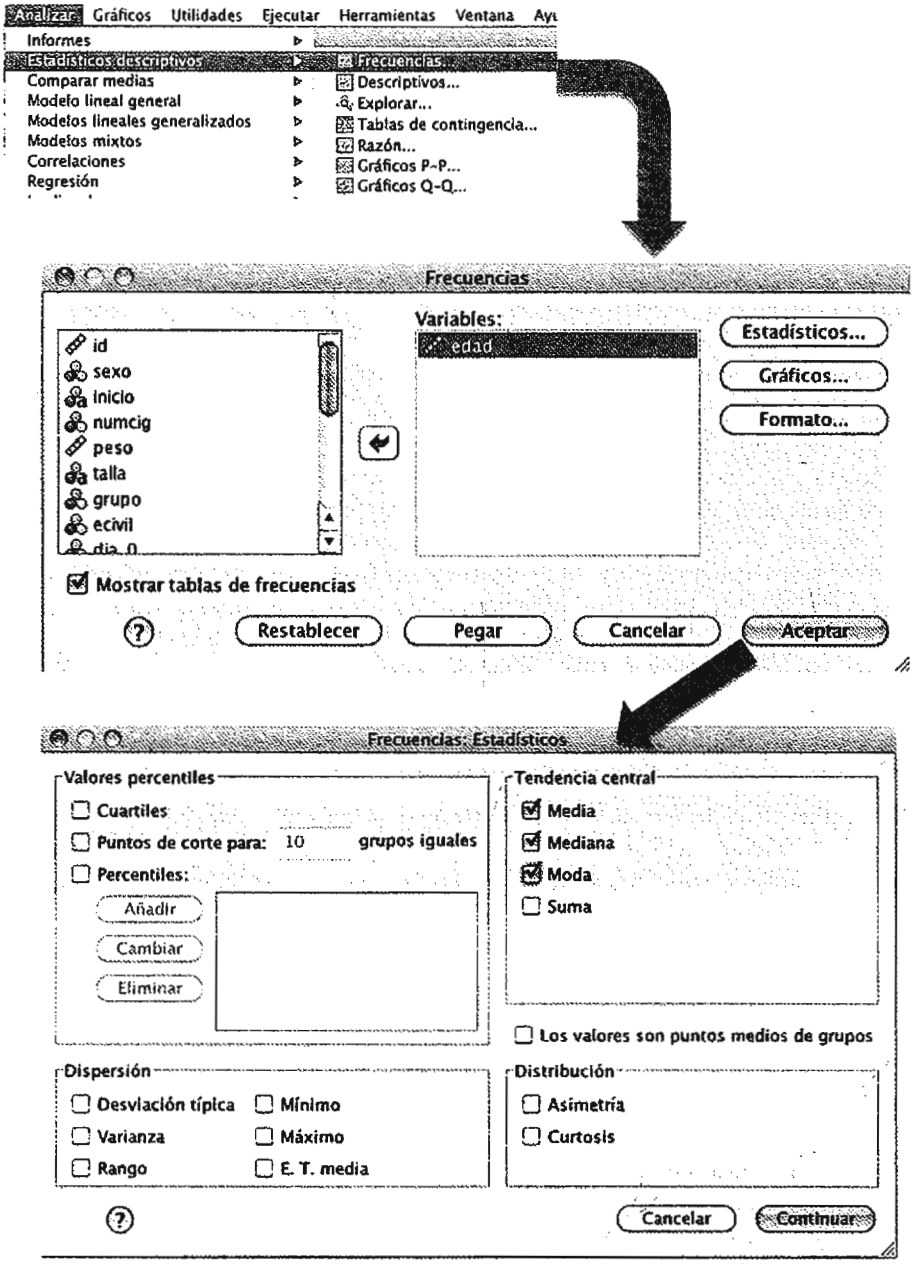



**Figura 2.22** Estadísticos descriptivos con SPSS.

El resultado que proporcionará el programa (una vez seleccionadas más opciones) es el que aparece el cuadro 2.2.

Se observa que la media aritmética de la edad es de 50,92 años. Se han utilizado 25 observaciones para calcularla. Como puede apreciarse, si se comparan estos resultados con los obtenidos en STATA, los programas de ordenador difieren a veces en sus resultados con muestras pequeñas para la asimetría

**CUADRO 2.2 MEDIDAS DE FRECUENCIA OBTENIDAS CON SPSS**

**Estadísticos**

**Edad**

N:
- Válidos: 25
- Perdidos: 0

Media: 50,92
Error típico de la media: 3,708
Mediana: 54,00
Moda: 51*
Desviación típica: 18,540
Varianza: 343,743
Asimetría: –0,490
Error típico de asimetría: 0,464
Curtosis: –0,972
Error típico de curtosis: 0,902
Mínimo: 19
Máximo: 77
Percentiles:
- 25: 36,00
- 50: 54,00
- 75: 66,50

**Existen varias modas. Se mostrará el menor de los valores.*

y los percentiles. SPSS siempre suele dar información sobre cuántos valores son válidos y cuántos se han perdido («Válidos», «Perdidos»), es decir, eliminando los datos faltantes o inválidos. Los valores inválidos son casillas que tienen el valor de la columna edad en blanco o que contienen una información que no corresponde a los valores que se hayan definido *a priori* como posibles para la edad.

Siempre es mejor dar los resultados de la media (y, en general, de la mayor parte de los resultados finales de una estadística) con pocos decimales, los que sean oportunos. No tendría sentido decir que la edad media de unos pacientes es 52,133 años. ¡Eso supondría que nos interesa separar edades que se diferencian no en horas, sino en minutos! En esto ha ido mejorando SPSS y otros programas, con las sucesivas versiones, ya que van redondeando el resultado de los índices descriptivos para presentar solo lo que es más razonable en la escala de medición empleada.

Al pedir la asimetría y la curtosis a SPSS aparecen dos nuevos índices que no se habían mencionado antes (ni se habían solicitado): el error estándar de la curtosis y el error estándar de la asimetría. ¿Para qué sirven? De momento puede decirse que, si el valor absoluto de la curtosis no llega a ser el doble de su error estándar, puede pensarse que la desviación de la curtosis con respecto a su valor nulo no es importante y se puede asumir que es prácticamente igual a 0 (en SPSS), es decir, normocúrtica. Lo mismo puede asumirse con el coeficiente de asimetría: si la asimetría es menor que dos veces su error estándar, puede considerarse nula desde el punto de vista práctico y asumir que la distribución es simétrica. De todos modos, esto hay que tomarlo con muchas reservas cuando el tamaño de muestra es pequeño (<30), ya que entonces los errores estándar suelen ser excesivamente grandes.

Cuando se usa SPSS, resulta difícil encontrar las medias geométricas y armónicas, pues no están en el menú de rutina, pero se pueden pedir con la sintaxis:

```
OLAP CUBES edad
 /CELLS=HARMONIC GEOMETRIC.
```

Es mejor hacerlo con sintaxis que con menús, ya que el menú (**Analizar → Informes → Cubos OLAP...**) está programado para pedir estos índices estadísticos separados por grupos.

## 2.15. DATOS TRUNCADOS O CENSURADOS

En ocasiones, sobre algunos datos o medidas solo se puede sospechar su valor, sin saberlo realmente con precisión. Así, cuando solo se sabe que un valor está situado más allá de un cierto límite, se clasificará tal observación como *truncada* (en inglés, *censored*, que, a veces, se traduce por *censurada*). Esto sucede, por ejemplo:

- Al estudiar la duración de la supervivencia de una serie de pacientes, se habrá observado en cada uno un determinado tiempo de vida, por ejemplo, desde el diagnóstico hasta la muerte. Pero, gracias a Dios, habrá pacientes cuyo tiempo de supervivencia no se pueda conocer con exactitud, ya que no habrán fallecido al final del estudio. Sin embargo, sí se sabe con seguridad que su supervivencia va a ser *superior* al tiempo de observación con el que hasta ahora se cuenta, pues estaban vivos al final del período de observación. Este tipo de datos se refiere como *datos truncados o censurados por la derecha.*
- Por otro lado, a veces, cuando se determinan los niveles en sangre de una determinada sustancia que suele estar presente en cantidades mínimas, habrá pacientes que presentarán niveles tan bajos que no sean medibles porque resultan indetectables con los aparatos de medida convencionales. Aunque se diga que la concentración de la sustancia es «no detectable», no debe interpretarse que la concentración sea exactamente igual a 0. Pueden existir razones biológicas para pensar que la concentración de esa sustancia en la sangre debe ser, en todo caso, ligeramente mayor que 0. A este tipo de datos se les llama *truncados o censurados por la izquierda.*

Ambos tipos de censura, la de derechas y la de izquierdas, requieren un tratamiento estadístico especial.

## 2.16. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Objetivo | STATA | SPSS |
|---|---|---|
| Recodificar | `recode numcig 1/19=1 /// 20/39=2 40/max=3, /// generate(ciggrup)` | Transformar → Recodificar en distintas variables |
| Describir | `summarize numcig, detail` | Analizar → Estadísticos descriptivos... → Frecuencias |
| Tabular | `tabulate numcig` | |
| Tablas de contingencia | `tabulate numcig ciggrup` | Analizar → Estadísticos descriptivos... → Tablas de contingencia |
| Gráfico de sector | `graph pie, over(sexo)` | **GRAPH /PIE = COUNT BY sexo.** |
| Histograma | `histogram numcig, /// percent /// start(5) width(10)` | **GRAPH /HISTOGRAM = numcig.** |
| Cajas según una sola variable | `graph box peso` | **EXAM numcig / PLOT = BOXPLOT .** |
| Cajas según otra variable | `graph box peso, over(sexo)` | **EXAM peso /PLOT = BOXPLOT /PANEL COLVAR = sexo.** |

## REFERENCIAS

1. Altman DG, Bland JM. Statistics notes: variables and parameters. BMJ 1999;318(7199):1667.
2. Greenhalgh T. Statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ 1997;315(7104):364-6.
3. Canga N, De Irala J, Vara E, Duaso MJ, Ferrer A, Martínez-González MA. Intervention study for smoking cessation in diabetic patients: a randomized controlled trial in both clinical and primary care settings. Diabetes Care 2000;23(10):1455-60.
4. Paul CA, Au R, Fredman L, Massaro JM, Seshadri S, Decarli C, et al. Association of alcohol consumption with brain volume in the Framingham study. Arch Neurol 2008;65(10):1363-7.
5. Greenland S. Analysis of polytomous exposures and outcomes. En: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008. p. 303-4.
6. Jolley D. The glitter of the t table. Lancet 1993;342(8862):27-9.
7. Gladwell M. Outliers: the story of success. New York: Little, Brown and Company; 2008.
8. Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996; 313(7066):1200.
9. Martínez-González MA, García-Arellano A, Toledo E, Salas-Salvadó J, Buil-Cosiales P, Corella D, et al. A 14-item Mediterranean diet assessment tool and obesity indexes among high-risk subjects: the PREDIMED trial. PLoS One 2012;7(8):e43134.

# PROBABILIDAD. DISTRIBUCIONES DE PROBABILIDAD

3

*E. Toledo, A. Sánchez-Villegas, M. Á. Martínez-González*

## 3.1. INTRODUCCIÓN

Habitualmente, los investigadores trabajan solo con una muestra, pero desean obtener conclusiones válidas y extensibles para una población compuesta por todos los individuos (universo o población diana) que presenten esas características. En el fondo, el concepto de población en estadística acaba apuntando hacia algo infinito. Para poder dar el paso de la muestra a la población (inferencia) es preciso utilizar conceptos de probabilidad.

## 3.2. CONCEPTOS DE PROBABILIDAD

Hay dos tendencias principales al concebir la probabilidad (**1-3**). La definición frecuentista mantiene que las probabilidades son los límites a los que tiende la proporción o frecuencia relativa con la que sucede un acontecimiento cuando el número de observaciones tiende a infinito. Por ejemplo, si se lanza una moneda un número muy elevado de veces, la frecuencia relativa con la que saldrá cara tenderá a ser del 50% (probabilidad = 0,5). Dado que las frecuencias relativas son observables empíricamente, se puede pensar que los límites a los que tienden pueden proponerse como propiedades de los sistemas o mecanismos que generan las secuencias de acontecimientos (**1**). Por eso, a esta concepción corresponde el concepto de la probabilidad como algo *objetivo* (probabilidad *física*). En su forma más pura, el pensamiento frecuentista niega todo significado a la probabilidad *individual* de un acontecimiento singular, ya que, al no engendrar una repetición, no es posible estimar empíricamente su frecuencia relativa. Este problema podría presentarse al aplicar la probabilidad al paciente individual, ya que «no hay enfermedades, sino enfermos». La respuesta frecuentista consiste en imputarle a cualquier paciente lo que ha ocurrido previamente con pacientes similares y siempre va acompañada de un cierto grado de incertidumbre, pero prescindiendo del hecho de que no hay dos seres humanos iguales.

La otra filosofía es la *bayesiana*, que maneja la probabilidad como algo *subjetivo*, es decir, el grado de certeza que se posee sobre un asunto, la «credibilidad» o la probabilidad personal. En el mundo de la estadística, los frecuentistas son más abundantes que los bayesianos (**4**). A los bayesianos se les acusa de cierto carácter arbitrario o irracional en sus probabilidades subjetivas (que se suelen llamar *a priori*), ya que estas probabilidades no son directamente medibles. También se podría decir que la verificación frecuentista de la probabilidad empírica nunca ha sido hecha. Nadie ha lanzado una moneda al aire *infinitas* veces.

Lo que sucede es que a veces se aplica un concepto teórico, deductivo, en vez de una inferencia empírica o inductiva de probabilidad, como muestra la figura 3.1.

### 3.2.1. Estimación teórica: ley de Laplace

En muchos casos, los distintos valores que una variable puede tomar suelen ser *equiprobables*, es decir, tienden a ocurrir con la misma frecuencia unos que otros. Así, al lanzar una moneda al aire se espera teóricamente obtener un 50% de cruces. La ley de Laplace establece que la probabilidad

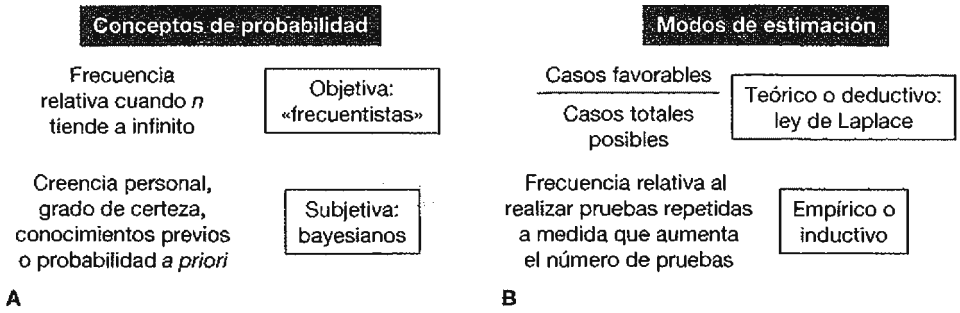



**Figura 3.1** Dos corrientes de pensamiento sobre la probabilidad y dos modos de hallar la probabilidad de un suceso.

de obtener cada resultado será igual al *cociente entre casos favorables y casos posibles*. Para hallar, por ejemplo, la probabilidad de que al lanzar un dado salga un número par habrá:

casos favorables = 3 (números pares)

casos posibles = 6 (todos los posibles)

probabilidad (par) = 3/6 = 0,5.

Sin embargo, debido al azar, en la práctica esto no siempre es así.

### 3.2.2. Modo empírico de hallar la probabilidad

Los *frecuentistas* consideran que la frecuencia relativa con que ocurre un suceso al realizar las pruebas en condiciones similares tiende a un valor de probabilidad a medida que aumenta el número de pruebas.

La práctica totalidad de las probabilidades que se manejan en ciencias de la vida se han obtenido por el método empírico. ¿Cuál es la probabilidad de que un paciente con cáncer de pulmón sobreviva 5 años? Esta probabilidad se determinará a partir de lo observado en pacientes con las mismas características hasta la fecha (y de los que se hayan recogido datos).

## 3.3. AXIOMAS Y PROPIEDADES DE LA PROBABILIDAD

### 3.3.1. Primer axioma

El valor de la probabilidad estará siempre comprendido en el intervalo [0,1], es decir, siempre será superior o igual a 0 e inferior o igual a 1 (fig. 3.2) (**1**).

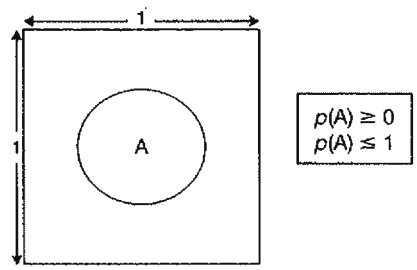



**Figura 3.2** Primer axioma de la probabilidad.

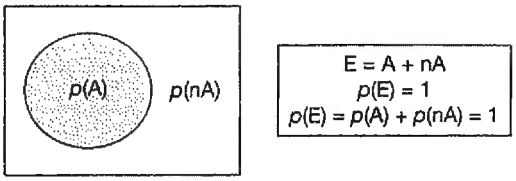



**Figura 3.3** Segundo axioma de la probabilidad. E, suceso seguro; *p*(nA), probabilidad de que A no ocurra (suceso complementario).

### 3.3.2. Segundo axioma

El suceso seguro tiene una probabilidad igual a la unidad, es decir, es aquel que ocurre en el 100% de las ocasiones[1] y se cumplirá necesariamente (p. ej., el que englobe como posibles resultados los seis lados de un dado) (fig. 3.3).

### 3.3.3. Tercer axioma

Dos sucesos A y B son *excluyentes o incompatibles*, por ejemplo, la «cara» y la «cruz» de una moneda, cuando no pueden acontecer simultáneamente. La probabilidad de que ocurra alguno de ambos sucesos, *p*(A∪B) en lenguaje matemático o bien *p* (A o B) en lenguaje convencional, es decir, que se cumpla bien un suceso o bien el otro, será igual a la suma de las probabilidades de cada uno por separado. En esto consiste la *propiedad aditiva* de la probabilidad.

$$p(A \text{ o } B) = p(A \cup B) = p(A) + p(B) \qquad \text{Si } A \cap B = \varnothing$$

donde *p*(A∪B) es la probabilidad de que ocurra *cualquiera* de los dos sucesos *(unión)* y A∩B es la intersección o coincidencia de *ambos* (fig. 3.4).

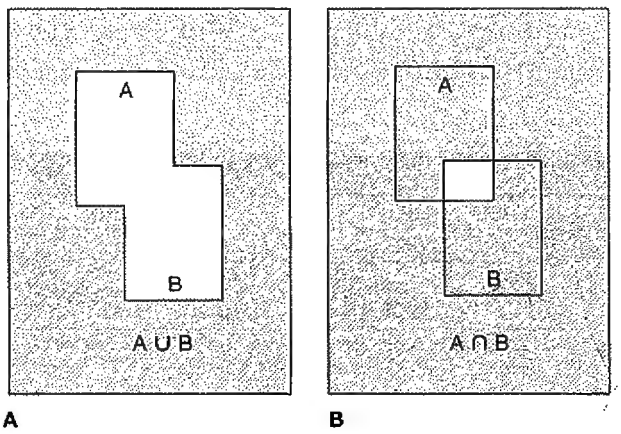



**Figura 3.4** Unión e intersección de sucesos. **A.** Unión: «o». **B.** Intersección: «y».

1 En lenguaje coloquial, las probabilidades suelen expresarse en tantos por ciento, mientras que al enunciar aquí formalmente los axiomas de la probabilidad se expresan en tantos por uno. Cuando la probabilidad se expresa como porcentaje, se diría que sus límites son el 0 y el 100%.

De estos axiomas se deducen las siguientes propiedades de la probabilidad:

- La probabilidad tiene siempre valores comprendidos entre 0 y 1: $0 \leq p(A) \leq 1$.
- La probabilidad del suceso complementario es igual a 1 menos la probabilidad del suceso inicial.

$$p(nA) = 1 - p(A)$$

- La probabilidad del suceso imposible es 0.
- Si un suceso está incluido en otro, la probabilidad del primero ha de ser menor o igual a la del segundo.
- Cuando los sucesos son incompatibles:
  - La probabilidad de su intersección es 0.
  - La probabilidad de su unión es la suma de las probabilidades de ambos:

$$p(A \cup B) = p(A) + p(B) \text{ (incompatibles)}$$

- Cuando los sucesos son compatibles:
  - La probabilidad de su intersección es mayor de 0.
  - La probabilidad de su unión es la suma de las probabilidades de ambos *menos* la probabilidad de la intersección:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) \text{ (compatibles)}$$

Por ejemplo, en una población de 1.000 sujetos con sospecha de sida se evalúan dos factores de riesgo: múltiples parejas sexuales (MPS) y uso de drogas parenterales (UDP). Habían estado expuestos a MPS 540 pacientes y a UDP 410. No habían estado expuestos a ningún factor 200 pacientes. Estar expuesto a MPS y a UDP no son hechos incompatibles. Se dan los siguientes datos:

$$p(\text{MPS}) = \frac{540}{1.000}; \qquad p(\text{UDP}) = \frac{410}{1.000}$$

$$p(\text{noMPS} \cap \text{noUDP}) = \frac{200}{1.000}$$

Podrá construirse una tabla sumando y restando a partir de estos datos, como muestra la figura 3.5.

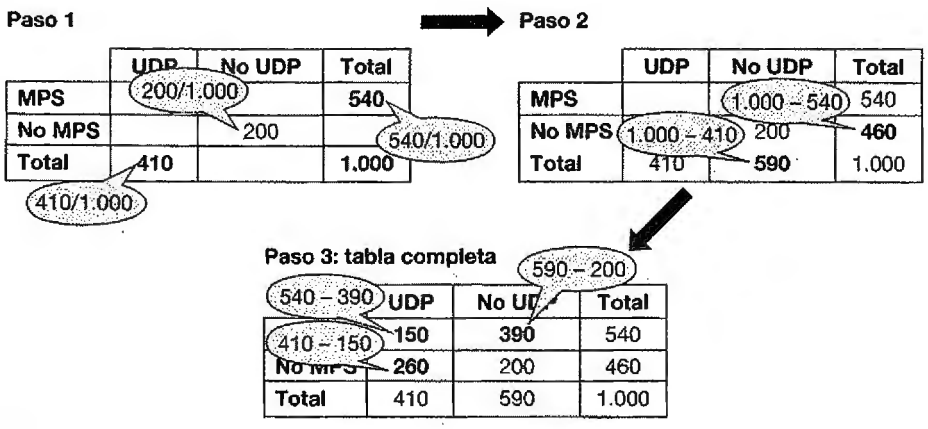
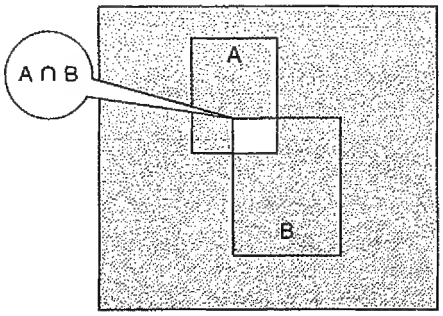



**Figura 3.5** Cómo construir una tabla a partir de datos de probabilidades.




$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

**Figura 3.6** Unión de sucesos compatibles.

Una vez que se obtiene la tabla completa, puede aplicarse la fórmula anterior para sucesos compatibles:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$p(MPS \cup UDP) = p(MPS) + p(UDP) - p(MPS \cap UDP)$$

$$\frac{150 + 390 + 260}{1.000} = \frac{540}{1.000} + \frac{410}{1.000} - \frac{150}{1.000} = \frac{800}{1.000}$$

Se cumple esta igualdad. También puede verse gráficamente, como se representa en la figura 3.6. Téngase en cuenta que, para sucesos incompatibles, la probabilidad de su intersección es 0.

## 3.4. CONCEPTO DE INDEPENDENCIA

Imagínese que se hizo una encuesta a 300 personas, 100 mujeres y 200 hombres, y se les preguntó si eran fumadores activos. Los resultados serían los de la tabla 3.1.

La probabilidad *marginal* de encontrar un fumador en el total de la muestra sería del 20%: $p$(fumar) = 0,2. Esta probabilidad sería idéntica para hombres —$p$(fumar) = 40/200 = 0,2— y mujeres —$p$(fumar) = 20/100 = 0,2—. Como esta probabilidad es la misma para hombres y para mujeres, se dirá que la probabilidad de fumar es *independiente* del sexo. La independencia entre dos variables implica que la información recibida sobre una variable no sirve en absoluto para predecir la otra (5).

## 3.5. PROBABILIDAD CONDICIONADA

Otro concepto importante es el de *probabilidad condicionada.*

Si se valorase la presencia de glucosa en orina (glucosuria) en pacientes diabéticos y en pacientes sin diabetes, podrían obtenerse los resultados de la tabla 3.2.

*Tabla 3.1 Hábito de fumar en una muestra según sexo*

| | FUMAN | NO FUMAN | TOTAL |
|---|---|---|---|
| **Mujeres** | 20 | 80 | **100** |
| **Hombres** | 40 | 160 | **200** |
| **Total** | **60** | **240** | **300** |

*Tabla 3.2 Presencia de glucosuria en pacientes diabéticos y sin diabetes*

| | DIABETES | NO DIABETES | TOTAL |
|---|---|---|---|
| Glucosuria | 60 | 8 | 68 |
| No glucosuria | 140 | 792 | 932 |
| Total | 200 | 800 | 1.000 |

En total se examinó a 1.000 pacientes, 200 de los cuales eran diabéticos y 800 no diabéticos. La probabilidad de que un diabético presente glucosuria —probabilidad de glucosuria condicional a diabetes, $p$(glucosuria | diabetes)— en esta muestra se puede calcular dividiendo el número de diabéticos con glucosuria (60) entre el total de diabéticos (200) y valdrá 0,3; es decir, el 30% de los diabéticos presentan glucosuria. En cambio, la probabilidad de que un *no diabético* presente glucosuria en esta muestra se hallará dividiendo el número de no diabéticos con glucosuria (8) entre el total de no diabéticos (800) y valdrá 0,01; es decir, solo el 1% de los no diabéticos presentan glucosuria. Ahora ya no se puede decir que exista independencia. Tener glucosuria *depende* de ser diabético, pues es 30 veces más frecuente encontrar glucosuria entre los diabéticos (30%) que entre los no diabéticos (1%). Si se posee información de un suceso (ser diabético), se puede adelantar que la probabilidad del segundo suceso (tener glucosuria) será mayor.

La probabilidad *condicionada* se puede definir como el cociente entre los casos favorables y los casos posibles dentro de aquellos que cumplen una condición. Es la probabilidad de ocurrencia de un fenómeno *dentro de un subgrupo*. La notación que indica cuál es la condición se expresa con una barra vertical (|). La probabilidad de presentar glucosuria *condicionada a ser diabético* será:

$$p(\text{glucosuria}|\text{diabetes}) = \frac{\text{casos con glucosuria y diabetes}}{\text{diabéticos}} = \frac{p(\text{glucosuria} \cap \text{diabetes})}{p(\text{diabetes})} = \frac{60/1.000}{200/1.000} = 0,3$$

Se cumple que:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Es decir, «la probabilidad de A condicionada al resultado ya ocurrido de B es igual a la probabilidad de la coincidencia de ambos sucesos dividida entre la probabilidad de que ocurra B». Lo opuesto a la probabilidad condicionada es la probabilidad marginal (o no condicionada, o probabilidad total). La probabilidad marginal de presentar glucosuria es:

$$p(\text{glucosuria}) = \frac{\text{casos con glucosuria}}{\text{total pacientes}} = \frac{68}{1.000} = 0,068$$

Solo si hay independencia coinciden las probabilidades condicionales y marginales.

Las tablas pueden representarse como árboles de probabilidad, como en la figura 3.7, donde los números corresponden a las probabilidades condicionadas al suceso que aparece inmediatamente antes (a la izquierda en el árbol). Por ejemplo, la probabilidad de ser diabético es 0,2 (es una probabilidad no condicionada). Se calcularía dividiendo 200/1.000. La probabilidad de no ser diabético es 0,8. Se cumple siempre que la suma de las probabilidades de las ramas que salen del mismo punto debe ser 1. Pueden irse calculando cada una de las probabilidades. Por ejemplo, la probabilidad de presentar glucosuria condicionada a ser diabético se calcularía dividiendo 60/200 = 0,3.

Para calcular las probabilidades de *intersección* de dos sucesos hay que ir multiplicando las probabilidades de cada rama hasta llegar al extremo del árbol que se desee. Recuérdese que la

A

| | Diabetes | No diabetes | Total |
|---|---|---|---|
| Glucosuria | 60 | 8 | 68 |
| No glucosuria | 140 | 792 | 932 |
| Total | 200 | 800 | 1.000 |

B

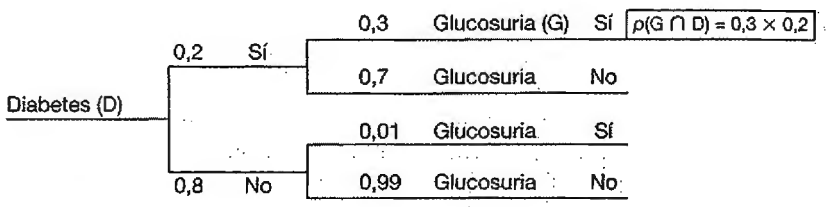



**Figura 3.7** Transformación de una tabla en árbol de probabilidades. **A.** Presencia de glucosuria en diabéticos y no diabéticos, presentado como tabla. **B.** Probabilidades de glucosuria en diabéticos y no diabéticos, presentado como árbol de probabilidad.

intersección es la ocurrencia *simultánea* de dos sucesos A y B, y se expresa como $p(A \cap B)$. Se expresó antes la siguiente equivalencia:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Por lo tanto:

$$p(A \cap B) = p(A \mid B) \times p(B)$$

Por ejemplo, la probabilidad de presentar simultáneamente diabetes y glucosuria, $p(G \cap D) = p(G \mid D) \times p(D)$, será $0,3 \times 0,2 = 0,06$.

Puede comprobarse que sucede así, ya que hay 60 diabéticos con glucosuria entre un total de 1.000 pacientes y $60/1.000 = 0,06$. Puede hacerse el ejercicio de calcular las siguientes probabilidades.

- Probabilidad de no presentar glucosuria y ser diabético: $p(\text{noG} \cap D)$.
- Probabilidad de no presentar glucosuria y no ser diabético: $p(\text{noG} \cap \text{noD})$.

Siempre se cumplirá la relación vista anteriormente: $p(A \cap B) = p(A \mid B) \times p(B)$, que se lee así: «la probabilidad de la intersección es igual a la probabilidad condicionada multiplicada por la probabilidad de la condición». En esto consiste *la ley multiplicativa de la probabilidad* o *ley de la probabilidad compuesta*, que establece que la probabilidad de que dos sucesos ocurran simultáneamente será igual a la probabilidad de uno de ellos, por la probabilidad condicionada del otro al resultado del primero. Dicho en verso:

La probabilidad de la intersección
es igual a la probabilidad condicionada
por la probabilidad de la condición.

Si se trata de dos sucesos independientes, la probabilidad condicionada $p(A \mid B)$, sería igual a la probabilidad marginal o total de $p(A)$, al no tener influencia el resultado de un suceso sobre el otro. Así, la probabilidad de la intersección se simplifica en:

$$p(A \cap B) = p(B) \times p(A) \text{ (para sucesos independientes)}$$

## 3.6. INVERSIÓN DE LAS CONDICIONES: TEOREMA DE BAYES

La llamada *fórmula de inversión de las condiciones* ayuda a definir el *teorema de Bayes,* que ha dado lugar a una nueva filosofía en la interpretación y metodología del análisis de datos (**2,3,6,7**).

En el ejemplo de la glucosuria y la diabetes, se asume que la probabilidad de que cualquier paciente que acuda a nuestra consulta presente diabetes es de 0,2, ya que el 20% de la muestra eran diabéticos. Por otro lado, la probabilidad de que un no diabético presente glucosuria es de 0,01, P(G | noD) = 0,01. Si se sabe que la probabilidad condicionada de presentar glucosuria en un diabético es de 0,3, P(G | D) = 0,3, se puede hallar lo que resulta verdaderamente interesante para un médico, que es la probabilidad de que un paciente sea diabético si se sabe que presentó glucosuria, es decir, P(D | G). En esto consiste la *inversión de las condiciones.*

Si se observa la figura 3.8, parece lógico pensar que la probabilidad de padecer glucosuria P(G) valdrá la suma de las probabilidades de tener glucosuria y diabetes, $P(G \cap D)$, más la de tener glucosuria y no tener diabetes, $P(G \cap nD)$:

$$P(G) = P(G \cap D) + P(G \cap nD)$$

que, por la ley multiplicativa, se convierte en la expresión (3.1):

$$P(G) = P(G \mid D) \times P(D) + P(G \mid nD) \times P(nD) \qquad (3.1)$$

La expresión (1) se conoce como *teorema de la probabilidad total,* y en ella aparece la probabilidad de G condicionada a D, P(G | D). ¿Cómo calcular la expresión inversa, P(D | G)? Según los conceptos de probabilidad condicionada:

$$P(D \mid G) = \frac{P(G \cap D)}{P(G)}$$

El numerador de esa expresión vale $P(G \cap D) = P(G \mid D) \times P(D)$ y el denominador corresponde a la expresión (3.1). Por lo tanto, para invertir las condiciones y calcular P(D | G):

$$P(D \mid G) = \frac{P(G \mid D) \times P(D)}{\left[P(G \mid D) \times P(D)\right] + \left[P(G \mid nD) \times (P(nD)\right]}$$

y así:

$$P(D \mid G) = \frac{0,3 \times 0,2}{(0,3 \times 0,2) + (0,01 \times 0,8)} = 0,882$$

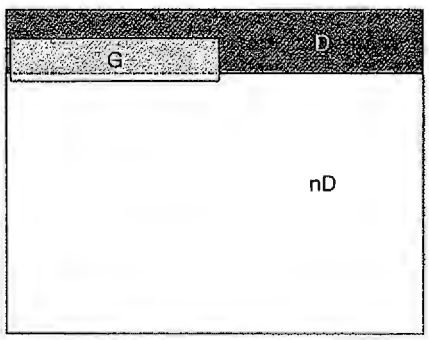



**Figura 3.8** Representación gráfica de las probabilidades de diabetes y glucosuria. D, diabetes; G, glucosuria; nD, no diabetes.

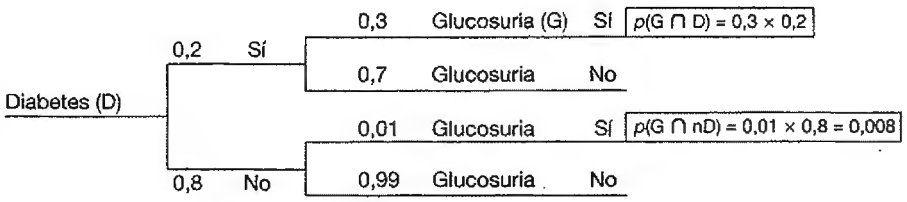



**Figura 3.9** Diagrama de árbol con las probabilidades de la intersección de dos sucesos.

Con esto se puede concluir que la probabilidad de que un paciente del que solo se sabe que tiene glucosuria sea diabético es de 0,882, es decir, de un 88,2%. Esto es comprobable en la tabla 3.2 o en las figuras 3.6 y 3.7, ya que 60/68 = 0,882.

Lo anterior se puede entender mejor con unos diagramas de árbol (fig. 3.9). Si la primera división del árbol viene dada por la diabetes (diabetes sí/diabetes no), la probabilidad de presentar glucosuria y simultáneamente diabetes será de 0,06 y la de presentar glucosuria y no ser diabético será de 0,008.

Si ahora se mira el árbol empezando por la glucosuria, se puede afirmar que, en total, la probabilidad de tener glucosuria es 0,06 + 0,008 = 0,068. Pues bien, de esa probabilidad de 0,068, una fracción que vale 0,06 pertenece a los diabéticos y la otra, de 0,008, pertenece a los no diabéticos. Por tanto, la probabilidad de ser diabético condicionada a tener glucosuria será de 0,06 dividido entre 0,068, que es 0,882. Este resultado coincide con lo demostrado antes teóricamente.

Otro modo (mucho más fácil de entender) de aplicar el teorema de Bayes consiste en preparar una sencilla tabla 2 × 2, imaginando que existiera, por ejemplo, un total de 1.000 individuos. Si se aplican las probabilidades conocidas, habría 200 diabéticos (de ellos 60 con glucosuria) y 800 no diabéticos (8 con glucosuria). Se debe multiplicar 1.000 por la probabilidad de ser diabético (1.000 × 0,2 = 200), después multiplicar 200 por la probabilidad de tener glucosuria condicionada a ser diabético (200 × 0,3 = 60) y, por último, multiplicar 800 por la probabilidad de tener glucosuria condicionada a no ser diabético (800 × 0,01 = 8). Los datos serían los que antes se comentaron, que se presentan de nuevo en la tabla 3.3.

Una vez preparada la tabla es fácil responder a todas las preguntas. Por ejemplo, la probabilidad de ser diabético condicionada a tener glucosuria sería:

$$P(D \mid G) = \frac{60}{68} = 0{,}882$$

y la probabilidad de no ser diabético condicionada a carecer de glucosuria sería:

$$P(nD \mid nG) = \frac{792}{932} = 0{,}85$$

En epidemiología clínica, a la probabilidad de presentar glucosuria condicionada a ser diabético se le llama *sensibilidad*, la probabilidad de no presentar glucosuria condicionada a no ser diabético recibe el nombre de *especificidad*, la probabilidad de ser diabético condicionada a tener glucosuria se conoce como *valor predictivo positivo* y la probabilidad de no ser diabético condicionada a carecer de glucosuria se denomina *valor predictivo negativo*.

***Tabla 3.3 Presencia de glucosuria en pacientes diabéticos y sin diabetes***

| | DIABETES | NO DIABETES | TOTAL |
|---|---|---|---|
| **Glucosuria** | 60 | 8 | 68 |
| **No glucosuria** | 140 | 792 | 932 |
| **Total** | 200 | 800 | 1.000 |

## 3.7. FACTOR BAYES PARA RELACIONAR LA *ODDS* PRETEST CON LA *ODDS* POSTEST

En primer lugar, hay que conocer el concepto de *odds*, que resulta más fácil de entender que de definir. Si el porcentaje de aprobados en un examen es del 75%, la *odds* de aprobar será 3, porque habrá tres aprobados por cada suspenso.

Si en un estudio que incluye a 1.000 pacientes, solo 200 de ellos padecen diabetes, la *odds* de desarrollar diabetes se calcularía dividiendo 200 entre 800 (*odds* = 1/4). Es decir, la *odds* es el cociente entre quienes tienen la característica de interés y quienes no la tienen. En una proporción, el numerador está incluido en el denominador; en la *odds*, el numerador **no** se incluye en el denominador. Una *odds* es, por tanto, la probabilidad *(p)* dividida por el complementario de la probabilidad $(1 - p)$:

$$Odds = \frac{p}{1-p}$$

También podría expresarse la *odds* como 1:4. Se interpreta que apareció un diabético por cada cuatro no diabéticos:

$$Odds = \frac{\text{diabéticos}}{\text{no diabéticos}} = \frac{200}{1.000} = \frac{1}{4}$$

Para transformar una *odds* en una proporción, se divide la *odds* por (1 + *odds*):

$$p = \frac{odds}{1+odds}$$

En el ejemplo de los diabéticos, $p = 0{,}25/(1 + 0{,}25) = 0{,}2$.

Se demuestra que la *odds a posteriori* (una vez que se sabe que se ha cumplido una condición) es igual a la *odds* marginal (no condicionada o *previa* a saber la condición), multiplicada por un factor, el «factor Bayes» (**8-9**). En la tabla 3.3, la *odds* de ser diabético previamente a saber la condición, es decir, no condicionada a la glucosuria, se basa en la *probabilidad total* o *marginal*, y se calcularía dividiendo 200 entre 800 (*odds* previa = 1/4 o bien 1:4); se interpreta diciendo que hay un diabético por cada cuatro no diabéticos:

$$Odds \text{ previa} = \frac{P(D)}{P(nD)} = \frac{200/1.000}{800/1.000} = \frac{200}{800} = \frac{1}{4}$$

Puede hablarse también de una *odds posterior*, que sería la *odds* condicionada a cumplir un requisito. En el ejemplo, la *odds* posterior sería la condicionada a tener glucosuria y valdría:

$$Odds \text{ posterior} = \frac{P(D|G)}{P(nD|G)} = \frac{60/68}{8/68} = \frac{60}{8} = 7{,}5$$

El teorema de Bayes demuestra que la *odds* posterior (condicional) es igual a la *odds* previa multiplicada por el «factor Bayes»:

$$Odds \text{ posterior} = odds \text{ previa} \times \text{factor Bayes}$$

El factor Bayes equivale a la razón de probabilidades de la condición, es decir, al cociente de probabilidades de presentar la condición (glucosuria) entre los que tienen el evento (diabéticos) y los que no lo tienen (no diabéticos). Ya se dijo que tener glucosuria era 30 veces más frecuente entre los diabéticos que entre los no diabéticos. A ese cociente se le llama factor Bayes:

$$\text{Factor Bayes} = \frac{P(G|D)}{P(G|nD)} = \frac{0{,}3}{0{,}01} = 30$$

El factor Bayes se interpreta como que es 30 veces más frecuente encontrar glucosuria en los diabéticos que en los no diabéticos. Se cumplirá la regla de Bayes:

$$\textit{Odds}\ \text{posterior} = \text{factor Bayes} \times \textit{odds}\ \text{previa} = 30 \times \frac{1}{4} = 7{,}5$$

Como se vio, la *odds* posterior era 7,5 y equivalía a dividir 60 entre 8. Esta *odds* posterior se interpreta como que entre los que tienen glucosuria hay 7,5 diabéticos por cada uno que no es diabético (es 7,5 veces más probable ser diabético si se cumple la condición de la glucosuria).

Si se transforma la *odds* en probabilidad, usando la expresión vista:

$$P = \frac{\textit{odds}}{1 + \textit{odds}}$$

$$\text{Probabilidad} = 7{,}5/(1 + 7{,}5) = 0{,}882$$

Esta es la probabilidad posterior, es decir, la probabilidad de diabetes condicionada a glucosuria, que ya se había calculado.

Si ahora se plantease, por ejemplo, que en una población con un 10% de diabéticos, el 40% de los diabéticos presentan glucemias por encima de 140 mg/dl y solo el 5% de los no diabéticos las presentan, simplemente aplicando el teorema de Bayes se puede responder a la pregunta de cuál es la probabilidad de que una persona con glucemia superior a 140 sea diabética:

$$\textit{Odds}\ \text{previa} = \frac{P(D)}{P(nD)} = \frac{0{,}1}{0{,}9} = \frac{1}{9}$$

$$\text{Factor Bayes} = \frac{P(\text{glucemia} > 140 \mid D)}{P(\text{glucemia} > 140 \mid nD)} = \frac{0{,}4}{0{,}05} = 8$$

$$\textit{Odds}\ \text{posterior} = \textit{odds}\ \text{previa} \times \text{factor Bayes} = 8 \times \frac{1}{9} = \frac{8}{9} = 0{,}889$$

La *odds* posterior valdría 8:9. Transformar una *odds* en una probabilidad consiste simplemente en añadir el numerador al denominador. Por lo tanto, la probabilidad de ser diabético si se presenta una glucemia > 140 mg/dl (este es el *valor predictivo positivo*) sería:

$$P(D \mid \text{glucemia} > 140) = 0{,}889 / (1 + 0{,}889) = 0{,}471$$

El 47,1% de los que presentan glucemias superiores a 140 mg/dl serán diabéticos en esa población. Este aspecto es lo que más le suele interesar al médico, saber cuál es la probabilidad de enfermedad condicional a que un test diagnóstico sea positivo, y se llama *valor predictivo positivo*. Corresponde a la probabilidad *a posteriori*. Su *odds* es la *odds* posterior.

En el contexto de las pruebas diagnósticas, el factor Bayes se denomina también razón de verosimilitud *(likelihood ratio)*.

## 3.8. PLANTEAMIENTO BAYESIANO, INFERENCIA BAYESIANA

Si se informa de que una enfermedad (E) se da en los pacientes que acuden a urgencias con una relativa baja frecuencia, en concreto, que hay una persona con esa enfermedad por cada 200 que no la padecen, esta (1/200) será la *odds* previa a cualquier otra información. La presunción de que la padezca algún paciente que llega a urgencias, antes de saber nada más de él, será que su *odds* de padecer esta enfermedad es 1:200. Se pueden realizar dos pruebas, cada una de las

***Tabla 3.4 Características de la prueba***

| | PRUEBA 1 | PRUEBA 2 |
|---|---|---|
| Sensibilidad | 90%: $p(\text{prueba}_1+ \mid E) = 0{,}9$ | 70%: $p(\text{prueba}_2+ \mid E) = 0{,}7$ |
| Especificidad | 80%: $p(\text{prueba}_1- \mid nE) = 0{,}8$ | 99%: $p(\text{prueba}_2- \mid nE) = 0{,}99$ |

cuales puede ser positiva o negativa. Las características de las pruebas 1 y 2 se muestran en la tabla 3.4. Si ambas pruebas diesen un resultado positivo, el planteamiento bayesiano será el que recoge la figura 3.10.

En la figura 3.10 se ve que la *odds* se va multiplicando por el factor Bayes de cada prueba a medida que las pruebas van dando resultados positivos. Este proceso podría continuar con sucesivos pasos, en los que siempre se parte de la *odds* obtenida en el paso previo. Pero, en el ejemplo, ya con dos pruebas positivas, la decisión debería ser que el paciente tiene la enfermedad, ya que es más probable que la tenga que lo contrario. Al principio se nos han dado las probabilidades de un resultado u otro de la prueba *condicionada* a la enfermedad. Al final, estimamos una *odds* de enfermedad *condicionada a los resultados observados en las pruebas*. Este es el fundamento de la estadística bayesiana (2). La estadística frecuentista convencional, como se verá en el capítulo 4, calcula las probabilidades de que ocurra un resultado condicionadas a que una hipótesis sea cierta (pero no valora cuál es la probabilidad de que esa hipótesis se cumpla). La estadística bayesiana, en cambio, calcula la probabilidad de que la hipótesis sea cierta condicionada a los resultados obtenidos.

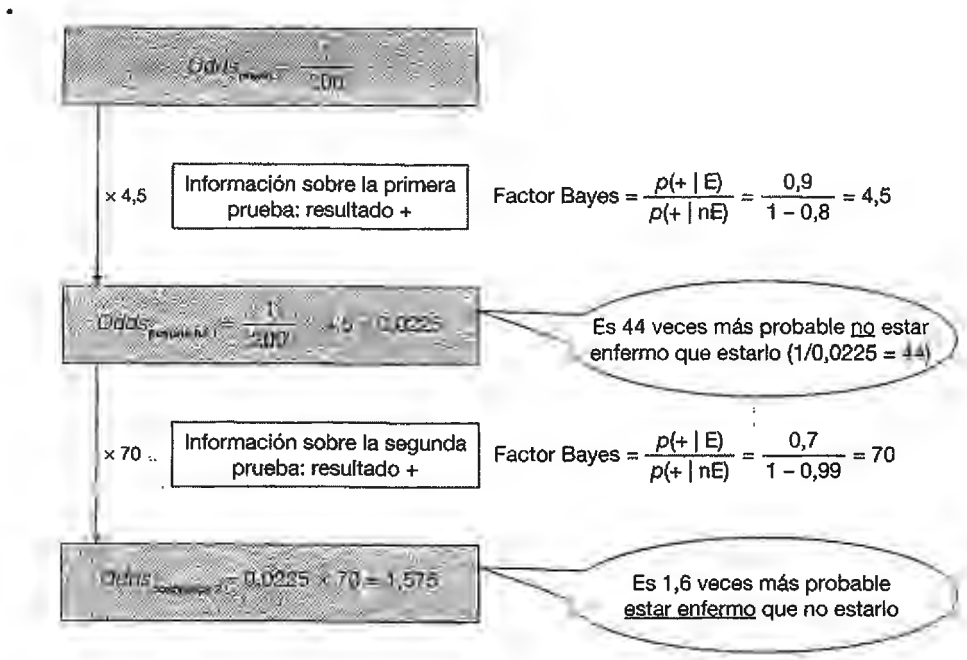



**Figura 3.10** Evolución sucesiva de la *odds* de enfermedad al resultar dos pruebas diagnósticas positivas.

## 3.9. DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

En estadística existen dos conceptos importantes que se deben conocer antes de profundizar en las distribuciones de probabilidad: *estimador* y *parámetro*. En la práctica, nunca se tiene acceso directo al estudio de la población total y se utiliza solo una *muestra* que procede de esa población teórica. En la muestra solo se pueden calcular estimadores ($\bar{x}$, s, etc.). Un *estimador* es una función de los valores de la muestra que permite obtener un valor aproximado de alguna característica de la población de la que se ha extraído dicha muestra. El valor de esa característica en la población se denomina *parámetro* ($\mu$, $\sigma$, etc.). Para los parámetros se suelen reservar las letras griegas. Así, mientras que la media poblacional ($\mu$) para una determinada característica de la población sería un parámetro imposible de calcular en la práctica, porque requeriría estudiar a todos los individuos (pasados, presentes y futuros) de la *superpoblación* a la que se aplicará la verdad científica obtenida, un estimador proporciona una idea aproximada, que sería, en este caso, la media muestral ($\bar{x}$) calculada en una muestra procedente de esa población.

La *distribución de frecuencias o distribución empírica* de una variable viene dada por la frecuencia con que se observan realmente en la muestra estudiada cada uno de los posibles valores que puede tomar esa variable. En cambio, se denomina *distribución de probabilidad* a aquella que presenta el conjunto de todos los valores que *teóricamente* podría tomar una variable, junto con sus correspondientes probabilidades calculadas ordinariamente siguiendo los principios de la ley de Laplace u otros análogos.

Si se lanzase una moneda infinitas veces, se esperaría obtener un 50% de cruces. Pero esto es teórico. Tras realizar en la práctica 20 lanzamientos, se obtienen, por ejemplo, 6 caras y 14 cruces en vez de 10 caras y 10 cruces. Se debe a la *variabilidad aleatoria* o al azar, pues se trata solo de una muestra. Lo encontrado en la muestra (6 caras y 14 cruces) sería la *distribución empírica*, mientras que lo teóricamente esperado (50% de caras y 50% de cruces) sería la distribución de probabilidad. Pero, si se lanzase la moneda más veces, la distribución empírica casi siempre se aproximará más a la teórica.

La mayoría de las distribuciones de probabilidad son descritas por uno o más *parámetros* (media, varianza, etc.). En estadística, se *asume* habitualmente que una muestra procede de una población que sigue una determinada distribución teórica de probabilidad. Esto no puede *comprobarse* de manera definitiva en ningún caso, pero en muchas ocasiones no hay grandes inconvenientes para *creérselo*. Cuando el método usado requiere asumir que los datos de una muestra pertenecen a una población con una distribución teórica conocida, suele decirse que dicho método es *paramétrico*. Se dice que es un método paramétrico porque se basa en los parámetros que definen esa distribución teórica. Si no se presupone nada acerca de la distribución de la población, se utilizan los métodos *no paramétricos* o de distribución libre, mala traducción acuñada por el uso de lo que sería en inglés *exentos de distribución (distribution-free)*. En general, se usan más los métodos paramétricos, pero en ocasiones debe recurrirse a los no paramétricos para no arriesgarse a asumir presupuestos erróneos.

Las distribuciones teóricas de probabilidad más usadas son la distribución binomial o la de Poisson para variables categóricas, y la distribución normal o de Gauss para las variables numéricas (cuantitativas continuas). Antes de estudiar cada una de ellas, es interesante conocer una distribución muy simple, que es la uniforme.

### 3.9.1. Distribución uniforme (discreta)

Imagínese que se lanza un dado 600 veces y se anotan las veces en que ha salido cada una de las seis caras. ¿Cuál sería la frecuencia teóricamente esperada para cada cara? Ahora ya no se trata de una variable con solo dos posibilidades. Existen seis posibles resultados. Si el dado es correcto, cada uno de sus seis lados tiene teóricamente la misma probabilidad de salir y lo esperado sería: 600/6 = 100. Si no existiese una cierta (aunque pequeña) variabilidad al azar, cada una de las seis

*Tabla 3.5 Distribución teórica y empírica observada al lanzar un dado 600 veces*

| | | FRECUENCIA ESPERADA | FRECUENCIA OBSERVADA |
|---|---|---|---|
| **Resultado** | 1 | 100 | 89 |
| | 2 | 100 | 125 |
| | 3 | 100 | 112 |
| | 4 | 100 | 88 |
| | 5 | 100 | 113 |
| | 6 | 100 | 73 |
| **Total** | | 600 | 600 |

caras saldría 100 veces. A esta distribución de probabilidad se le llama «uniforme», porque otorga la misma probabilidad a todos los sucesos (tabla 3.5).

Esta distribución teórica uniforme casi nunca se observa en la realidad exactamente así, pues existe una variabilidad debida al azar. Al hacer el experimento se obtuvo la distribución empírica de frecuencias de la tabla 3.5.

Las diferencias entre lo observado y lo esperado no son grandes, pero existen. En esta diferencia entre lo observado y lo *esperado* (lo que se esperaría si los datos siguiesen exactamente una *distribución teórica* de probabilidad) se basa la mayor parte de los test estadísticos. El balance entre *efecto* y variabilidad atribuible al azar está en el núcleo del razonamiento estadístico: cuanto más variable sea el suceso, más difícil es apreciar un efecto sobre él.

### 3.9.2. Distribución binomial

La distribución binomial se refiere a sucesos en los que solo existen dos posibilidades, como el lanzamiento de una moneda, el hecho de que un paciente padezca o no diabetes, etc. Se trata de *dos sucesos mutuamente excluyentes.* Al lanzar una moneda, la distribución de probabilidad consistiría en esperar un 50% de caras ($\pi = 0,5$) y un 50% de cruces ($\pi = 0,5$). Pero ¿qué pasa si se lanzan dos monedas a la vez? Existen 4 posibilidades teóricas (2 caras, 1 cara y 1 cruz, 1 cruz y 1 cara, y 2 cruces), cada una con una probabilidad de 0,25 (1/4). Si lo que interesa es el *número* de veces que sale cruz, los posibles resultados se simplificarían (0 cruces, 1 cruz y 2 cruces) y se podrían expresar como la probabilidad de que salgan *k* cruces: P(cruces = *k*) (tabla 3.6).

Si en vez de una moneda ($\pi = 0,5$) se trata de un dado con seis caras y lo que interesa es que salga un 6 o no salga, la probabilidad teórica *($\pi$)* será $\pi = 1/6$. Para usar la distribución binomial en este último ejemplo deben dicotomizarse los posibles resultados (obtener un 6 frente a cualquiera de los otros cinco resultados que se agrupan juntos como única alternativa). Habrá siempre, por tanto, dos características («parámetros») que definen una distribución binomial:

- El *número* (n) *de intentos o de unidades* (cantidad de lanzamientos de dados, individuos observados, intentos de curar a un paciente, etc.).
- La *probabilidad* ($\pi$) *teórica de éxito en cada intento.*

Se suele llamar *n* al número de intentos y $\pi$ a la probabilidad de éxito en cada intento. Finalmente, falta fijar otra característica, a la que se llamará *k*, que es el número de éxitos que se alcanzarán.

*Tabla 3.6 Distribución teórica de la probabilidad de obtener un cierto número de cruces al lanzar dos monedas*

| P (CRUCES = *K*) | N.° DE CRUCES | PROBABILIDAD |
|---|---|---|
| P (cruces = 0) | 0 | 1/4 |
| P (cruces = 1) | 1 cruz | 1/4 + 1/4 = 1/2 |
| P (cruces = 2) | 2 cruces | 1/4 |

Por ejemplo, si en una población la probabilidad de fumar es del 20%, ¿cuál es la probabilidad de que al entrevistar a dos personas ambas sean fumadoras? Esto supone:

$$n = 2$$
$$\pi = 0,2$$
$$k = 2$$

Se podría contestar esta pregunta con un diagrama de árbol.

Pero si se pregunta: ¿cuál es la probabilidad de que al entrevistar a 10 personas haya tres fumadores? El diagrama de árbol se complicaría mucho. Afortunadamente, existe la ecuación de la distribución binomial, que resuelve este problema[2]:

$$p(k) = \pi^k (1-\pi)^{n-k} \frac{n!}{(n-k)!\,k!}$$

La distribución binomial se simboliza como $B(n,\pi)$ y permite modelizar la distribución de probabilidad de un cierto carácter dicotómico (del tipo éxito o fracaso) que se observaría al extraer aleatoriamente (con reposición) muestras de tamaño $n$ de una población en la que se conoce la proporción $\pi$ de dicho carácter. Siempre que se conozca cuál es la probabilidad $(\pi)$ de obtener un resultado, la expresión matemática calcula cuál es la probabilidad $p$ de obtener $k$ resultados (éxitos = $k$) después de $n$ intentos. En el segundo ejemplo, las circunstancias eran $k = 3$, $n = 10$ y $\pi = 0,2$. Aplicando la ecuación, se obtiene que esto ocurrirá en algo más del 20% de las ocasiones, exactamente en el 20,13%:

$$p(k=3) = \pi^k (1-\pi)^{n-k} \frac{n!}{(n-k)!\,k!} = 0,2^3 (0,8)^7 \frac{10!}{7!\,3!} = 0,2013$$

Si lo que interesa es la probabilidad de que se encuentren *como mucho* tres fumadores, habrá que sumar a 0,2013 la probabilidad de hallar dos fumadores, la probabilidad de hallar solo un fumador y la de no encontrar ningún fumador:

$$p(k \leq 3) = p(k=0) + p(k=1) + p(k=2) + p(k=3)$$

La distribución binomial es asimétrica, pero, a medida que la proporción teórica $(\pi)$ se aproxima a 0,5, se va haciendo más simétrica. Sucede lo mismo a medida que aumenta $n$. En todos estos casos, la distribución binomial tiende a ser una distribución teórica que tiene forma de campana, es simétrica y mesocúrtica.

La esperanza matemática (viene a equivaler a la media) y la varianza de una variable que siga una distribución binomial vienen dadas por:

$$\text{Esperanza matemática } (\sim \text{media}) = n\pi \qquad \text{Varianza} = n\pi(1-\pi).$$

### 3.9.3. Distribución de Poisson

En una distribución binomial en la que $n$ es grande ($n \geq 100$) y la característica de interés es poco frecuente ($\pi \leq 0,05$), existe una buena aproximación más sencilla de calcular, que es la distribución de Poisson. Su expresión es:

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

2 Un número, por ejemplo 5, acompañado por un signo de admiración (5!) se lee 5 factorial y equivale a multiplicar 5 por 4 por 3 por 2 por 1: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. En general, $n! = n \times (n-1) \times (n-2) \times \ldots 2 \times 1$. Hay que tener en cuenta que $0! = 1$.

donde $\lambda = n\pi$ y «e» es la base de los logaritmos naturales (e = 2,71828).

Para aplicar la distribución de Poisson, basta con conocer dos datos:

- Número de éxitos observados: $k$.
- Número de éxitos esperados: $\lambda$.

Por ejemplo, entre 200 pacientes estudiados, se desea saber cuál es la probabilidad de encontrar un paciente con un polimorfismo genético cuya probabilidad $(\pi)$ es de 0,01. Primero se obtendrá $\lambda$:

$$\lambda = 200 \times 0{,}01 = 2$$

Este número $\lambda = 2$ se interpreta como el número *esperado* de personas con ese polimorfismo. Se esperaría encontrar dos personas con el polimorfismo, pero el problema pregunta cuál es la probabilidad de encontrar solo una ($k = 1$). Para eso se aplica la fórmula:

$$p(k=1) = \frac{2^1}{1!} e^{-2} = 0{,}271$$

La interpretación es que la probabilidad de encontrar exactamente un polimorfismo en esa muestra de 200 pacientes estudiados es del 27,1%. En cambio, si se hubiesen encontrado 20 personas con ese polimorfismo entre los 200 pacientes estudiados, la probabilidad de explicar este hallazgo solo por el azar sería muy pequeña y podría decirse que hay una gran diferencia entre lo observado y lo esperado:

$$p(k=20) = \frac{2^{20}}{20!} e^{-2} = 5{,}8 \times 10^{-14}$$

Este cálculo indica que la probabilidad de explicar este hallazgo por el azar es mínima (seis de cada 100 billones de veces que se repitiese el muestreo). Esto nos haría sospechar que tal vez el valor *teórico* que se ha asumido ($\pi = 0{,}01$) no sea el correcto en la población de la que procede esta muestra de 200.

Este procedimiento de calcular lo esperado por la distribución de Poisson se puede aplicar siempre que se cumplan dos condiciones:

1. Tamaño de muestra grande ($n \geq 100$).
2. Suceso poco frecuente ($\pi \leq 0{,}05$).

Cuanto más frecuente sea el suceso y más pequeña la muestra, peor será la aproximación. En cambio, la distribución binomial puede aplicarse siempre y proporciona resultados exactos.

Una característica interesante de la distribución de Poisson es que su media y su varianza coinciden:

$$\mu = \sigma^2 = \lambda$$

Como la distribución de Poisson se aproxima bien a una binomial cuando $n$ es grande y $\pi$ es pequeña, esta distribución de Poisson está especialmente indicada cuando hay sucesos *raros*, como ocurrencias de casos de una enfermedad por unidad de tiempo (se asume que ocurren de modo homogéneo en el tiempo y que todo suceso es independiente del resto de sucesos) o número de partículas en un medio (se asume que están repartidas al azar por todo el medio).

## 3.10. DISTRIBUCIONES DE PROBABILIDAD CONTINUAS: DISTRIBUCIÓN NORMAL

Imagínese que se lanzan 1.000 veces dos dados simultáneamente y *se suman* los resultados. Se simuló el experimento y se obtuvo la distribución que aparece en el diagrama de barras de la figura 3.11.

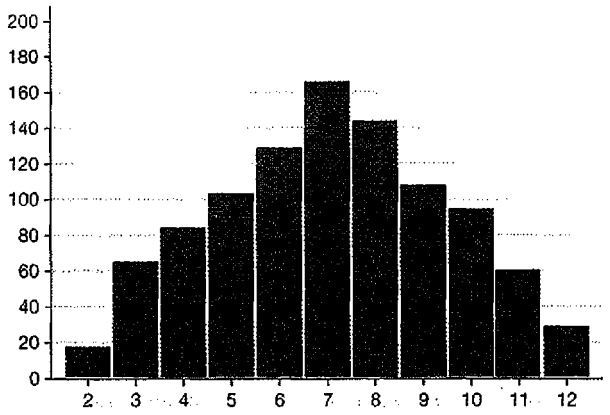



**Figura 3.11** Resultados empíricos (sumas) obtenidos realmente al lanzar dos dados 1.000 veces.

Aparece una distribución de los datos que se aproxima a la forma de una campana, es simétrica y mesocúrtica. Para entenderlo habría que empezar por preguntarse cuál es la probabilidad de que un dado obtenga el 1. Si el dado tiene seis caras, la probabilidad teórica (según la ley de Laplace) de cada una de ellas es un sexto (1/6); por tanto, la probabilidad de que en un dado obtenga un 1 es también 1/6. ¿Cuál es la probabilidad de que los dos dados obtengan simultáneamente la cara que tiene un 1? La probabilidad de que simultáneamente obtengan un 1 los dos dados es la de un dado multiplicada por la del otro, es decir, $1/6 \times 1/6 = (1/6)^2 = 1/36$.

El valor máximo alcanzable teóricamente con los dos dados sería aquel en que ambos obtuviesen un 6, sumando en total 12. La probabilidad de obtener una suma de 12 también sería de 1/36. Tanto para obtener una suma igual a 2 como para obtener una suma igual a 12 se requiere que se produzca un único resultado en ambos dados. No hay *ninguna otra posible combinación* que pueda lograr esa suma. En cambio, es muy fácil que la suma de ambos dados sea igual a 7, ya que diferentes combinaciones conducen al mismo resultado: 1 + 6; 2 + 5; 3 + 4; 4 + 3; 5 + 2; 6 + 1. Algo similar sucede con una suma que sea 6 u 8. Esto explica por qué la distribución va aproximándose a la forma de una campana a medida que aumenta el número de lanzamientos, ya que los valores centrales pueden ser el fruto de combinaciones que tienen mayor probabilidad de darse (tabla 3.7).



***Tabla 3.7 Modos de obtener cada posible suma al lanzar dos dados***

| SUMA | MODOS DE OBTENERLA | PROBABILIDAD |
|---|---|---|
| 2 | (1 + 1) | (1/6)(1/6) = 1/36 |
| 3 | (1 + 2) (2 + 1) | 2(1/36) = 1/18 |
| 4 | (1 + 3) (2 + 2) (3 + 1) | 3(1/36) = 1/12 |
| 5 | (1 + 4) (2 + 3) (3 + 2) (4 + 1) | 4(1/36) = 1/9 |
| 6 | (1 + 5) (2 + 4) (3 + 3) (4 + 2) (5 + 1) | 5(1/36) = 1/7,2 |
| 7 | (1 + 6) (2 + 5) (3 + 4) (4 + 3) (5 + 2) (6 + 1) | 6(1/36) = 1/6 |
| 8 | (2 + 6) (3 + 5) (4 + 4) (5 + 3) (6 + 2) | 5(1/36) = 1/7,2 |
| 9 | (3 + 6) (4 + 5) (5 + 4) (6 + 3) | 4(1/36) = 1/9 |
| 10 | (4 + 6) (5 + 5) (6 + 4) | 3(1/36) = 1/12 |
| 11 | (5 + 6) (6 + 5) | 2(1/36) = 1/18 |
| 12 | (6 + 6) | (1/6)(1/6) = 1/36 |

***Tabla 3.8 Resultados obtenidos al sumar cinco dados lanzados 1.000 veces***

| SUMA | FRECUENCIA |
|---|---|
| 6 | 1 |
| 7 | 10 |
| 8 | 11 |
| 9 | 16 |
| 10 | 34 |
| 11 | 40 |
| 12 | 52 |
| 13 | 57 |
| 14 | 60 |
| 15 | 64 |
| 16 | 99 |
| 17 | 85 |
| 18 | 70 |
| 19 | 62 |
| 20 | 55 |
| 21 | 57 |
| 22 | 59 |
| 23 | 49 |
| 24 | 43 |
| 25 | 38 |
| 26 | 15 |
| 27 | 13 |
| 28 | 3 |
| 29 | 7 |
| Total | 1.000 |

Así como es muy fácil conseguir sumas de los valores centrales, los valores *extremos* se dan muy poco. Supóngase ahora que ampliamos el anterior experimento y, en vez de lanzar 1.000 veces dos dados, ahora se lanzan cinco dados. Esto es lo que se simula en la tabla 3.8. y en la figura 3.12.

Siempre que existan muchos factores independientes que determinan un resultado, los valores extremos no suelen darse prácticamente *nunca* en la realidad. Este hecho está en la base de un teorema que se llama *teorema central del límite* (**10**). Se debe a que, para que se den valores extremos, tienen que coincidir muchos factores independientes que apunten todos en la misma dirección,

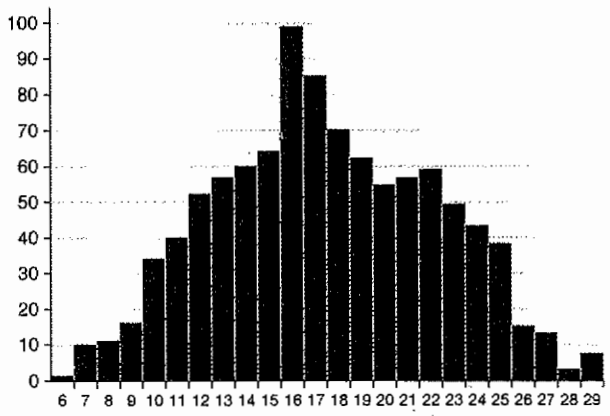



**Figura 3.12** Resultados obtenidos al sumar cinco dados lanzados 1.000 veces.

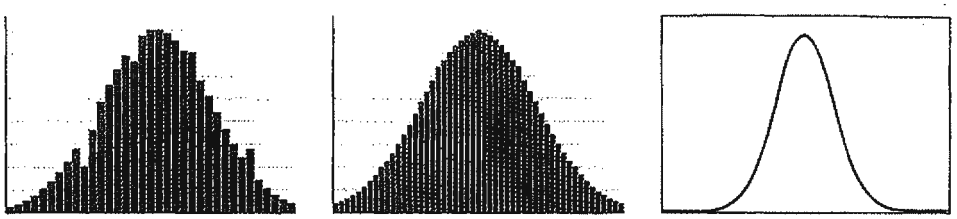


**Figura 3.13** Cambios que se van produciendo en la distribución de frecuencias al aumentar el número de factores independientes y las repeticiones de la experiencia.

lo cual es poco probable. Lo más probable es que apunten en direcciones dispares. En el ejemplo, los valores que más se han producido en las 1.000 tiradas de los cinco dados son el 16 y el 17, que son centrales y han salido 99 y 85 veces, respectivamente. Esto es debido a que los valores centrales son más fáciles de conseguir, pues existen muchas combinaciones que suman 16 o 17.

Teóricamente podría ocurrir que los cinco dados en una misma tirada sacasen un 1 y, entonces, su suma fuese 5, pero esto no ha sucedido ninguna de las 1.000 veces.

¿Cuál es la probabilidad de que los cinco dados obtengan un 1? Al ser sucesos independientes, la probabilidad de que simultáneamente salga un 1 en los cinco dados es la de que salga en un dado (1/6) multiplicada por la de que salga en otro y así hasta 5:

$$1/6 \times 1/6 \times 1/6 \times 1/6 \times 1/6 = (1/6)^5 = 0{,}0001286$$

Esto equivale a dividir 1 entre 7.776 ($6^5 = 7.776$). Teóricamente, haría falta, pues, lanzar los cinco dados 7.776 veces para obtener cinco caras con un uno simultáneamente. Lo mismo ocurriría para una suma total de 30 (obtener un 6 en los cinco dados).

A medida que van aumentando las categorías, el diagrama contiene más barras, que son cada vez más estrechas y, al aumentar las repeticiones de la experiencia, el contorno se alisa hasta llegar a una curva sin saltos (fig. 3.13). Así se representa en la tercera gráfica situada más a la derecha, lo que correspondería a la distribución teórica que se obtendría si el número de repeticiones fuese infinito. Tiene forma de campana, es simétrica, mesocúrtica y, en ella, la media, la mediana y la moda coinciden. Esta distribución teórica de probabilidad es conocida como *distribución normal o campana de Gauss* en honor del matemático que la describió.

En la tercera gráfica de la figura 3.13, el eje horizontal o de abscisas corresponde a cada uno de los valores posibles de la variable que se estudia (p. ej., niveles de colesterol), mientras que podría pensarse que el eje vertical (ordenadas) corresponde a la frecuencia con que ocurre ese valor; sin embargo, la probabilidad de que suceda un valor individual aislado teóricamente es 0, y solo el área que queda bajo la curva correspondiente a un cierto intervalo de valores se interpreta como la probabilidad de que ocurra alguno de los valores contenidos en ese intervalo. La probabilidad de un valor concreto es 0 porque el modelo matemático de la distribución normal es el de una variable *continua* y, en esta situación, hablar de un valor concreto supondría una exactitud absoluta y radical.

Por ejemplo, si se sabe que el nivel de colesterol total de una población sigue una distribución normal y se pregunta cuál es la probabilidad de que alguien tenga un colesterol = 200 mg/dl, dicha probabilidad es 0, porque no se estará refiriendo a que tenga un nivel de colesterol entre 199,5 y 200,5 mg/dl, ni entre 199,9 y 200,1 mg/dl, sino exactamente 200,000000000000, y habría que ampliar los dígitos decimales hasta el infinito. La solución con la distribución normal es valorar siempre la probabilidad para un rango o intervalo entre dos límites. Esto sí puede hacerse, por cercanos que sean los límites. Es preciso advertir, por tanto, que las variables continuas (colesterol,

tensión arterial, edad) tienen (teóricamente) un número infinito de posibles valores. En estos casos no tiene sentido hablar de la probabilidad de que un sujeto presente un determinado valor puntual de colesterol, puesto que, como tal, no puede ser calculada. En estos casos se habla de *densidad de probabilidad,* un concepto que se refiere a la probabilidad de que un sujeto tenga un valor incluido en un determinado intervalo entre dos valores. Aunque, como se verá más adelante, es importante desde el punto de vista conceptual entender la densidad de probabilidad, su valor absoluto (valor de la altura del eje vertical) tiene escasa utilidad práctica.

En la tercera gráfica de la figura 3.13, la suma de toda el área bajo la curva (desde $-\infty$ hasta $+\infty$) tiene una probabilidad = 1 (el 100% de los valores están entre esos límites).

La distribución normal es *continua* y, en cambio, la distribución binomial es *discreta.* La distribución normal es la indicada para datos que siguen una escala, al menos en teoría, continua (peso, talla, edad, colesterol, tensión arterial, ácido úrico, bilirrubina, etc.), y posee la ventaja adicional de que otras distribuciones, en ciertas condiciones, acaban por aproximarse a ella, como se vio antes para la distribución binomial cuando $\pi$ se acerca a 0,5 y $n$ es grande. También se aproxima a la normal la suma de varias distribuciones uniformes. Esto es lo que se acaba de ver (suma de varios dados).

La distribución normal teórica nunca se da exactamente en la realidad. Solo existen aproximaciones a ella, pero se puede expresar como ecuación matemática. No es preciso conocer esta expresión matemática para resolver la mayor parte de los problemas relativos a la distribución normal. Al ser expresada como un modelo o ecuación, la distribución se hace continua y teóricamente hay infinitos puntos, es decir, infinitos valores posibles. En teoría, su eje horizontal (abscisas) se extiende desde menos infinito hasta más infinito.

Cuando una variable $x$ sigue una distribución normal de media $\mu$ y varianza $\sigma^2$, se representa $x \in N(\mu;\sigma^2)$ y se lee: *x pertenece a una normal, con media* $\mu$ *(mu) y varianza* $\sigma^2$ *(sigma cuadrado).*

En general, una distribución normal se caracteriza por (fig. 3.14):

1. Tener forma de campana.
2. Ser simétrica (asimetría = 0).
3. No ser excesivamente plana ni excesivamente picuda (mesocúrtica).
4. Coincidir en ella la media, la mediana y la moda.

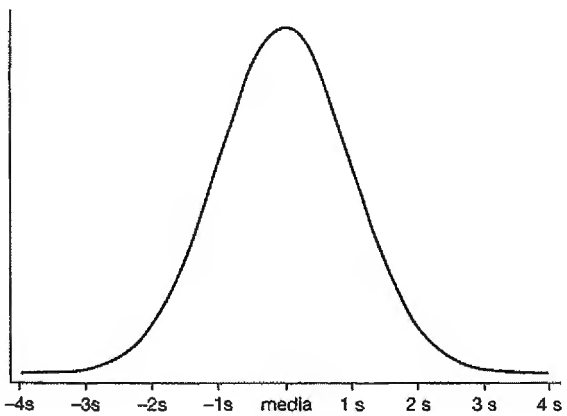



**Figura 3.14** Representación gráfica de la distribución normal.

5. Tener aproximadamente el 95% de sus valores dentro del intervalo $\mu \pm 2\ \sigma$ (media ± 2 desviaciones estándar). Exactamente, el 95% de los individuos se encuentra dentro del intervalo comprendido por $\mu \pm 1{,}96\ \sigma$. Además, casi el 100% de los valores está dentro del intervalo $\mu \pm 3\ \sigma$.
6. Ser la distribución muestral que siguen los índices o estimadores estadísticos calculados en una muestra. Esto es lo más importante.

Desde el punto de vista práctico es importante adquirir familiaridad con el procedimiento de *tipificar* o estandarizar la normal. Consiste en *transformar una determinada distribución normal en otra que tenga una media igual a 0 y una desviación estándar igual a 1*. Así se consigue que cualquier valor represente una distancia a la media expresada como el número de desviaciones estándar en que ese valor se aleja de la media. Este número en estadística se llama $z$. Habrá valores positivos de $z$, los que están por encima de la media, y valores negativos, por debajo de la media.

Para obtener $z$ se emplea la siguiente ecuación:

$$z = \frac{x - \mu}{\sigma}$$

donde $x$ es la antigua variable y $z$ la nueva variable con media = 0 y desviación estándar = 1. Para devolver la transformación a su estado original, se usará:

$$x = \mu + \sigma z$$

Por ejemplo, si la media de tensión arterial sistólica de una población es 120 mmHg y la desviación estándar es 25 mmHg, y *se asume que sigue una distribución normal*, se puede responder a diversas preguntas con estas sencillas fórmulas. Así, para conocer la proporción de personas que tienen una tensión arterial sistólica superior a 170 mmHg (fig. 3.15), habrá que calcular el valor $z$ que corresponde a 170:

$$z = \frac{x - \mu}{\sigma} = \frac{170 - 120}{25} = +2$$

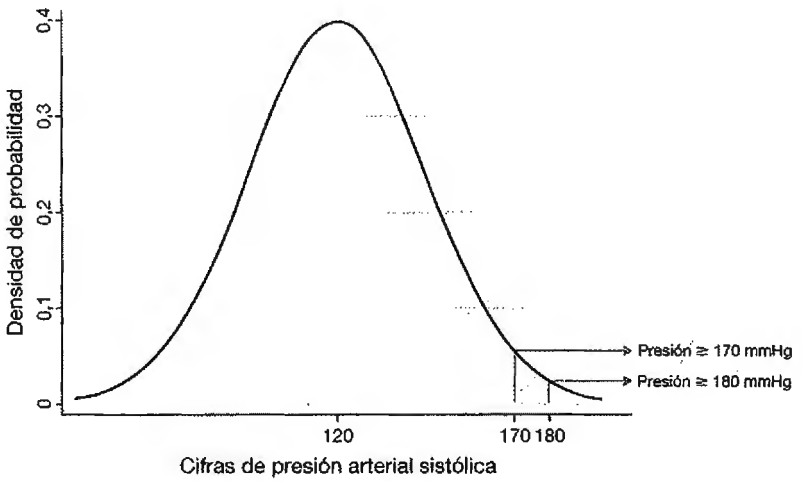



**Figura 3.15** Probabilidad de encontrar a un individuo en el intervalo entre +2 y +2,4 desviaciones estándar de la media en una distribución normal.

¿Qué significa saber que 170 mmHg corresponde a un valor de $z = +2$? En primer lugar, se sabrá que 170 mmHg está dos desviaciones estándar por encima de la media. En segundo lugar, existen tablas de la distribución normal que indican cuál es la probabilidad de que se dé un valor superior o inferior a cada valor de $z$. Estas tablas están encabezadas por una frase que dice:

*Dentro de la tabla se proporciona el valor de* p *para +z o para –z.*

Para buscar la probabilidad de la cola que está por encima de $z = +2{,}00$, hay que localizar el valor correspondiente a 2,0 (en la primera columna) y 0,00 (en la primera fila). La probabilidad *(p)* correspondiente al área de la cola que está a la derecha de un valor de $z = +2$ es 0,0228. Es decir, el 2,3% de los individuos tendrán valores superiores a 170 mmHg.

Para saber cuántas personas presentan tensiones arteriales sistólicas entre 170 y 180 mmHg, habrá que hacer lo mismo con 180 ($z$ será +2,40 entonces) y, a continuación, se busca en la tabla el área bajo la curva que queda a la derecha del valor de $z = +2{,}40$ correspondiente a 180. Este valor (0,0082) corresponde a la probabilidad de tener valores por encima de $z = +2{,}40$ y llevará a concluir que el 0,8% de esa población tiene tensiones de 180 o incluso más. Lo único que queda es restar 0,8% de 2,3%:

- Si hay un 0,8% de resultados por encima de 180.
- Y hay un 2,3% por encima de 170.
- ¿Cuántos habrá entre 170 y 180?

$$2{,}3\% - 0{,}8\% = 1{,}5\%$$

La solución es que el 1,5% de las personas de esa población tienen tensiones sistólicas comprendidas entre esos valores (de 170 a 180 mmHg).

También se puede plantear qué valor deja al 90% de las personas por debajo de él. Esto supone exactamente preguntarse cuál es el percentil 90. Para hacerlo hay que recorrer un camino similar, pero al revés, empezando por mirar las tablas:

1. Buscar en las tablas qué valor de $z$ deja 0,1 por encima. Este valor es $z = 1{,}28$.
2. Calcular $x$ a partir de $z$. Este valor es 152 mmHg.

$$x = \mu + z\sigma = 120 + (1{,}28 \times 25) = 152$$

Hay algunos valores clave de la normal que es interesante conocer de memoria y que se presentan en la tabla 3.9.

Es importante subrayar que, a pesar de su nombre (distribución *normal*), existen muchas variables biológicas que *no* siguen una distribución normal. Es más, lo más frecuente, especialmente en medicina clínica, es que las variables no se adapten perfectamente al modelo teórico matemático de la distribución normal. Es frecuente que unos pocos sujetos se desvíen mucho por encima de la normal por problemas de asimetría positiva. Suele suceder porque hay siempre algunas personas que —por estar enfermas— tienen valores muy altos (así sucede, por ejemplo, cuando se habla de que alguien tiene el colesterol sérico «por las nubes»).

***Tabla 3.9 Valores frecuentemente usados de la distribución normal***

| Z | P (UNA COLA) |
|---|---|
| 1,28 | 0,1 |
| 1,645 | 0,05 |
| 1,96 | 0,025 |
| 2,32 | 0,01 |

## 3.11. TEOREMA DEL LÍMITE CENTRAL

Aunque los valores que presenten los individuos de una población no sigan una distribución normal, *la distribución de los estimadores que se calculan en sucesivas muestras que se obtengan de estas poblaciones (distribución del estimador muestral) sí que seguirá aproximadamente una distribución normal.* La condición es que la muestra sea suficientemente grande. Esto figura en el núcleo de muchos métodos estadísticos y se conoce como *teorema del límite central o teorema central del límite.* Hace posible que se puedan realizar inferencias estadísticas a partir de muestras usando las propiedades de la distribución normal, aunque la población de la que procedan no siga la normal.

La única condición para que lo anterior se cumpla es que la muestra sea grande y extraída aleatoriamente de la población. Cuanto más grande sea la muestra, mejor se cumple este teorema. Por encima de 60 individuos, la adaptación de la distribución muestral de estimadores a la distribución normal es muy buena. Entre 30 y 60 individuos es aceptable. Por debajo de 30 individuos en la muestra empiezan a aparecer problemas (**11**).

Una consecuencia concreta del teorema central del límite se explica con la siguiente experiencia. Imagínese que alguien tiene a su disposición la lista completa con todos los valores de colesterol sérico de toda la población de Navarra (640.000 valores de colesterol). Cada día, esa persona elige al azar 30 valores de entre los 640.000 navarros y calcula su media. Diariamente se repite el cálculo con una muestra diferente, siempre con 30 valores de colesterol de personas elegidas al azar entre todos los navarros. Lo único que se guarda es la media calculada en la muestra que se extrae cada día. Al cabo de 1.000 días repitiendo la experiencia, se tendrán 1.000 medias obtenidas en 1.000 muestras de 30 individuos cada una y se podrá construir una base de datos *solo con las medias* muestrales. Se tendría una nueva distribución de valores en la que, en vez de calcular frecuencias de valores individuales, se obtendrán frecuencias de ocurrencia de cada media muestral. Esta situación se conoce como *distribución muestral de medias.* Se ha hecho la simulación por ordenador de esta experiencia suponiendo que la población de Navarra tiene un colesterol medio de 200 mg/dl, con una desviación estándar de 35 mg/dl, y que existe asimetría positiva en la población, con lo que no puede asumirse que la distribución poblacional de colesterol siga una normal (fig. 3.16).

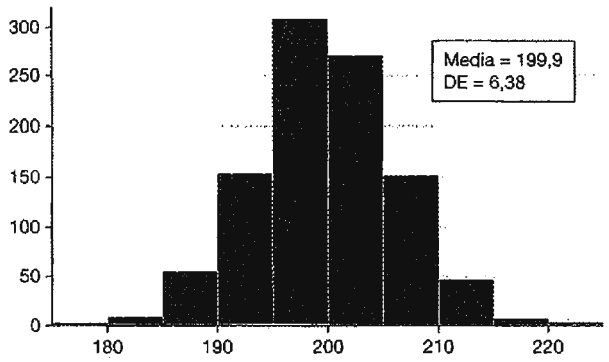



**Figura 3.16** Distribución muestral de medias (los valores representados no son individuales, sino las medias de 1.000 muestras). DE, desviación estándar.

Lo primero que llama la atención es que se observa, de nuevo, la forma de campana y demás características de la distribución normal. En este supuesto, el colesterol *en la población no seguía una distribución normal,* pero la distribución de las medias muestrales de colesterol *sí* que la sigue. La segunda característica es que la media de las muestras es prácticamente la misma que la media poblacional (199,9 mg/dl ~200 mg/dl). La tercera es que se ha reducido mucho la «desviación estándar». En la población total, la desviación estándar del colesterol era 35 mg/dl; en la distribución muestral de las medias, solo es 6,38. Esta nueva desviación estándar corresponde, según el teorema central del límite, al *error estándar de la media*, que vale $s / \sqrt{n}$ (v. capítulo 2). Efectivamente, se comprueba que, si se divide 35 por la raíz cuadrada de 30, se encontrará que el error estándar es 6,39, muy parecido a la desviación estándar de la distribución muestral de medias obtenida empíricamente por la simulación anterior.

Por lo tanto, si se tipifica esta nueva distribución muestral, se obtiene la distancia a la cual está de la media poblacional cada media muestral. Esta distancia estará medida en unidades de error estándar. Si antes, al hablar de individuos, un valor *z* se interpretaba como el número de desviaciones estándar en que un sujeto se separaba de la media, ahora, en el contexto de las muestras, un valor *z* se interpreta como el *número de errores estándar* en que una media muestral se separa de la media poblacional. También se entenderá ahora por qué el 95% de las medias calculadas en las muestras estarán aproximadamente en el intervalo ± 2 errores estándar de la media poblacional. Se abre paso así a la posibilidad de extraer conclusiones válidas acerca de la población a partir del estudio de muestras, como se hace en los test de hipótesis e intervalos de confianza basados en la distribución normal.

Se recomienda ver el vídeo titulado «Teorema central del límite (explicación, PowerPoint)», disponible en http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

## 3.12. CONDICIONES, PRUEBAS Y GRÁFICOS DE NORMALIDAD

Existen diversos test para comprobar si los valores de una variable siguen o no la distribución normal. Cuando resultan *significativos* (valor *p* de significación estadística < 0,05), se *rechaza* la hipótesis de normalidad. Estos test se deben interpretar con cautela, siempre a la luz del número de individuos para los que existan datos sobre esa variable. Si el número de individuos de nuestra base de datos fuese muy grande, bastará una pequeña desviación de la normalidad para que el test arroje un resultado significativo y se rechace la normalidad. Por el contrario, cuando hay pocos datos, casi nunca se dispondrá de evidencias para rechazar la normalidad y los test no serán significativos, a pesar de que existan desviaciones patentes de la normalidad. Es decir, son pruebas que se dejan influir notablemente por el tamaño de muestra. Funcionan mejor con tamaños de muestra intermedios, por ejemplo, entre $n = 20$ y $n = 200$ (estas cifras deben entenderse de modo flexible). No obstante, si el tamaño de muestra fuese mayor de 200, en algunas ocasiones, con pequeñas desviaciones de la normalidad sin importancia práctica, se obtendrán resultados significativos. Por el contrario, si el tamaño de muestra fuese pequeño, menor de 20, a pesar de graves desviaciones de la normalidad, se obtendrán ocasionalmente falsas seguridades, pues el test no resultará significativo.

Por este motivo es conveniente usar *siempre* una combinación de enfoques para juzgar la adaptación de una variable a la normalidad.

Algunos test diseñados para comprobar la normalidad son:

- Test de Shapiro-Wilk W.
- Test de Shapiro-Francia W'.
- Test de D'Agostino.
- Test de Kolmogorov-Smirnov.
- Test de Lilliefors.

Todos estos procedimientos son contrastes de hipótesis (v. capítulo 4) y dan como resultado final una probabilidad (valor *p*) correspondiente a una *significación estadística*. ¿Cómo se interpretan? Son pruebas que calculan cuál sería la probabilidad de encontrar esta distribución de los datos (o una todavía más alejada de la normalidad) bajo la hipótesis nula de que en la población de la que procede la muestra esa variable siguiese una distribución normal perfecta. Para estas pruebas, *la hipótesis nula es la normalidad.* Por tanto, si la probabilidad de encontrar estos datos en el supuesto de que siguen una normal perfecta fuese alta (valor $p > 0{,}05$), no habría evidencias para rechazar la hipótesis nula y se podría asumir la normalidad. Pero, cuando el valor *p* de cualquiera de estas pruebas sea inferior a 0,05, es posible que existan dificultades para que pueda asumirse la normalidad.

De todos modos, cuando la muestra es grande ($n > 60$), con frecuencia se puede asumir la normalidad para *la distribución muestral de estimadores,* aunque estas pruebas arrojen un valor $p < 0{,}05$, ya que los estimadores calculados en muestras grandes, según se deriva del teorema central del límite, se aproximan a la distribución normal (**11**).

En el programa STATA se puede obtener el test de Shapiro-Wilk a través de:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Shapiro-Wilk normality test**

y seleccionar luego en el despegable la variable cuya normalidad se desee comprobar. Si se aplicase, por ejemplo, para la glucemia en una muestra de 50 sujetos, los resultados podrían ser:

```
. swilk glucemia
```

Shapiro-Wilk W test for normal data (Valor *p*)

| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| glucemia | 50 | 0.97160 | 1.336 | 0.617 | 0.26857 |

Se puede observar que, efectivamente, había 50 observaciones para la variable glucemia y que el valor *p* de significación estadística del test calculado es superior a 0,05, por lo que la normalidad se admitiría o sería asumible. Este tamaño de muestra ($n = 50$) se encuentra dentro del intervalo en que estas pruebas de normalidad funcionan relativamente bien.

El programa STATA también ofrece la opción de elegir el test de normalidad de la asimetría y la curtosis:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Skewness and kurtosis normality test**

Se seleccionará la variable para la cual se desea realizar el test. En el mismo ejemplo anterior, se obtendría:

```
. sktest glucemia
```

Skewness/Kurtosis tests for Normality (Valor *p*)

| Variable | Obs | Pr(Skewness) | Pr(Kurtosis) | joint adj chi2(2) | joint Prob>chi2 |
|---|---|---|---|---|---|
| glucemia | 50 | 0.2044 | 0.4004 | 2.45 | 0.2944 |

Se llegaría a la misma conclusión.

También existen procedimientos gráficos que permiten valorar si los datos se adaptan bien o no a una distribución normal. Son especialmente útiles cuando el tamaño muestral es pequeño. Primero debe observarse el histograma (fig. 3.17A).

El histograma de la glucemia podría sugerir cierto apartamiento de la normalidad, pero existen otros métodos gráficos más específicos para valorar dicha normalidad, como el gráfico

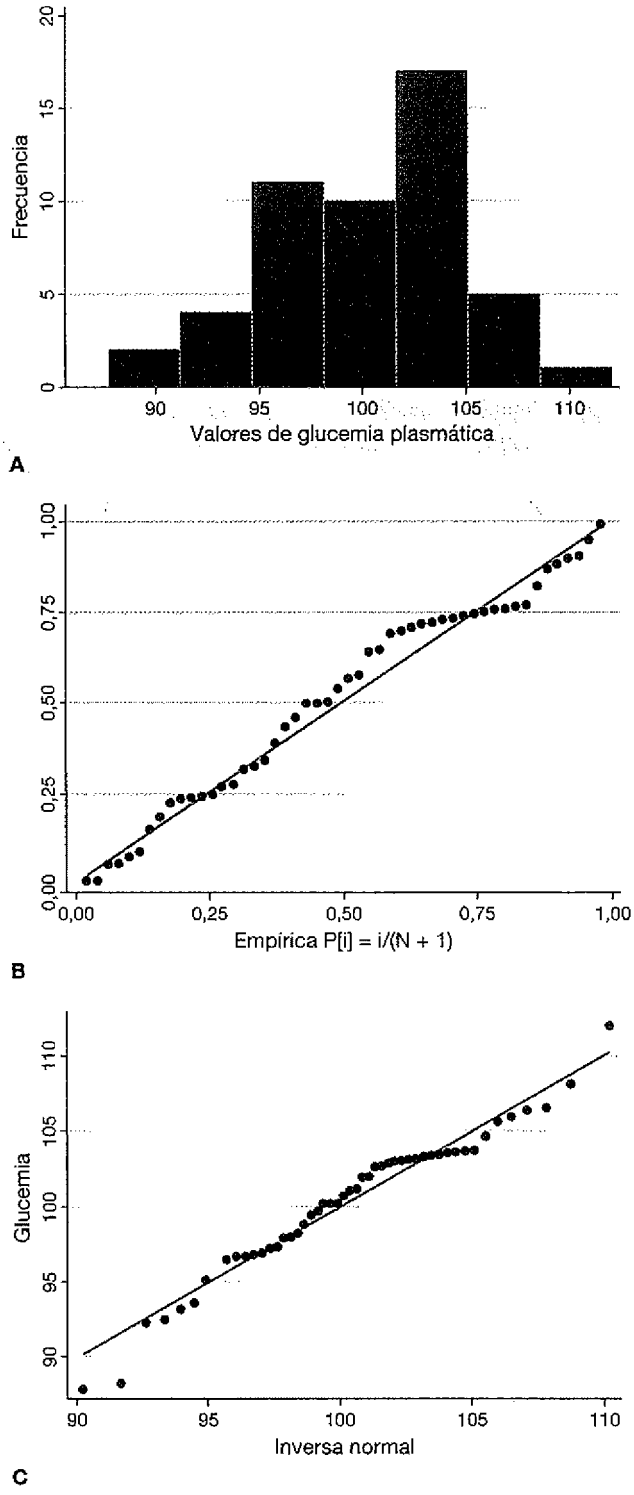



**Figura 3.17** Estudio de la distribución de la variable glucemia. **A.** Histograma. **B.** Gráfico P-P. **C.** Gráfico Q-Q.

estandarizado de probabilidad normal o *gráfico percentil-percentil (P-P)*. En el eje de las abscisas (horizontal) se representan los percentiles (porcentajes acumulados) de la distribución *observada* empíricamente, mientras que el vertical (ordenadas) corresponde a los percentiles *esperados* si la distribución siguiese una normal perfecta. Cuando lo observado coincida con lo esperado, los puntos para cada observación se situarán en la diagonal. En el programa STATA se obtendría así (fig. 3.17B):

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Normal probability plot, standardized**

También puede pedirse así:

```
pnorm glucemia
```

El primer punto que aparece tiene un valor correspondiente al porcentaje acumulado observado de 0,02 en el eje horizontal (es la primera de 50 observaciones; por lo tanto, representa el 2% de las mismas) y un valor de 0,006, que sería el porcentaje (0,6%) esperado de sujetos que tendrían glucemias inferiores o iguales a 87,9 mg/dl si la distribución fuese perfectamente normal. Visualmente, lo importante, también en el gráfico P-P, es que cuanto más se alejen de la diagonal los puntos, más se aleja la distribución de esa variable de la normal.

Otro gráfico es el *cuantil-cuantil* (QQ), que compara los valores *observados* (ahora en eje de ordenadas) con respecto a los valores *esperados* (eje de abscisas), que corresponderían a esas observaciones si la variable siguiese la distribución normal. Así, en una distribución normal de media 100,2 y desviación estándar 4,9, el valor esperado para el percentil 2 sería 90,1, y el mínimo valor observado en esta serie de 50 datos, 87,9 mg/dl (fig. 3.17C). Lo importante al interpretarlo es que, cuando haya apartamiento de los puntos con respecto a la diagonal, existirá alejamiento de la normalidad. En este ejemplo, los datos se adaptan bastante bien a la diagonal.

En el programa STATA, este gráfico se obtendría así:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Normal quantile plot**

También puede pedirse así:

```
qnorm glucemia
```

En resumen, y desde el punto de vista práctico, las gráficas P-P y Q-Q, así como el histograma y los test de normalidad, sugieren que se puede asumir la aproximación a la normalidad de la variable glucemia.

¿Qué ha de hacerse cuando no se puede asumir la normalidad y se desea utilizar un método que exija la normalidad como supuesto? Hay dos opciones. Una alternativa realista y pragmárica consiste en emplear un método no paramétrico, que no requiere la normalidad, y comparar los resultados con los del método paramétrico. Si no hay diferencias, se utilizan los paramétricos. La otra opción es intentar una transformación de la variable; la más empleada es la transformación logarítmica, que suele conseguir aproximar variables con asimetría positiva a la distribución normal.

En la figura 3.18 se muestra la distribución de la variable colesterol total en una muestra de 176 sujetos. Se observa que esta variable presenta asimetría positiva, ya que la cola de la derecha es más gruesa que la de la izquierda y la curva no es simétrica.

En STATA se pueden probar varias transformaciones simultáneamente con:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Ladder-of-powers**

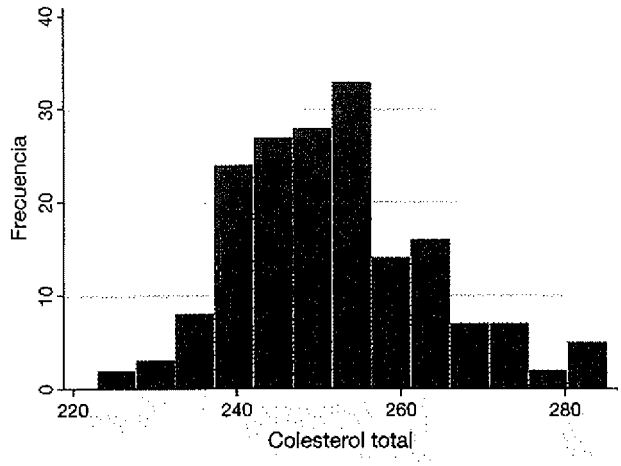



**Figura 3.18** Colesterol total (apartamiento de la normalidad con asimetría positiva).

También puede pedirse así:

```
ladder colesterol_total
```

Con esta orden, STATA probará automáticamente diversas transformaciones y realizará un test de normalidad para cada una de ellas. En concreto, trata la variable como si estuviese elevada al cubo o elevada al cuadrado, como si se transformase usando su raíz cuadrada, como si se tomasen logaritmos, como si se calculase el inverso de la raíz cuadrada, el inverso de la variable original, el inverso de la variable al cuadrado o el inverso de la variable al cubo. Lo que STATA denomina «identity» es la variable en su escala original:

```
. ladder colesterol_total

Transformation          formula                 chi2(2)     P(chi2)
-------------------------------------------------------------------
cubic                   coles~al^3                13.73     0.001
square                  coles~al^2                 9.77     0.008
identity                coles~al                   6.40     0.041
square root             sqrt(coles~al)             4.96     0.084
log                     log(coles~al)              3.45     0.178
1/(square root)         1/sqrt(coles~al)           2.18     0.335
inverse                 1/coles~al                 1.20     0.548
1/square                1/(coles~al^2)             0.12     0.941
1/cubic                 1/(coles~al^3)             0.30     0.862
```

En este listado se observa que cualquiera de las últimas seis transformaciones daría lugar a una variable de la que se podría asumir que sigue una distribución normal.

Este mismo resultado se puede visualizar gráficamente con la orden:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Ladder-of-powers histograms**

También puede pedirse así:

```
gladder colesterol_total
```

Se obtendría el gráfico mostrado en la figura 3.19. Se observa así que las últimas seis transformaciones generan distribuciones que se adaptan bien a una distribución normal.

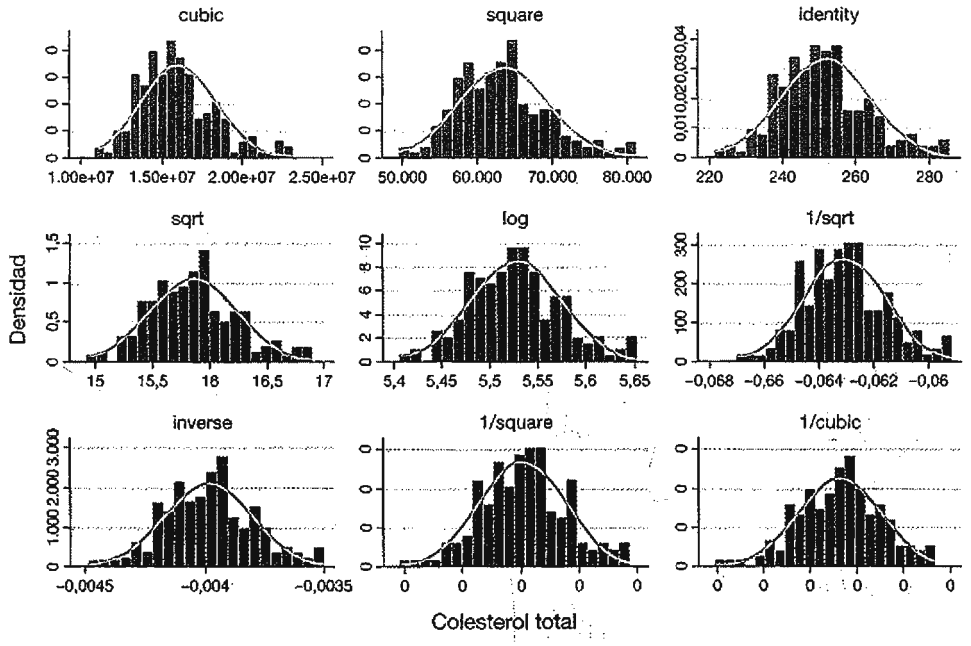



**Figura 3.19** Gráficos obtenidos con la orden *gladder* para la variable colesterol total.

## 3.13. LAS DISTRIBUCIONES BINOMIAL, DE POISSON Y NORMAL EN STATA

### 3.13.1. La distribución binomial en STATA

Para calcular las probabilidades de encontrar un número *k* de fumadores en una población donde la prevalencia ($\pi$) del uso del tabaco sea 0,2 en STATA, se empleará la orden **di binomial**(n, k, $\pi$). Así, la probabilidad de hallar 0 fumadores en 10 intentos en una población donde la prevalencia del uso del tabaco es de 0,2 será:

```
. di binomial(10,0,0.2)
.10737418
```

La orden **di binomial** de STATA da siempre como resultado la probabilidad *acumulada desde* k = *0 hasta el valor de* k *que se indique*. Esto implica que la orden **di binomial(10,2,0.2)** dará como resultado la probabilidad de hallar $k \leq 2$ fumadores (es decir, la probabilidad de hallar 0, más la de hallar 1, más la de hallar 2) entre 10 personas extraídas de una población con un 20% de fumadores. Por lo tanto, para hallar exactamente la probabilidad de $k = 2$ fumadores, a esta probabilidad de $k \leq 2$ podría restársele la probabilidad de hallar $k \leq 1$ fumador:

```
. di binomial(10,2,0.2)-binomial(10,1,0.2)
.30198989
```

Mejor alternativa es emplear otra orden **di binomialp,** que no da la probabilidad acumulada, sino individual, para un valor específico de *k*. Se obtendrá el mismo resultado que con la orden anterior:

```
. di binomialp(10,2,0.2)
.30198989
```

Al igual que en Excel, la orden **di binomial** de STATA da como resultado siempre la cola *de la izquierda.* En el caso de querer conocer la probabilidad de hallar al menos ocho fumadores (es decir, ocho *o más*) en 10 personas extraídas de una población con un 20% de fumadores, se estará pidiendo la cola de la derecha, y se procederá así:

```
. di 1-binomial(10,7,0.2)
.00007793
```

Con esta orden, se resta a 1 la probabilidad de hallar siete o menos fumadores. Esto será equivalente a la probabilidad de hallar ocho o más fumadores. Para obtener directamente la cola de la derecha, se puede usar alternativamente la orden **di binomialtail**. Se puede observar que con esta orden se obtiene lo mismo:

```
. di binomialtail(10,8,0.2)
.00007793
```

### 3.13.2. La distribución de Poisson en STATA

Se empleará la orden **di poisson**(λ, k). En el caso de querer hallar la probabilidad de encontrar 0 enfermos en una población donde el número esperado de enfermos es cinco:

```
. di poisson(5,0)
.00673795
```

Del mismo modo que sucede con la distribución binomial, la orden **di poisson** dará siempre la probabilidad *acumulada* desde $k = 0$ hasta el valor de $k$ que se indique (cola de la izquierda). Para hallar la probabilidad de hallar exactamente $k$ eventos, se empleará la orden **di poissonp**(λ, k), y para calcular la probabilidad de encontrar al menos $k$ eventos, se podrá emplear la orden **di poissontail**(λ, k).

### 3.13.3. La distribución normal en STATA

Para calcular la probabilidad de observar a alguien con un valor de colesterol inferior a 180, en una distribución normal cuya media es 200 y su desviación estándar es 35, puede obtenerse fácilmente con STATA con la orden **di normal (z)**, pero antes debe calcularse $z$ como $(x - \mu)/\sigma$. STATA devuelve siempre el área bajo la cola de la izquierda. Se verá con dos ejemplos:

| Se introduce | STATA devuelve |
|---|---|
| di normal((180-200)/35) | .28385458 |
| di normal((220-200)/35) | .71614542 |

También existe otra función que usa la distribución normal de modo inverso, es decir, si se da a STATA la probabilidad (área bajo la cola izquierda), con la orden di invnormal(p), el programa devolverá el valor de $z$.

| Se introduce | STATA devuelve |
|---|---|
| di invnormal(.28385458) | –.57142858 |
| di invnormal(.71614542) | .57142858 |

A partir de este valor de $z$, se puede calcular el valor de $x$ como $x = \mu + z\sigma$. En los ejemplos: $x = 200 + (-0{,}571)*35 = 180$ y $x = 200 + 0{,}571*35 = 220$.

## 3.14. LAS DISTRIBUCIONES BINOMIAL, DE POISSON Y NORMAL EN EL PROGRAMA EXCEL

### 3.14.1. La distribución binomial en Excel

Para calcular las probabilidades de encontrar un número *k* de fumadores en una población donde la prevalencia *(π)* del uso del tabaco sea 0,2, se obtiene una tabla para todos los posibles valores de *k* con el programa Excel con los siguientes pasos:

1. Crear tres columnas: la primera, que corresponde a la letra «A», contendrá el número de «éxitos» (A = *k*); la segunda (B), el número de ensayos *n* (B = *n*), y la tercera, la probabilidad esperada de éxito (C = *π*). Se reserva la primera fila (casillas A1, B1 y C1) para los nombres de las columnas y se introducen los valores 1, 10 y 0,2 en las casillas A2, B2 y C2.
2. Introducir en la cuarta casilla (D2) la expresión: =DISTR.BINOM(A2;B2;C2;0).

En la casilla A2 debe figurar un valor de *k;* en la B2, el valor de *n*; en la C2, *π*. El último hueco de esta ventana exige que se introduzca la palabra FALSO (sustituible por un 0) o VERDADERO (sustituible por un 1). Como dice la ayuda, si se introduce FALSO, se calculará la probabilidad de encontrar exactamente *k* éxitos; si se introduce VERDADERO, se determinará la probabilidad de encontrar ≤ *k* éxitos. Una vez que se han completado los huecos de la ventana, basta con cambiar las casillas A2, B2 o C2 para que el ordenador calcule inmediatamente la probabilidad según el método binomial exacto. También se pueden copiar todas las casillas, incluida la fórmula, para los cálculos para diferentes valores de *k*. En la tabla 3.10 se recoge un ejemplo del resultado que se obtendría.

### 3.14.2. La distribución de Poisson en Excel

Para calcular las probabilidades de encontrar un número *k* de enfermos en una población donde el número esperado de enfermos es cinco, puede obtenerse fácilmente una tabla para todos los posibles valores de *k* con el programa Excel, con los siguientes pasos (figs. 3.20A y 3.20B).

1. Crear dos columnas: la primera, que corresponde a la letra «A», contendrá el número de «éxitos» (A = *k*); la segunda (B), el número de sucesos esperados *(λ)*. Reservamos la primera fila (casillas A1 y B1) para los nombres de las columnas e introducimos los valores 0 y 5 en las casillas A2, B2 y C2.
2. Introducir en la tercera casilla (D2) la expresión =POISSON(A2;B2;FALSO).

Al igual que con la binomial, si se introduce FALSO (=0), Excel calculará la probabilidad de encontrar exactamente *k* éxitos; si se introduce VERDADERO (=1), determinará la probabilidad de encontrar ≤ *k* éxitos.

***Tabla 3.10 Resultados obtenidos en 10 intentos con una probabilidad de éxito esperada de 0,2 mediante Excel***

| K | N | π | PROB. (K) | PROB. ACUM. |
|---|---|---|---|---|
| 0* | 10 | 0,2 | 0,10737 | 0,10737 |
| 1 | 10 | 0,2 | 0,26844 | 0,37581 |
| 2 | 10 | 0,2 | 0,30199 | 0,67780 |
| 3 | 10 | 0,2 | **0,20133** | **0,87913** |
| 4 | 10 | 0,2 | 0,08808 | 0,96721 |
| 5 | 10 | 0,2 | 0,02642 | 0,99363 |
| 6 | 10 | 0,2 | 0,00551 | 0,99914 |

En la penúltima columna se introdujo
=DISTR.BINOM(A2;B2;C2;FALSO) y en la última
=DISTR.BINOM(A2;B2;C2;VERDADERO)
*Interpretación:* por ejemplo, si la probabilidad de fumar es 0,2, la probabilidad de encontrar tres fumadores entre 10 sujetos es 0,2013. La probabilidad de encontrar tres o menos es 0,8791.

**Casilla A2.*

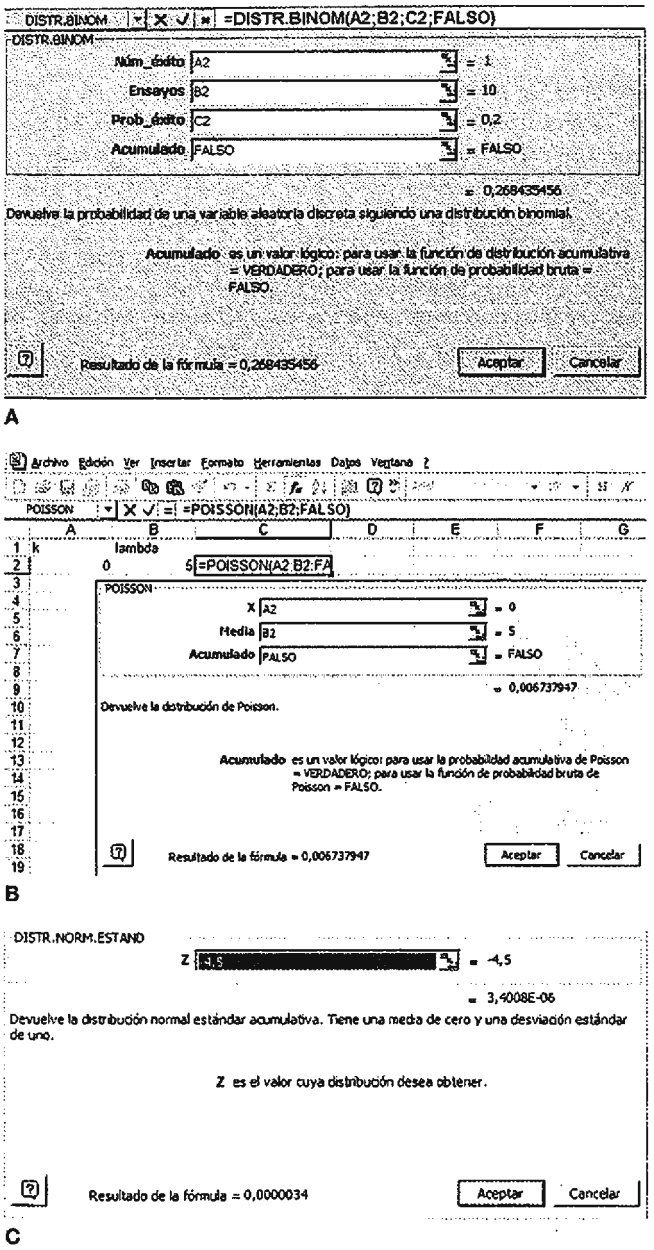



**Figura 3.20** Cómo realizar con Excel los cálculos de distribución binomial, de Poisson y normal. **A.** Distribución binomial. **B.** Distribución de Poisson. **C.** Distribución normal.

### 3.14.3. La distribución normal en Excel

Para calcular la probabilidad de observar a alguien con un valor de colesterol inferior a 180 en una distribución normal cuya media es 200 y cuya desviación estándar es 35, puede obtenerse fácilmente el resultado con el programa Excel introduciendo la siguiente expresión:

=DISTR. ORM(180;200;35;verdadero). Excel devuelve siempre el área bajo la cola de la izquierda, como muestran estos ejemplos:

| Se introduce | Excel devuelve |
|---|---|
| =DISTR.NORM(180;200;10;verdadero) | 0,0228 |
| =DISTR.NORM(220;200;10;verdadero) | 0,9772 |

También existe otra función que usa la distribución normal de modo inverso; es decir, si se da a Excel la probabilidad (área bajo la cola izquierda), el programa devuelve el valor de *z*.

| Se introduce | Excel devuelve |
|---|---|
| =DISTR.NORM.INV(0,0228;200;10) | 180,0 |
| =DISTR.NORM.INV((1-0,0228);200;10) | 220,0 |

De modo interesante, Excel contiene las funciones para todas las posibles situaciones de una distribución normal estandarizada.

Este procedimiento sustituye con gran ventaja a las habituales tablas de la distribución normal. Si se desea buscar, en la tabla de la normal, qué área deja a su izquierda un valor $z = -1{,}00$, en el interior de la tabla, se encuentra que el área es 0,1587. Con Excel, se obtiene usando la expresión = DISTR.NORM.ESTAND(–1) (fig. 3.20C).

| Se introduce | Excel devuelve |
|---|---|
| =DISTR.NORM.ESTAND(–1) | 0,1587 |
| =DISTR.NORM.ESTAND(+1) | 0,8413 |
| =DISTR.NORM.ESTAND(–4,5) | 0,000003 |

De nuevo, se comprueba que Excel se diferencia de la tabla en que siempre proporciona el valor de la cola de la izquierda. Si *z* es positivo, no dará la cola de la derecha, sino que habrá que calcular $1 - p$ para saber cuál es la probabilidad de la cola de la derecha. El último ejemplo, =DISTR. NORM.ESTAND(–4,5), se *sale* de las tablas. Nunca se hubiese podido resolver usando tablas. Por eso es interesante conocer cómo puede hacerse con Excel.

También se puede usar Excel para determinar el valor *z* que deja a su izquierda un área (probabilidad) determinada, como se recoge en los siguientes ejemplos.

| Se introduce | Excel devuelve |
|---|---|
| =DISTR.NORM.ESTAND.INV(0,1587) | –1,000 |
| =DISTR.NORM.ESTAND.INV(0,8413) | 1,000 |
| =DISTR.NORM.ESTAND.INV(0,000003) | 4,54 |

## 3.15. LAS DISTRIBUCIONES BINOMIAL, DE POISSON Y NORMAL EN OTROS PROGRAMAS

### 3.15.1. Las distribuciones binomial, de Poisson y normal en R/SPlus

En la tabla 3.11 se describe cómo usar las distribuciones binomial, de Poisson y normal en R/SPlus.

*Tabla 3.11 Distribuciones binomial, de Poisson y normal en R/SPlus*

| SE INTRODUCE | R DEVUELVE | COMENTARIO |
|---|---|---|
| **Distribución normal** | | |
| >**pnorm**(−1,96) | 0,02499790 | Para una $z = -1{,}96$, el área de la cola de la izquierda es 0,025 |
| >**pnorm**(1,96) | 0,9750021 | Para una $z = +1{,}96$, el área de la cola de la izquierda es 0,975 |
| >1-**pnorm**(1.96) | 0,02499790 | Para una $z = +1{,}96$, el área de la cola de la derecha es 0,025 |
| >**pnorm**(180,200,10) | 0,02275013 | Si $\mu = 200$ y $\sigma = 10$, la probabilidad de $x \leq 180$ es 0,0228 |
| >**pnorm**(220,200,10) | 0.9772499 | Si $\mu = 200$ y $\sigma = 10$, la probabilidad de $x \leq 220$ es 0,977 |
| >**qnorm**(0.025) | −1.959964 | Para un error alfa = 0,025, $z = -1{,}96$ |
| >**qnorm**(0.025,200,10) | 180.4004 | 180,4 deja una cola izquierda con $p = 0{,}025$, si $\mu = 200$ y $\sigma = 10$ |
| >**rnorm**(1000,200,10) | Crea 1.000 valores al azar de una distribución normal con $\mu = 200$ y $\sigma = 10$ | |
| **Distribución binomial** | | |
| >**dbinom**(3,10,0.2) | 0.2013266 | La probabilidad de tres éxitos en 10 ensayos con $\pi = 0{,}2$ es $p = 0{,}201$ |
| >**pbinom**(3,10,0.2) | 0.8791261 | La probabilidad de $k \leq 3$ éxitos en 10 ensayos con $\pi = 0{,}2$ es $p = 0{,}879$ |
| >**qbinom**(0.87,10,0.2) | 3 | En 10 ensayos con $\pi = 0{,}2$, el resultado con $p = 0{,}87$ es $k \leq 3$ |
| >**qbinom**(0.88,10,0.2) | 4 | En 10 ensayos con $\pi = 0{,}2$, el resultado con $p = 0{,}88$ es $k \leq 4$ |
| >**rbinom**(100,10,0.2) | Crea 100 valores al azar de una distribución binomial con $n = 10$ y $\pi = 0{,}2$ | |
| **Distribución de Poisson** | | |
| >**ppois**(0,5) | 0.006737947 | La probabilidad de $k = 0$ éxitos esperando $\lambda = 5$ es $p = 0{,}0067$ |
| >**ppois**(3,5) | 0.2650259 | La probabilidad de $k \leq 3$ éxitos esperando $\lambda = 5$ es $p = 0{,}265$ |
| >**dpois**(3,5) | 0.1403739 | La probabilidad de $k = 3$ éxitos esperando $\lambda = 5$ es $p = 0{,}140$ |
| >**qpois**(0.2650259,5) | 3 | Si lo esperado es $\lambda = 5$, el resultado con $p = 0{,}265$ es $k \leq 3$ |
| >**qpois**(0.265026,5) | 4 | Si lo esperado es $\lambda = 5$, el resultado con $p = 0{,}265$ es $k \leq 4$ |
| >**rpois**(100,5) | Crea 100 valores al azar de una distribución de Poisson con $\lambda = 5$ | |

## 3.16. APROXIMACIÓN A LAS DISTRIBUCIONES BINOMIAL Y DE POISSON CON LA DISTRIBUCIÓN NORMAL

En ciertas condiciones, se pueden resolver rápidamente problemas de la distribución binomial usando una aproximación mediante la normal. Todo se basa en aplicar la expresión ya conocida de la normal:

$$z = \frac{x - \mu}{\sigma}$$

pero sustituyendo la media por su esperanza matemática en una binomial y la desviación estándar por la raíz cuadrada de la varianza de una binomial. En la distribución binomial la esperanza matemática (~media) valdría:

$$\mu = n\,\pi$$

y la varianza sería:

$$\sigma^2 = n\,\pi\,(1-\pi)$$

Por lo tanto, la aproximación a la normal será:

$$z = \frac{x - n\,\pi}{\sqrt{n\,\pi(1-\pi)}}$$

Esta aproximación solo es válida cuando ambos productos, $n\,\pi$ y $n\,(1-\pi)$, son >5.

Se verá mejor con un ejemplo. Supóngase que se desea saber cuál es la probabilidad de encontrar 180 o menos hipertensos en una población de 1.000 personas, donde la prevalencia esperada $(\pi)$ de hipertensión es 0,20:

$$z = \frac{x - (n\,\pi)}{\sqrt{n\,\pi(1-\pi)}} = \frac{180 - (1.000 \times 0{,}2)}{\sqrt{1.000 \times 0{,}2 \times 0{,}8}} = \frac{-20}{12{,}65} = -1{,}58$$

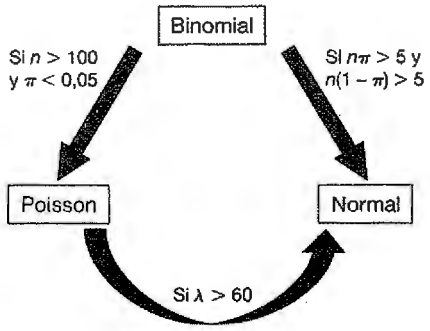



**Figura 3.21** Aproximaciones de una a otra distribución.

Mirando las tablas de la normal (o consultando Excel), se sabrá que, para $z = -1,58$, la probabilidad (área de la cola izquierda) es 0,057.

Si se calculase este mismo problema con la binomial, por ejemplo, introduciendo en Excel =DISTR.BINOM(180;1000;0,2;VERDADERO), la probabilidad obtenida sería 0,06. La aproximación no es exacta, pero aceptable. No obstante, siempre que esté disponible el ordenador, se deben resolver estos problemas con la binomial, ya que no solo es la distribución apta para variables discretas, sino que, además, el resultado que proporciona es exacto. La distribución de Poisson también se aproxima a la normal a medida que aumenta la muestra, y entonces se puede usar su media y su desviación estándar para hacer predicciones. Ahora, la media y la varianza corresponden a $\lambda$, y la expresión para calcular $z$ será:

$$z = \frac{x - \lambda}{\sqrt{\lambda}}$$

Si, por ejemplo, se desea conocer la probabilidad de observar 100 o más casos de una enfermedad en una población (grande, $n > 10.000$) donde se espera observar 85 casos:

$$z = \frac{x - \lambda}{\sqrt{\lambda}} = \frac{100 - 85}{\sqrt{85}} = +1,63$$

Para un valor $z = +1,63$, la distribución normal proporciona una probabilidad (área bajo la cola de la derecha en este caso) de $p = 0,052$. Si se calcula este problema por la distribución de Poisson, habría que escribir en Excel lo siguiente: =1-POISSON(99;85;1), y se obtendrá $p = 0,0607$. Sucede así porque Excel produce la cola de la izquierda y ahora se requiere la de la derecha. Se puso $k = 99$ dentro del paréntesis porque el 100 ya estaría incluido en la cola de la derecha, cuya área se desea calcular. Por eso no se resta de 1.

Este procedimiento solo es válido cuando $\lambda$ es grande, *al menos superior a 60.*

La figura 3.21 resume las aproximaciones desde una a otra distribución.

## 3.17. MEDIA Y DESVIACIÓN ESTÁNDAR DE UNA PROPORCIÓN

En las expresiones analizadas para calcular la esperanza matemática (~media) y la varianza de una distribución binomial, lo que interesaba era el *número absoluto* de éxitos *(k)*. Sin embargo, a veces se pretende conocer más bien la *proporción* de éxitos (porcentaje de caras al lanzar una moneda, de casos presentes en una población) y, entonces, las expresiones cambian tal como se recoge en la tabla 3.12, donde $p$ es la proporción observada en la muestra y $q = 1 - p$.

*Tabla 3.12 Índices estadísticos de una proporción*

| ÍNDICE ESTADÍSTICO | MUESTRAL | POBLACIONAL |
|---|---|---|
| Proporción (análoga a la media) | $p$ | $\pi$ |
| Varianza | $pq$ | $\pi(1-\pi)$ |
| Desviación estándar | $\sqrt{pq}$ | $\sqrt{\pi(1-\pi)}$ |
| Error estándar | $\sqrt{pq/n}$ | |

Se hace posible así, con muestras grandes (si $n\pi > 5$ y también $n(1-\pi) > 5$), realizar estimaciones mediante aproximaciones que usan la distribución normal. Por ejemplo, si en una población la proporción de fumadores fuese 0,25, ¿cuál sería la probabilidad de que en una muestra de tamaño 100 se obtuviese una proporción de 0,20 o todavía inferior? Como se pregunta por una muestra[3], se utilizará en el denominador *el error estándar* en vez de la desviación estándar, y $z$ valdrá:

$$z = \frac{p-\pi}{\sqrt{\frac{pq}{n}}} = \frac{0,2-0,25}{\sqrt{\frac{0,2\times 0,8}{100}}} = \frac{-0,05}{0,04} = -1,25$$

Mirando en las tablas de la normal, la cola que queda a la izquierda de $z = -1,25$ tiene un área de 0,1056. Esa será la probabilidad de encontrar muestras con $p \leq 0,2$, si se asume que la proporción poblacional $(\pi)$ es 0,25.

## REFERENCIAS


1. Greenland S. Probability logic and probabilistic induction. Epidemiology 1998;9(3):322-32.
2. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural bayesians. BMJ 2005;330(7499):1080-3. Erratum in: BMJ 2005; 330(7504):1369.
3. Martínez-González MA, Seguí-Gómez M, Delgado-Rodríguez M. ¿Cómo mejorar los intervalos de confianza? Med Clin (Barc) 2010;135(1):30-4.
4. Bland JM, Altman DG. Bayesians and frequentists. BMJ 1998;317(7166):1151.
5. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
6. Davidoff F. Standing statistics right side up. Ann Intern Med 1999;130:1019-21.
7. Goodman SN. Bayesian methods for evidence evaluation: are we there yet? Circulation 2013;127(24):2367-9.
8. Berry DA. Bayesian approaches for comparative effectiveness research. Clin Trials 2012;9(1):37-47.
9. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 1999;130(12):1005-13.
10. Altman DG, Bland JM. Statistics notes: the normal distribution. BMJ 1995;310(6975):298.
11. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23(1):151-69.


3 Matemáticamente sería más correcto usar en el denominador la raíz de $\pi(1-\pi)/n$ en vez de $pq/n$. Lo que sucede es que, en la práctica, la información de la que se dispone es la de la muestra ($pq$) y no la de la población.

# INTERVALOS DE CONFIANZA Y CONTRASTE DE HIPÓTESIS

4

*M. Á. Martínez-González, J. B. Toledo, J. López-Fidalgo*

## 4.1. ERROR SISTEMÁTICO Y ERROR ALEATORIO

Un primer paso imprescindible en toda investigación consiste en *medir* las variables. El segundo paso es estimar la asociación entre ellas. El método epidemiológico se ha convertido en una herramienta principal de investigación, y la epidemiología se concibe también como un *ejercicio de medición* de la ocurrencia de la enfermedad (o de la salud) en una población (**1**). Ya sea porque se estén recogiendo las variables básicas que serán los sillares sobre los que se construirá el trabajo de investigación, o bien porque se deseen identificar los determinantes de la ocurrencia de enfermedad, se necesitará realizar siempre un ejercicio de medición. El objetivo común es estimar determinados parámetros con el menor error posible, es decir, conseguir la máxima *exactitud* al medir.

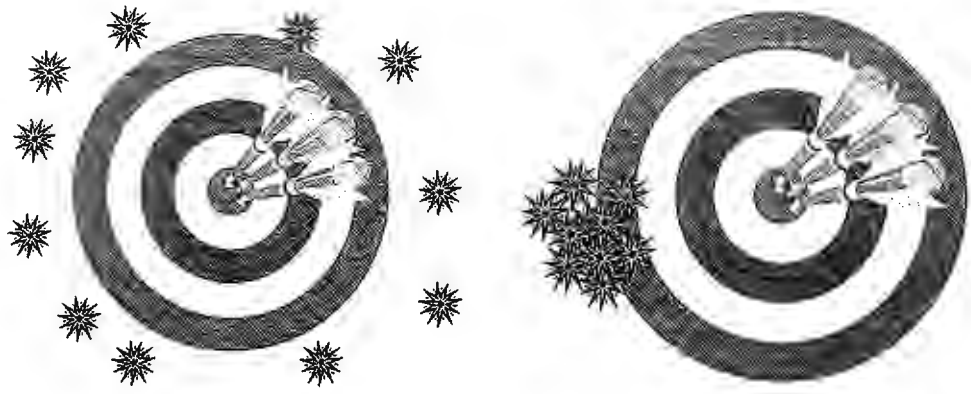
Los errores (faltas de exactitud) pueden clasificarse en dos tipos: sistemáticos o aleatorios. Por eso, la exactitud tiene dos componentes, *validez* y *precisión*. La validez evita los *errores sistemáticos*, y la precisión, los *errores aleatorios*. Los errores sistemáticos están producidos, por ejemplo, por un defecto del instrumento de medida o por una tendencia errónea del observador y, por tanto, tienden a registrarse en el mismo sentido; solo pueden ser puestos de manifiesto cambiando de aparato de medida o de observador. En cambio, los errores aleatorios o accidentales son aquellos debidos a pequeñas causas imponderables e imposibles de controlar; entre ellos se incluye el error cometido al extraer una muestra para sacar conclusiones que se apliquen a toda la población a partir de la misma *(error de muestreo)*.

Un ejemplo ayudará a entender la diferencia entre *validez* y *precisión*. Imagínese a un individuo que dispara a una diana. Si tiene mala puntería, sus disparos estarán muy separados unos de otros e irán a la periferia de la diana. Comete errores, pero son errores que pueden ir en cualquier dirección, son impredecibles. Si solo el tirador ve la diana, pero sus espectadores pueden ver los disparos, tras muchos disparos, los espectadores adivinarían dónde está el centro de la diana, al estar enmarcado por los disparos.

Supóngase ahora otra situación. Un tirador (ahora con buena puntería) usa una escopeta con un defecto de fábrica y sistemáticamente desvía los tiros hacia abajo y a la izquierda. Esta vez, si el tirador efectúa muchos disparos, estos quedarán muy juntos entre sí, pero seguirán lejos del centro de la diana. Si hay espectadores que solo ven los impactos, sin poder ver la diana, se equivocarán pensando que el centro de la diana está abajo y a la izquierda, es decir, en medio del espacio que circunscriben los disparos; es más, parecería falsamente que es más fácil saber dónde está el centro de la diana en esta situación. La segunda situación es más peligrosa que la primera. No solo comete errores el tirador, sino que induce a cometerlos a quienes le observan y, además, transmite la falsa imagen de que acierta casi siempre (fig. 4.1).

La primera situación se conoce como *error aleatorio* (falta de *precisión*); la segunda se denomina error sistemático (falta de *validez*) (tabla 4.1).

Las variaciones introducidas por una mala medición o un mal diseño de un estudio y que conducen a un error que tiende a desviarse de la verdad siempre en el mismo sentido se conocen por errores sistemáticos o *sesgos*, y conducen a una falta de *validez* (**2-5**). Las variaciones que ocurren por azar se llaman errores aleatorios y determinan el mayor o menor grado de *precisión* de un resultado.



A B

**Figura 4.1** **A.** Error aleatorio. **B.** Error sistemático = sesgo.

***Tabla 4.1 Diferencias entre error aleatorio y error sistemático***

| ERROR ALEATORIO | ERROR SISTEMÁTICO |
|---|---|
| 1. Impredecible | 1. Predecible |
| 2. Simétrico | 2. Asimétrico |
| 3. Inevitable, aunque estimable | 3. Corregible |
| 4. Equivale a falta de precisión | 4. Equivale a falta de validez |
| 5. Estimación y control → Estadística | 5. Prevención y control → Epidemiología |

El azar es un concepto muy usado, pero mal definido. A menudo, en estadística se equipara el azar a aquello que no puede explicarse, e incluso podría ser sinónimo de nuestra ignorancia (**1**). Cuanto mayor sea el error aleatorio (por azar), menos precisa resultará la estimación (más se alejará de la verdad). Los errores aleatorios producen observaciones desviadas en cualquier dirección, tanto por encima como por debajo del valor real, de modo que la media de los valores se acercará al valor real. El error que se comete al utilizar una muestra que se extrae a partir de una población se llama *error de muestreo* y, en principio, será aleatorio, siempre que la muestra se haya extraído al azar. El error aleatorio no es predecible y no puede ser eliminado, pero sí reducido mediante diseños más *eficientes* (que proporcionen mayor información sin necesidad de observar a más sujetos) o aumentando el tamaño de la muestra estudiada. El error aleatorio que persista puede ser estimado estadísticamente.

La *estadística* estima y controla el error aleatorio (**6,7**), mientras que la *epidemiología* se ocupa preferentemente de prevenir y controlar los sesgos o errores sistemáticos a través de un correcto diseño de las investigaciones y de las estrategias de recogida de datos (**1,3,8**). Para estimar y tener en cuenta el error aleatorio, en estadística se usan dos procedimientos, que son caras de una misma moneda: intervalos de confianza y pruebas de contraste de hipótesis. A ellos se dedica este capítulo.

## 4.2. MUESTREO ALEATORIO O SELECCIÓN ALEATORIA

Se suele trabajar con una muestra, no con toda la población. Como no suele ser factible, por motivos prácticos, determinar o medir la característica en *todas* las personas de la población, se usará solo un subgrupo, que se denomina muestra para, a partir de ella, describir la población. Además, esto no supone perder mucha información. A un investigador interesado en conocer la prevalencia de diabetes en la población adulta de Navarra, le interesaría una determinada precisión, pero se conformaría con saber que esta prevalencia se encuentra, por ejemplo, entre 0,07 (7%) y

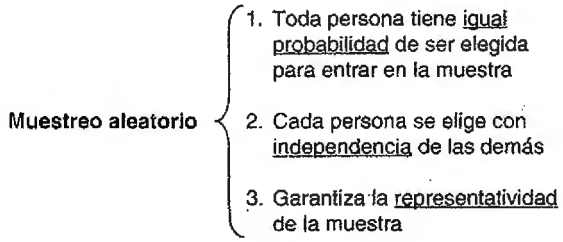



**Figura 4.2** Muestreo aleatorio.

0,08 (8%). Se quedaría conforme y deduciría que aproximadamente el 7,5% de la población, en cifras redondas, es diabética. En cambio, sería un despilfarro que, para mejorar esta precisión, se dedicasen recursos a determinar en *toda* la población adulta de Navarra la glucemia para detectar a *todos* los diabéticos. Resultaría ineficiente, porque con una muestra de varios miles de sujetos ya se lograría suficiente precisión.

El problema práctico más relevante en esta situación es que hay que elegir una muestra que sea *representativa*. Esto significa que todos los candidatos para su elección deben estar representados por igual en los incluidos finalmente en la muestra y, también, que la selección de una persona no influya en la probabilidad de que otra persona también sea seleccionada (independencia). Cuando el proceso de obtención de una muestra cumple estas dos características, se habla de muestreo aleatorio y se garantiza la representatividad (fig. 4.2).

El conjunto de todos los elegibles recibe el nombre de «universo» o población diana. Por *marco muestral* se entiende, sin embargo, la parte de la población de la que realmente se va a extraer la muestra y que debe garantizar la representatividad. Sería el caso de un listado completo de toda la población, como un censo o un padrón. Para que el muestreo sea científico, todas las personas de la población diana deben tener una probabilidad conocida de ser elegidos. Estas probabilidades serán iguales para todas ellas si se trata de un muestreo aleatorio simple. A veces puede interesar que un grupo esté más representado, en cuyo caso se hará el muestreo de modo que los sujetos de ese grupo tengan mayor probabilidad de ser elegidos; esta decisión debe tenerse en cuenta después al analizar e interpretar los datos.

Una muestra aleatoria se crea mediante la asignación de un identificador (número) a cada persona del marco muestral. Posteriormente se seleccionarán los identificadores mediante un sistema que proceda al azar, como la tabla de números aleatorios o una secuencia al azar generada por ordenador (fig. 4.3).

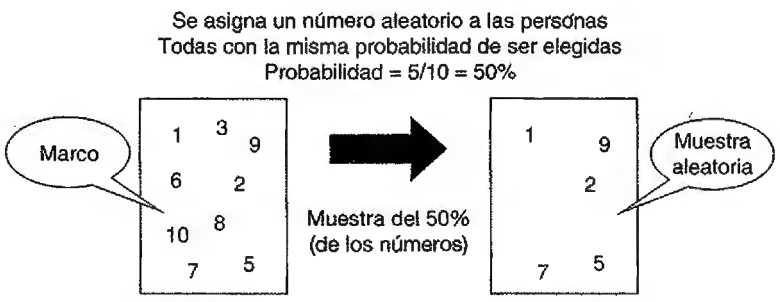



**Figura 4.3** Muestra aleatoria simple.

Una *tabla de números aleatorios* debe cumplir dos condiciones:

- Cada número (p. ej., el 3 o el 9) tiene la misma probabilidad de aparecer en ella que los demás.
- La elección de cada número es independiente de la de cualquier otro número de la tabla (no existe ninguna secuencia o combinación que pueda servir para predecirlos).

La tabla 4.2 recoge una secuencia de 300 números aleatorios.

***Tabla 4.2 Secuencia de 300 números aleatorios***

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 9 | 0 | 2 | 3 | 7 | 1 | 3 | 5 | 7 | 7 | 4 | 9 | 4 |
| 4 | 6 | 5 | 2 | 7 | 8 | 1 | 2 | 1 | 4 | 1 | 3 | 2 | 6 | 4 |
| 9 | 6 | 3 | 3 | 0 | 7 | 2 | 4 | 4 | 7 | 6 | 5 | 1 | 5 | 3 |
| 4 | 6 | 8 | 9 | 1 | 5 | 2 | 4 | 8 | 5 | 2 | 1 | 8 | 6 | 4 |
| 6 | 0 | 2 | 5 | 0 | 7 | 8 | 5 | 0 | 8 | 2 | 1 | 0 | 0 | 3 |
| 2 | 1 | 3 | 9 | 7 | 1 | 1 | 4 | 5 | 2 | 9 | 5 | 2 | 0 | 8 |
| 1 | 9 | 8 | 3 | 5 | 5 | 3 | 1 | 5 | 2 | 4 | 1 | 9 | 6 | 8 |
| 0 | 2 | 5 | 9 | 7 | 1 | 9 | 8 | 2 | 7 | 6 | 7 | 5 | 3 | 5 |
| 0 | 6 | 8 | 0 | 8 | 0 | 9 | 7 | 9 | 5 | 4 | 8 | 2 | 7 | 2 |
| 1 | 5 | 8 | 1 | 4 | 7 | 0 | 8 | 9 | 2 | 6 | 5 | 4 | 8 | 5 |
| 9 | 0 | 7 | 9 | 3 | 9 | 9 | 6 | 4 | 9 | 1 | 0 | 0 | 7 | 7 |
| 7 | 3 | 9 | 3 | 1 | 0 | 7 | 6 | 8 | 4 | 5 | 1 | 4 | 8 | 5 |
| 4 | 7 | 4 | 8 | 1 | 1 | 1 | 6 | 2 | 0 | 7 | 4 | 8 | 3 | 7 |
| 2 | 4 | 4 | 6 | 9 | 4 | 3 | 8 | 6 | 7 | 2 | 1 | 5 | 9 | 5 |
| 2 | 0 | 6 | 3 | 0 | 0 | 7 | 8 | 3 | 6 | 0 | 4 | 1 | 2 | 3 |
| 3 | 3 | 3 | 7 | 6 | 4 | 3 | 8 | 2 | 6 | 6 | 9 | 4 | 3 | 5 |
| 5 | 2 | 9 | 5 | 5 | 1 | 8 | 0 | 3 | 8 | 1 | 4 | 9 | 1 | 6 |
| 3 | 3 | 5 | 6 | 7 | 9 | 7 | 7 | 0 | 3 | 7 | 0 | 0 | 6 | 9 |
| 3 | 0 | 8 | 8 | 4 | 2 | 0 | 8 | 8 | 2 | 6 | 1 | 2 | 7 | 6 |
| 0 | 9 | 2 | 9 | 8 | 6 | 9 | 0 | 6 | 5 | 1 | 4 | 9 | 6 | 6 |

¿Cómo puede usarse esta tabla para seleccionar a los participantes que formarían parte de la muestra para investigar la prevalencia de diabetes tipo 2 en Navarra? Se formaría un listado de los candidatos a participar en la muestra y a cada uno se le asignaría un número consecutivo, por ejemplo, del 1 al 500.000. Si se desease obtener una muestra de 20.000 participantes, dentro de cada millar se deberían elegir 50 personas para formar parte de la muestra. ¿Qué 50 personas se eligen? Aquellas cuyos tres últimos dígitos coincidan con las tres primeras columnas de la tabla 4.2; así, en el primer millar serán elegidas como participantes las personas cuyos números en el listado sean el 239, el 023, el 713, el 577, el 494, etc., hasta completar 50 sujetos extraídos de los 1.000 primeros. Habrá que repetir este proceso hasta obtener los 20.000 participantes.

### 4.2.1. Muestra al azar con Excel

Habitualmente, es más sencillo recurrir a un ordenador, ya que muchos programas contienen funciones que producen números seudoaleatorios, como la función matemática =ALEATORIO() incorporada en Excel. Al aceptar esta función, aparecerá en esta celda un número aleatorio con un valor entre 0 y 1. A continuación, situando el cursor en la esquina inferior derecha de esa primera celda (el cursor se transforma entonces en un signo positivo), se arrastra hasta la última persona en la base de datos. Aparecerán números aleatorios que se distribuyen uniformemente entre las personas de la base de datos.

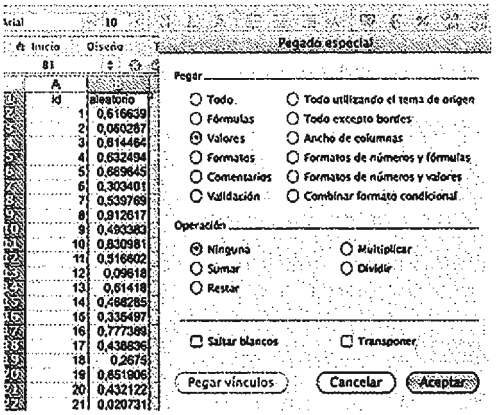
Imagínese que se dispone de 5.000 filas y se desea seleccionar al azar el 0,5% de ellas, es decir, 25. Una vez creada la columna de números aleatorios, como se ha dicho antes, se copiará esta columna y se pegará en la misma columna siguiendo las instrucciones:

Edición → Copiar

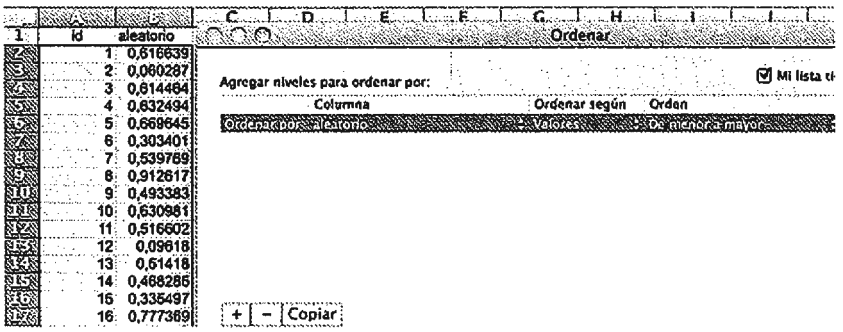
Edición → Pegado especial → Pegar → Valores

Así, en cada celda quedará un único valor del número aleatorio sin la función subyacente responsable de crear dicho número. Basta con ordenar a continuación los valores en orden ascendente: la muestra aleatoria estará compuesta por las 25 primeras filas.

=aleatorio() y después copiar y pegado especial.




Ordenar de menor a mayor por la segunda columna.




Seleccionar los 25 sujetos con el menor valor aleatorio.

| id | aleatorio |
|---|---|
| 70 | 0,002818 |
| 361 | 0,004208 |
| 73 | 0,005274 |
| 185 | 0,00888 |
| 403 | 0,011651 |
| 57 | 0,014508 |
| 120 | 0,016199 |
| 481 | 0,0162 |
| 21 | 0,020731 |
| 203 | 0,023914 |
| 489 | 0,02707 |
| 299 | 0,032901 |
| 131 | 0,034757 |
| 275 | 0,042461 |
| 308 | 0,048544 |
| 471 | 0,051003 |
| 178 | 0,052028 |
| 343 | 0,055906 |
| 217 | 0,05654 |
| 318 | 0,056673 |
| 47 | 0,057408 |
| 2 | 0,060287 |
| 229 | 0,060635 |
| 143 | 0,061266 |

Hasta aquí los 25 seleccionados, la muestra al azar incluirá los id 70, 361, 73, ..., 47, 2 y 229

### 4.2.2. Muestra al azar con STATA

La orden **sample** (que puede ser peligrosa y no debe aplicarse hasta haber leído todo este apartado) selecciona una muestra al azar. Debe ir seguida de un número que indique el porcentaje de sujetos que se desea que queden en la muestra aleatoria; *los demás se borrarán.* También puede ir seguida de un número más la opción **count;** entonces, el número indicará el número de sujetos que quedarán exactamente en la muestra. En una base de datos con 5.000 sujetos, para extraer una muestra del 0,5% (25 sujetos) se logra lo mismo con cualquiera de las dos órdenes siguientes:

```
sample 0.5

sample 25, count
```

Es una orden *peligrosa,* porque *borra* toda la base de datos salvo la muestra aleatoria que se obtenga. Hay que salvar antes a buen recaudo la base de datos original y guardar la nueva muestra al final *con otro nombre.*

### 4.2.3. Muestra al azar con SPSS

Este procedimiento también puede llevarse a cabo con SPSS, con los siguientes pasos:

**Datos → Seleccionar Casos → Muestra aleatoria de casos → Aproximadamente |_|% casos**

Con sintaxis, basta escribir:

```
COMP muestra=(UNIFORM(1)<=.005).

FILTER BY muestra.

EXE.
```

### 4.2.4. Muestra al azar con R

En R se usa la instrucción **sample.** En este caso se guarda la muestra en un nuevo vector (*Sujetos-Muestreados*). Dentro de la función **sample,** se incluye primero el vector que se desea muestrear (*TodosSujetos*) y, a continuación, el número de sujetos que se desea seleccionar al azar. Para obtener un porcentaje se multiplica el número de sujetos, obtenido mediante la función **length,** por el porcentaje, y se eliminan decimales con la función **round.**

```
SujetosMuestreados<- sample(TodosSujetos,25)

SujetosMuestreados<- sample(TodosSujetos,
+round(0.05*length(TodosSujetos)))
```

### 4.2.5. Interpretación de las muestras obtenidas al azar

Será interesante detenerse ahora y realizar como prueba un muestreo aleatorio de una base de datos creada personalmente.

Por ejemplo, de una base de datos total ***(marco)*** que incluía a 21.325 personas, se seleccionó una muestra aleatoria del 0,5%. Las medias de la talla fueron 168,4 en el universo y 168,2 en la muestra. La mediana fue 168 en ambas. La diferencia entre lo obtenido en el universo y en la muestra se denomina *error de muestreo.* La diferencia (0,2 = 168,4 – 168,2) entre la media poblacional y la media muestral es el *error de estimación de la media* debido al muestreo. Los *errores de muestreo* no se desvían siempre en la misma dirección, sino en direcciones imprevisibles; en este ejemplo, la muestra se desvió hacia abajo. Otra nueva muestra podría desviarse

hacia arriba. El *error de muestreo* no es sistemático, sino aleatorio. La estadística gira en torno al error de muestreo. La selección aleatoria implica que solo el azar será responsable de que unos individuos estén en la muestra y otros no. Garantiza la representatividad (a no ser que el azar juegue una mala pasada).

En el fondo, viene a equivaler a un sorteo o lotería en el que el *premio* consiste en ser elegido como participante en un estudio de investigación científica. Todos tienen las mismas oportunidades de ganar. El error sistemático, en cambio, sería como jugar sucio (una rifa o sorteo donde hubiese truco), para que unos tengan más probabilidades de ganar el premio que otros.

En la práctica, hay dificultades para contar con un acceso completo al *universo* desde el cual extraer la muestra, salvo en casos en que se pueda contar con un *censo* exhaustivo. Lo habitual es que el marco muestral utilizado *no sea exhaustivo* y resulte imposible un muestreo aleatorio estricto. Por ejemplo, extraer una muestra de la guía telefónica sería usar un marco muestral no exhaustivo. Además, estar incluido en la guía telefónica implica diferencias en clase social, situación laboral, edad, etc. También influye la participación selectiva (nunca acepta participar el 100% y siempre participan los más entusiastas). Debe reconocerse, por tanto, que casi nunca existen muestras verdaderamente *aleatorias y representativas* de una ciudad, región o país, pero este defecto ha de minimizarse convenientemente, o al menos tenerse en cuenta en los análisis correspondientes.

### 4.2.6. Juicio crítico sobre las muestras representativas

Las muestras representativas escasean. En la vida real ninguna muestra es verdadera y estrictamente representativa de una población. ¿Qué problemas provoca esto? Las consecuencias pueden ser graves cuando el objetivo del estudio es responder a preguntas *descriptivas* (¿cuál es el colesterol medio en la población?, ¿qué porcentaje de mujeres usan el método sintotérmico?, etc.). El objetivo de las investigaciones *descriptivas* no es realizar comparaciones, sino calcular medias o proporciones. Exigen representatividad.

En cambio, el problema no suele ser tan grave cuando se trata de estudios analíticos o inferenciales, que se asume que no pretenden describir a la población, sino obtener comparaciones científicamente válidas. Buscan la verdad abstracta y universal.

Hay que distinguir entre validez *interna* y *externa*. La validez *interna* consiste en que lo hallado coincida con la verdad en la muestra que se estudia. La validez *externa* reside en que coincida con la verdad en una población diana mucho más amplia que la muestra. Los estudios analíticos o inferenciales tratarán de asegurar, sobre todo, la validez *interna* para los participantes en el estudio (¿es mejor el tratamiento médico o el tratamiento quirúrgico en estos pacientes?; ¿puede ser que el consumo de café proteja frente a la diabetes?; ¿aumenta el riesgo de cáncer de mama si es tardío el primer parto?, etc.). En estas situaciones, aunque la muestra no sea perfectamente representativa, puede solventarse en un segundo paso el problema de la extrapolación o generalización a otras poblaciones. Los investigadores no deberían incurrir en una preocupación obsesiva por buscar metas utópicas en pro de garantizar la representatividad estadística. Esta preocupación por la *representatividad* a veces ha podido hacer más daño que bien en estudios analíticos, tal como ha denunciado Rothman, quien llega a afirmar taxativamente que:

> *La representatividad es una falacia que ha contaminado los estudios epidemiológicos durante décadas*[1] *(9).*

Quizá se requiera matizar esta afirmación, pero, incluso con matices, no deja de ser cierto que muchos de los descubrimientos más importantes de la epidemiología analítica se han realizado en

1 Literalmente, «representativeness is a fallacy that has plagued epidemiologic studies for decades».

muestras que *no eran representativas* en el sentido estadístico del término. Por ejemplo, la fuerte relación tabaco-cáncer de pulmón se puso de manifiesto en una muestra en la que solo había varones y todos ellos eran médicos. Evidentemente, no era una muestra representativa. Así ha sucedido también con otras asociaciones importantes para la salud pública. Para la búsqueda de una asociación en estudios inferenciales o analíticos, que no pretenden una mera descripción sino ir más allá y encontrar relaciones científicamente sólidas y universalizables, puede optarse (*y muchas veces debe* optarse) por una muestra *no* representativa. Se actúa así para garantizar la validez *interna,* por ejemplo, para seleccionar sujetos con características homogéneas entre ellos y que sean más comparables entre sí, o para restringir el estudio a voluntarios altamente seleccionados que proporcionarán una información autorreferida de óptima calidad.

Estas y otras restricciones previenen muchos sesgos que amenazan la validez interna (**10**). En estudios inferenciales o analíticos debe asegurarse, sobre todo, la validez *interna,* ya que sin validez interna no tiene sentido preguntarse por la externa. La validez interna es el requisito previo e imprescindible. Después, debe tenerse en cuenta que generalizar *no* es un proceso estadístico *automático.* Unas conclusiones con alta validez interna se suelen generalizar a toda la humanidad en un segundo paso en función del *conocimiento biológico,* no de una aplicación automática de principios estadísticos o por el mero criterio de representatividad. A veces será preferible que participen en un estudio solo las personas más motivadas y conscientes, porque serán las que proporcionarán una información más fiable, aunque la baja participación reduzca la *representatividad* en el sentido estadístico del término. Por eso, a veces (no siempre, desde luego), los estudios con una baja proporción de candidatos que aceptan participar pueden ser más válidos que si participase una alta proporción de los invitados (**11**). La ciencia que se adquirió sobre tabaco-cáncer de pulmón en una muestra de médicos varones es generalizable a toda la humanidad como conocimiento universal y abstracto por las siguientes razones:

1. No hay ningún motivo biológico para pensar que si el tabaco causa cáncer en los varones no lo vaya a causar en mujeres. ¿Es acaso distinto el tejido pulmonar en los hombres y en las mujeres en cuanto a su susceptibilidad a cancerígenos? Responder afirmativamente a esta pregunta supondría olvidar la histopatología pulmonar común en los dos sexos, hombre y mujer.
2. No hay ningún motivo biológico para pensar que si el tabaco provoca cáncer en quienes son intelectuales o de clase alta no lo vaya a hacer también en clases bajas o en quienes no estudiaron medicina. ¿Cambia el tejido pulmonar al acabar la carrera de medicina?
3. Sucesivos estudios de otros países realizados en otras muestras, que tampoco son por sí mismas representativas, corroboraron que cuanto más se exponían las personas al tabaco mayor era su riesgo de cáncer pulmonar (consistencia).
4. Estudios en animales demostraron que el humo del tabaco era cancerígeno (y eso reafirmó la causalidad, aunque, lógicamente, los estudios en animales tampoco son *representativos* de los seres humanos).
5. El análisis químico del humo del tabaco halló altas concentraciones de cancerígenos.
6. En quienes dejaban de fumar se observó una reducción del riesgo de cáncer pulmonar y en personas que nunca fumaron pero convivieron mucho con fumadores (fumadores pasivos) también aumentaba el riesgo de cáncer pulmonar.

Todas estas razones no son de *representatividad* rígidamente considerada en sentido automático, sino de conocimiento biológico básico y de epidemiología y estadística bien entendidas. Son la base de la validez externa y permiten fundamentar el establecimiento de conclusiones científicas válidas, universales y abstractas. Por otra parte, no debe olvidarse que

los mejores ensayos clínicos que más han cambiado la práctica médica nunca usaron muestras representativas.

## 4.3. DIFERENCIA ENTRE SELECCIÓN ALEATORIA Y ASIGNACIÓN ALEATORIA (ALEATORIZACIÓN)

El término aleatorización (*randomization*, en inglés) abunda en la bibliografía biomédica. También se habla de diseños *aleatorizados* (la palabra «randomizado» no existe en castellano). Aleatorización es el *reparto* o asignación de una muestra al azar en dos o más grupos. Si el tamaño de cada grupo es suficientemente grande, tal asignación al azar garantizará que los grupos así obtenidos serán idénticos en sus características conocidas, y también en las *desconocidas;* en variables medidas y no medidas. En consecuencia, esos grupos son intercambiables y perfectamente comparables entre sí. Si una vez obtenidos estos grupos idénticos por aleatorización se introduce solo en uno de ellos un factor diferencial, entonces todas las diferencias que se observarán más tarde entre los grupos se deberán *causalmente* a ese factor diferencial y no pueden ser achacables a ninguna otra causa, porque los grupos eran idénticos en todo lo demás. Este es el fundamento de los *ensayos clínicos aleatorizados* (**12**). La aleatorización es la herramienta más poderosa en investigación para lograr demostrar relaciones causa-efecto. No siempre será ético o factible usar un diseño aleatorizado, pero los estudios aleatorizados y correctamente realizados, cuando son éticos y factibles, tienen la última palabra para establecer una relación causa-efecto.

No debe confundirse la aleatorización *(asignación)* con la *selección* aleatoria antes vista. La selección aleatoria consiste en extraer un pequeño subgrupo de sujetos (muestra) al azar desde una población. Se obtiene una muestra que garantice que todos los que estaban en la población tenían la misma probabilidad de entrar en la muestra. La aleatorización en cambio reparte toda la muestra en grupos iguales. La selección aleatoria se usa para obtener una muestra representativa en estudios descriptivos. La aleatorización pretende crear varios grupos equiparables entre sí (intercambiables) para hacer estudios analíticos.

¿Qué tamaño ha de tener la muestra para garantizar que los grupos sean intercambiables? La respuesta sensata es admitir que *cuanto más grande, mejor.* Como regla rápida:

- Con menos de 100 sujetos por grupo (200 en total si son 2 grupos) puede haber diferencias de al menos un 10% en más de un 20% de las variables y, probablemente, la aleatorización no conseguirá del todo su propósito.
- Entre 100 y 300 sujetos (total: 200 o 600, respectivamente), la aleatorización habrá ayudado, pero los investigadores deben ser conscientes de que seguirá habiendo variables que no se distribuyen por igual en los grupos comparados.
- Si *cada* grupo tiene más de 300 sujetos (por ejemplo, un ensayo con >600 participantes y >300 en cada grupo), la intercambiabilidad puede darse por asumida desde el punto de vista práctico, tanto para variables medidas como no medidas.

### 4.3.1. Asignación aleatoria (aleatorización) con STATA

Se pueden usar las siguientes instrucciones con STATA:

```
generate grupo=round(uniform())
```

Esta orden creará una columna con números al azar que solo podrán ser el 0 o el 1. Se asignarán los sujetos con 0 a un grupo y los sujetos con 1 al otro grupo. Si se desease asignar 3 grupos al azar, se haría así:

```
generate grupo3 =1+floor(3*uniform())
```

Los posibles valores para ***grupo3*** serán el 1, el 2 y el 3, con el mismo número de sujetos aproximadamente en cada grupo. Se sugiere realizar el siguiente ejercicio: abrir un *Do-file* en STATA y ejecutar la siguiente secuencia de órdenes:

```
clear                    #borra datos previos

set obs 1000             #crea base de datos vacía con 1.000 filas

gen id=_n                #asigna id =.n.º de fila a cada uno

set seed 1234            #fija números aleatorios repetibles
                           (si se vuelve a ejecutar dará idéntico
                           resultado para números aleatorios)

g grup2=1+floor(2*uniform()) #genera 2 grupos al azar
                              de ≈ tamaño

g grup3=1+floor(3*uniform()) #genera 3 grupos al azar
                              de ≈ tamaño

g grup4=1+floor(2*uniform()) #genera 4 grupos al azar
                              de ≈ tamaño

tabl grup2-grup4         #tabula los grupos al azar

                         (lo que sigue a tab es uno;
                         sirve para pedir frecuencias)

tabl grup*               #logra lo mismo que la orden anterior
                           (el asterisco funciona como comodín)
```

Se logra también con un bucle, con las siguientes órdenes:

```
clear

se ob 1000

se se 1234

foreach k of numlist 2/5      {

g grup`k'=1+floor(`k'*uniform())

ta grup`k'

}
```

### 4.3.2. Asignación aleatoria (aleatorización) con R

Usando de nuevo **sample,** se indica la secuencia de números que se desea obtener desde el nivel inferior, seguido de «:» hasta el nivel superior; a continuación se indica el número de observaciones

que se desea generar. Finalmente, se señala que se utilizarán muestras con reemplazamiento para que los números puedan ser seleccionados más de una vez:

```
grupo2< -sample(0:1, 100,replace=TRUE)

grupo3< -sample(0:2, 100,replace=TRUE)

grupo4< -sample(0:3, 100,replace=TRUE)
```

## 4.4. CONCEPTOS GENERALES SOBRE ESTIMACIÓN DE PARÁMETROS

La bibliografía biomédica estima medidas de asociación entre exposiciones y sus efectos sobre la salud. Ofrece así una gran contribución a la salud pública, pues un mejor conocimiento puede traducirse en medidas preventivas. Por ejemplo, en 2011 se publicó que una dieta mediterránea suplementada con aceite de oliva virgen reducía al cabo de 1 año el grado de arteriosclerosis en la carótida comparada con una dieta control. Entre quienes al inicio tenían engrosada la íntima media carotídea (>0,9 mm) se encontraba una reducción de –0,093 mm con la dieta mediterránea y aceite de oliva virgen, superior a la reducción de –0,014 mm observada en el grupo control (**13**). Los autores acompañaban esta estimación de una frase que podría expresarse así: *con un intervalo de confianza del 95%, la reducción de la íntima media carotídea en el grupo de dieta mediterránea con aceite de oliva virgen iba de –0,146 a –0,039*. Proporcionaban una estimación puntual (–0,093 mm) y unos márgenes de error hacia abajo y hacia arriba (de –0,146 a –0,039 mm). Habían hecho una *estimación*, en concreto, una estimación *por intervalo*. Indicaban un intervalo de confianza, que es el *rango de posibles efectos compatibles con los datos* (**14**). Este intervalo da una idea de la magnitud que puede tener el efecto. Se confía en que, si se pudiese comparar a *todas* las personas que siguen esta dieta y tienen arterioesclerosis, la verdadera reducción de la íntima media carotídea estará en algún punto dentro del rango entre una reducción de –0,039 y –1,146 mm.

Una estimación consiste en basarse en la información contenida en la muestra para *apostar* por un valor para un parámetro que es real y que existe en toda la población, pero se desconoce. Se calcula un rango u horquilla de posibles valores, entre los cuales puede situarse el verdadero parámetro poblacional, y se confía (con una confianza cuantificada) en acertar. A esa horquilla o rango de valores suele llamársele *intervalo de confianza*. Un *intervalo de confianza* es un *rango de valores en que se confía que contenga el parámetro poblacional* (**15**).

Los intervalos de confianza permiten presentar un resultado acompañándolo de un margen de error, con un límite superior y otro inferior (**16-18**). Estos intervalos se usan como procedimiento habitual para estimar parámetros de una población.

Todo el proceso parte de los correspondientes *estimadores* muestrales. Un *estimador* es una función de los valores de una muestra que se elabora para indagar acerca del valor de un parámetro desconocido de la población de la que procede la muestra (**19**). Los *estimadores* son conocidos y se calculan en muestras. Los *parámetros* se refieren a la población y suelen ser desconocidos. Para ellos se suelen reservar las letras griegas. Ejemplos de parámetros y estimadores son la media poblacional y la muestral, la proporción poblacional y muestral, la mediana poblacional y muestral, o la varianza poblacional y muestral. Los intervalos de confianza se construyen a partir de los estimadores.

El *error estándar* es el error de estimación. Es un concepto central en los intervalos de confianza y se usa habitualmente para calcularlos. A menudo, para hallar un intervalo de confianza al 95% bastará con sumar y restar dos veces *(aproximadamente)* el error estándar al estimador muestral:

$$\text{Intervalo de confianza al 95\%} = \text{estimador} \pm (2 \times \text{error estándar})$$

Un error estándar (p. ej., el error estándar de la media, que se calcula dividiendo la desviación estándar entre la raíz cuadrada de *n*) no expresa la variabilidad de los *individuos*, sino la variabilidad de las *medias* (estimadores) calculadas en muchas posibles *muestras* que se tomen de una población, todas ellas de tamaño *n* (v. apartados 2.6.4 y 3.11). Esto mismo se aplica a otros estimadores distintos de la media: proporciones, varianzas, diferencias de medias, diferencias de proporciones, asimetría, curtosis, riesgos relativos, etc.

*El error estándar es al estimador muestral lo que la desviación estándar es al individuo.* El error estándar mide el grado de incertidumbre con respecto a la capacidad del estimador muestral para estimar el parámetro poblacional. Si la edad media de una población de universitarios es 22 años y su desviación estándar es 10, y se toman repetidas muestras (p. ej., 500), todas ellas de tamaño 100, el error estándar de la media valdrá 10/100^0,5 = 1 y se esperará que las 500 medias formen una distribución normal cuya desviación estándar será 1; por tanto, el 95% de estas muestras (475 muestras) tendrán medias entre 20 y 24 años.

Lo interesante de esta propiedad es que se puede aplicar en camino inverso cuando *no se sabe* cuál es realmente la media poblacional y solo se conoce una media muestral, digamos $\bar{x} = 21$ años. En esta situación, que es la que ocurre en la práctica, la única posibilidad es *confiar* en que esta media muestral ($\bar{x} = 21$) esté entre ese 95% de medias muestrales situadas en el entorno de ±2 errores estándar de la verdadera media poblacional. Simplemente sumándole y restándole 2 errores estándar se obtiene un rango de valores (de 19 a 22 años) en el que se *confía* (con un 95% de confianza) que se encontrará la verdadera media poblacional. Si se *repite todo el proceso 100 veces*, aproximadamente *95 intervalos así calculados contendrán la verdadera media poblacional.* Con un intervalo de confianza se puede acertar o errar. El acierto consiste en que el parámetro esté incluido en el intervalo, y la equivocación, en que el parámetro quede fuera. Cuando se calculan intervalos de confianza al 95%, acertar depende de si la muestra concreta que se ha usado para calcular el intervalo es una de ese 95% de muestras que caen a menos de 2 errores estándar del verdadero parámetro o es de ese otro 5% que se sitúa más lejos. Si alguien todos los días durante 100 días calculase un intervalo de confianza al 95%, en 95 de esos días su intervalo contendrá el verdadero parámetro (habrá tenido *un buen día*), pero en los otros 5 días la verdad poblacional o parámetro quedará fuera de su intervalo y su predicción o apuesta sobre el parámetro será errónea (tendrá *un mal día*).

La expresión más general para hacer estimaciones calculando intervalos de confianza se basa en sumar y restar al estimador muestral *z* veces el error estándar (EE):

$$\text{Parámetro} \in \text{estimador} \pm z \times \text{EE del estimador}$$

Aquí, *z* es el valor correspondiente de la distribución normal (v. apartado 3.10). Si el intervalo de confianza es al 95% (en general, $1 - \alpha$), habrá un 5% de error (en general, $\alpha$). Este error *alfa* ($\alpha$) se reparte en dos colas, una a cada lado. Cada cola valdría el 2,5% ($\alpha/2 = 0{,}025$); en ese caso, *z* valdría 1,96. A veces, en vez de *z* se usa otra cantidad (*t* de Student, como se verá en los apartados 4.6.2 y 6.1), debido a que no se dispone de *sigma* (desviación estándar *poblacional*), sino solo de $\sigma$ (muestral). Si el intervalo de confianza fuese al 90%, entonces, $\alpha = 0{,}10$, $\alpha/2 = 0{,}05$ y $z = 1{,}645$.

## 4.5. ESTIMACIÓN DE UNA PROPORCIÓN

El uso de proporciones, expresadas coloquialmente como *porcentajes o tantos por ciento* (en vez de tantos por uno), es muy frecuente, sobre todo en medicina: la proporción o porcentaje de fumadores que desarrollarán cáncer de pulmón, el porcentaje de usuarias de contraceptivos que presentarán una trombosis, la proporción de habitantes de un país africano que están infectados por el virus del sida, la proporción de alumnos de bioestadística que aprobarán la asignatura en primera convocatoria, etc.

La epidemia del siglo XXI es la obesidad. Interesa saber qué proporción de un país tiene obesidad (*prevalencia* de la obesidad). Un estudio ejemplar llamado ENRICA trabajó con una muestra representativa de toda la población española de 18 y más años. En esa muestra se valoró la obesidad en 12.036 participantes (**20**). Se encontró una proporción de obesidad de 0,229 o, lo que es lo mismo, el 22,9%. Se utilizó una muestra y se desea extrapolar el resultado a toda la población española de esas edades, ya que el total de la población es inabarcable. Hay que *asumir* que la muestra es representativa de toda la población de 18 años o mayor. Es decir, estos 12.036 individuos no tienen ni más ni menos probabilidad de ser obesos que el resto de los habitantes de España de su misma edad. Para contestar a la pregunta acerca de la *representatividad,* debe valorarse el diseño del método de muestreo que se ha empleado, el porcentaje de no respondedores, los posibles sesgos de selección, etc., que son conceptos que pertenecen al método epidemiológico. Además, es interesante comprobar si coinciden las medias y proporciones de la muestra con las medias y proporciones conocidas de la población española para algunas variables (sexo, edad, nivel de estudios y otras).

Con los datos ya presentados, es fácil saber que en la muestra ($n$ = 12.036) se encontraron 2.756 obesos (22,9%). ¿Cómo se calcula este número?

$$12.036 \times 0,229 = 2.756,2$$

Se debe redondear al entero más próximo, ya que es lógico que el 22,9% se haya obtenido dividiendo el número de obesos (son personas, no admiten decimales) entre el total:

$$2.756 / 12.036 = 0,229$$

Este 22,9% es la estimación *puntual* hallada en la muestra *(estimador o proporción muestral),* pero se desea saber entre qué rango de valores podría encontrarse la verdadera proporción poblacional *(parámetro).* Con toda seguridad podrá decirse que no será exactamente del 22,9%. Habrá que proporcionar un rango de valores creíbles para el verdadero parámetro (horquilla de valores que incluya la prevalencia real de obesidad en la población española). ¿Qué anchura debe tener ese intervalo? ¿Podría valer del 1 al 99%? Si se diesen tales límites, se estaría seguro casi al 100% de que dentro de ellos estará incluida la verdadera proporción de obesos del país. Aunque decir que la proporción de obesos se encuentra entre el 1 y el 99% garantiza acertar, equivale a no decir nada. Tal intervalo sería poco informativo. Además, es poco asumible que en la población haya un 99% de personas con obesidad si en esta muestra solo hay un 22,9%. Lo mismo podría decirse respecto al 1%. Podría limitarse el rango un poco más, pero, a medida que se reduce el rango, se irá perdiendo seguridad y podría suceder que la proporción verdadera se situara fuera del intervalo (y se tendría *un mal día*). Los científicos suelen usar intervalos en los que tienen una confianza del 95% de incluir el parámetro.

El problema del intervalo de confianza se resuelve sumando y restando una cantidad a la proporción (0,229) calculada en la muestra. Una vez sumada y restada esta cantidad, podrá afirmarse, con una confianza del 95%, que la proporción de obesos españoles está entre un 22,1 y un 23,7% en la población de la que procede la muestra (personas ≥ 18 años). En una presentación se presentaría del modo siguiente:

Prevalencia de obesidad: 22,9% (intervalo de confianza al 95%: 22,1 a 23,7%)

En el apartado 4.5.2 se verán los cálculos. De momento interesa fijarse en que el intervalo es *simétrico,* hay una distancia del 0,8% hacia arriba y otro 0,8% hacia abajo. Este intervalo puede contener la verdadera proporción o tal vez no la contenga. Con los datos aportados no se sabe ni se está seguro, solo se *confía* en ello. ¿Con cuánta confianza? Con mucha: el 95% (confianza = 95%). *Confianza no es probabilidad.* Si se constatara que este intervalo sí contenía la proporción poblacional, su probabilidad de incluir el parámetro hubiese sido del 100%. Si, por el contrario, la verdadera proporción poblacional fuese, por ejemplo, del 22%, la probabilidad de que el intervalo

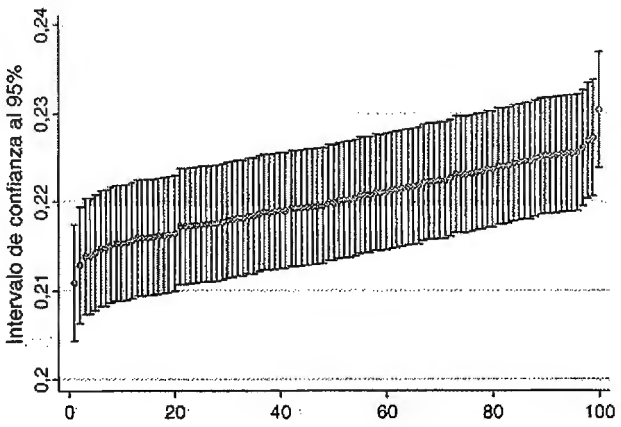



**Figura 4.4** Representación de 100 intervalos de confianza al 95% calculados en 100 muestras distintas obtenidas de la misma población. La verdadera proporción poblacional era 0,22. Las 100 muestras contenían, cada una, 12.000 sujetos. En cada una de las 100 muestras se calculó un intervalo de confianza al 95%.

incluyese el parámetro habría sido del 0%. Sin embargo el intervalo fue calculado con una confianza del 95%. Ningún intervalo contiene el parámetro con una probabilidad del 95%. Simplemente lo contiene (probabilidad = 100%) o no lo contiene (probabilidad = 0%), pero *subjetivamente se tiene una confianza* del 95% en que lo contendrá. Esta confianza está fundada en saber que, si se repitiese la experiencia 100 veces y en todas ellas se calculase un intervalo de confianza al 95%, habría 95 intervalos que contendrían el parámetro y 5 que no lo contendrían (**21**).

Esta experiencia se ha simulado en la figura 4.4: se obtuvieron 100 muestras aleatorias y en cada una se calculó un intervalo de confianza.

Aparecen 100 intervalos de confianza, definidos por un punto central (estimador muestral, proporción de cada muestra o estimación puntual) y unas barras laterales que delimitan la extensión del intervalo. Esta experiencia asumió el supuesto de que la verdadera proporción poblacional de obesidad en España era 0,22. Esto es totalmente ficticio, pues —insistimos— en la práctica se desconoce la verdad poblacional.

En el estudio real se calculó un intervalo al 95%, que valía de 0,221 a 0,237, con la esperanza de que el intervalo obtenido fuese uno de los 95 de cada 100 que contienen el parámetro. Con el supuesto ficticio que ahora se ha asumido, ese intervalo sería de los pocos (5 de cada 100, como se ve en la figura 4.4) que no contienen el verdadero parámetro. Si fuese verdad que la proporción poblacional ($\pi$ *(pi)* en letras griegas) fue 0,22, los investigadores del estudio ENRICA habrían tenido *un mal día* y esa muestra concreta que ellos obtuvieron, no apoyaría —debido al azar— la verdad del parámetro poblacional. Aun así, esto no suele ocurrir casi nunca, solo 5 de cada 100 veces, como se ve en la figura 4.4, y lo razonable siempre es *confiar* en que el intervalo encontrado sí contendrá el verdadero parámetro y aceptar el intervalo del ENRICA.

### 4.5.1. Supuestos que se deben verificar para calcular el intervalo de confianza de una proporción

Con muestras grandes, la estimación de una proporción se suele hacer por aproximación a la normal. Cuanto más cercana a $p = 0,5$ (porcentaje = 50%) sea la proporción, mejor será la aproximación. La aproximación a la normal es razonable siempre que $p$ y $1 - p$ sean superiores a

$5/n$ (15). Si no se cumplen las condiciones de aplicación por tratarse de una muestra pequeña o cuando las proporciones son próximas al 0 o al 100%, el problema debe ser resuelto a través del método exacto por la distribución binomial mediante el uso, preferentemente, de algún *software* estadístico (STATA, SPSS...), como se verá más adelante.

### 4.5.2. Cálculo del intervalo de confianza de una proporción (usando la distribución normal)

1. Cálculo de la proporción:

$$p = c/n$$

donde p es la proporción muestral; $c$ es el número de sujetos con el carácter que se estudia y $n$ es el total de la muestra. En el ejemplo:

$$p = 2756/12036 = 0,229$$

2. Comprobación de las condiciones de aplicación:

$$p > 5/n \rightarrow 0,229 > 5/12.036$$

$$(1-p) > 5/n \rightarrow (1-0,229) > 5/12.036$$

3. Cálculo del error estándar de la proporción (EEP):

$$\text{EEP} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}}$$

Con mucha frecuencia, en los textos de estadística a $(1 - p)$ se le llama $q$, como aparece en la segunda forma de la expresión.

$$\text{EEP} = \sqrt{\frac{0,229 \times (1-0,229)}{12.036}} = 0,00383$$

4. Búsqueda en la distribución normal (tablas, Excel o STATA) de un valor $z$ adecuado para el error alfa del intervalo de confianza que se desee calcular.

Alfa *(α)* (en tanto por uno) es lo que le falta al intervalo de confianza para llegar al 100%. Por ejemplo, si el intervalo fuese al 95%, el valor de alfa total sería 0,05, sumando las dos colas ($\alpha/2 = 0,025$ en cada cola). La distribución normal dice que, si se desea dejar en cada cola una probabilidad del 2,5% ($\alpha/2 = 0,025$), entonces $z_{\alpha/2} = 1,96$. Hay que tener siempre en cuenta que:

a. El *error* $\alpha$ en cada cola es (1 – nivel de confianza)/2.
b. El *error* $\alpha$ en ambas colas es (1 – nivel de confianza).

5. Sumar y restar $z$ veces el EEP a la proporción hallada.

$$\text{IC}(\pi) = p \pm z_{\alpha/2}\ \text{EEP}$$

$$\text{IC proporción poblacional}\ (\pi) = 0,229 \pm 1,96(0,00383) = 0,221 \text{ a } 0,237$$

### 4.5.3. Intervalos de confianza exactos de una proporción (binomial) con STATA

Cuando la aproximación anterior no es razonable porque no se cumplen las condiciones de aplicación, se debe usar un método exacto basado en la distribución binomial, que sería laborioso ejecutar *a mano*. Es mejor recurrir a STATA. Se aplicó este procedimiento exacto con STATA sin

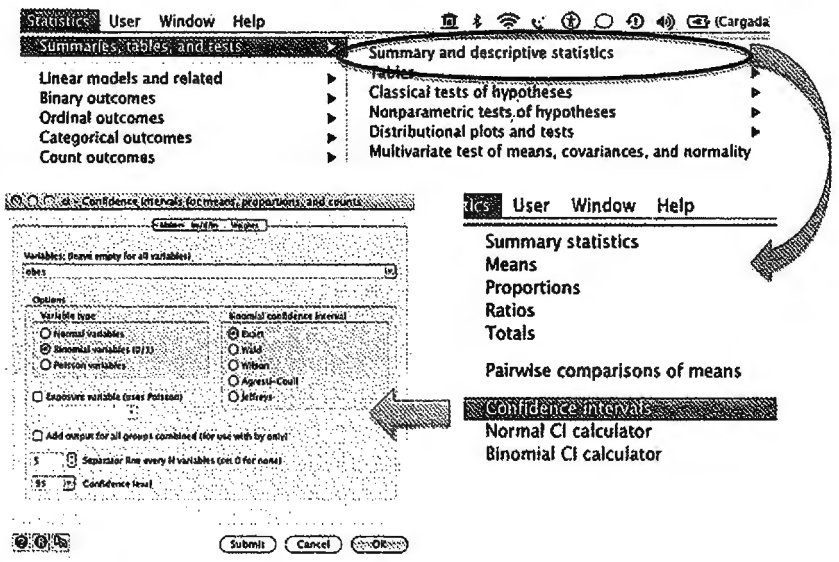
necesidad de introducir los datos, uno a uno (no hubo que escribir 12.036 filas); simplemente, tras la orden **cii** se escribe el total muestral (12036) y los que tenían obesidad (2756).

```
. cii 12036 2756
```

| Variable | Obs | Mean | Std. Err. | — Binomial Exact — [95% Conf. | Interval] |
|---|---|---|---|---|---|
| | 12036 | .2289797 | .0038299 | .2214979 | .2365928 |

***Interpretación:*** *con una confianza del 95%, puede decirse que la proporción poblacional se encuentra en el intervalo (0,2215 a 0,2366), es decir, se confía en que la verdadera proporción de obesidad esté entre el 22,15% y el 23,66% de la población española de 18 o más años. Al tratarse de una muestra grande, sus resultados coinciden con la aproximación a la normal antes explicada.*

Si la base de datos estuviese ya introducida con un sujeto en cada fila (se habrían escrito 12.036 filas) y una columna de ceros y de unos (0 = no obeso; 1 = obeso) llamada ***obes,*** se podría recurrir al menú para seleccionar *Statistics,* después *Summary, tables and tests,* en tercer lugar *Summary and descriptive statistics* y, por último, *Confidence intervals.*




Al pulsar en *Confidence intervals* aparece otro menú donde debe escribirse el nombre de la variable ***(obes).*** Si una vez seleccionada la opción *Binomial confidence interval → Exact* se oprime el botón *Submit,* se encontrará la misma solución que antes:

```
. ci obes, binomial
```

| Variable | Obs | Mean | Std. Err. | — Binomial Exact — [95% Conf. | Interval] |
|---|---|---|---|---|---|
| obes | 12036 | .2289797 | .0038299 | .2214979 | .2365928 |

Al usar la distribución binomial con muestras más pequeñas, el intervalo que resulta puede ser *asimétrico,* algo que no ocurre nunca al usar la normal. Puede verse en otro ejemplo distinto.

Imagínese que hubo dos sujetos con una mutación en una muestra de 18 pacientes. STATA dará un intervalo de confianza al 95% asimétrico para la proporción (2/18 = 0,11) que va desde 0,014 a 0,347.

```
. cii 18 2
```

| Variable | Obs | Mean | Std. Err. | — Binomial Exact — [95% Conf. | Interval] |
|---|---|---|---|---|---|
| | 18 | .1111111 | .0740741 | .0137512 | .3471204 |

La distancia entre la estimación puntual (0,11) y el límite superior (0,347) es mayor que la que existe entre la proporción (0,11) y el límite inferior (0,014). Esto puede suceder con el método binomial exacto.

En este segundo ejemplo, la aproximación con la distribución normal no sería válida, ya que *no* se cumple que $p > 5/n$, pues $0{,}111 < 5/18$. Usar la normal aquí daría problemas; por ejemplo, estimaría un límite inferior de confianza negativo, lo cual es absurdo.

Puede entenderse intuitivamente el uso de la distribución binomial imaginando que consiste en que el ordenador busca, a través de *rondas* o recorridos reiterativos, aquellos valores cuya probabilidad binomial acumulada sería de 0,025 en cada cola (es decir, dejan fuera del intervalo una probabilidad global del 5% y dentro el 95% restante). Si este procedimiento se hiciera a mano, el trabajo necesario sería enorme y no compensaría. Por ello, es preciso recurrir al ordenador.

Por omisión, si no se especifica nada más, STATA usará el 95% de confianza. Si se desea otro nivel de confianza, por ejemplo del 90%, se puede añadir la opción **level** ***(confianza):***

```
. cii 18 2, level(90)
```

| Variable | Obs | Mean | Std. Err. | — Binomial Exact — [90% Conf. | Interval] |
|---|---|---|---|---|---|
| | 18 | .1111111 | .0740741 | .0201107 | .3102627 |

***Interpretación:*** *con una confianza del 90% puede decirse que la proporción poblacional está en el intervalo (0,020 a 0,310), es decir, se confía en que entre el 2,0% y el 31,0% de la población de esos enfermos tendrá esa mutación.*

### 4.5.4. Intervalos de confianza exactos de una proporción (binomial) con R

Se utiliza la función **binom.test,** indicando, en primer lugar, el número de eventos y, en segundo lugar, el número total de observaciones. Se obtendrá la probabilidad de la proporción frente a una probabilidad esperada de 0,5, el intervalo de confianza al 95% y la proporción.

```
binom.test(2756,12036)
```

```
data:  2756 and 12036
number of successes = 2756, number of trials = 12036, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2214979 0.2365928
sample estimates:
probability of success
             0.2289797
```

Para un intervalo de confianza del 90%, se indicaría:

```
binom.test(27 56,12036,conf .level = 0.90)
```

## 4.6. ESTIMACIÓN DE UNA MEDIA

Se publica que el valor medio del índice de masa corporal (IMC) en varones españoles de 16 a 26 años fue de 23,6 kg/m² y su desviación estándar fue de 3,2 kg/m² (**22**). Supóngase que había 1.024 varones de estas edades en la muestra. Se desea estimar la media poblacional *(μ)* del IMC. Los datos serían:

$\bar{x} = 23$

$s = 3,2$

$n = 1.024$

A partir de la media muestral ($\bar{x}$), cuyo valor es 23,6 kg/m², se calcula el intervalo de confianza para la media poblacional *(μ)*.

### 4.6.1. Muestras muy grandes: intervalo de confianza de una media usando la distribución normal

El error estándar de la media (EEM) se calcula como:

$$\text{EEM} = \frac{s}{\sqrt{n}} = \frac{3,2}{\sqrt{1.024}} = 0,1$$

Para m*uestras muy grandes* ($n > 500$) como esta, una vez calculado el error estándar de la media (EEM), es preciso multiplicarlo por la cantidad $z$ tomada de la normal, y después sumarlo y restarlo a la media muestral. Si se desea un intervalo de confianza al 95%, el valor de $z_{\alpha/2} = z_{0,025}$ será 1,96.

$$\text{IC}(1-\alpha) = \mu \in \bar{x} \pm z_{\alpha/2}\ \text{EEM} = \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\text{IC } 95\% = 23,6 \pm \left(1,96 \times \frac{3,2}{\sqrt{1.024}}\right) = 23,6 \pm (1,96 \times 0,1) = 23,4 - 23,8$$

Se tiene una confianza del 95% de que dentro del intervalo 23,4 a 23,8 kg/m² esté contenida la verdadera media del IMC de la población de varones de esas edades.

### 4.6.2. Muestras pequeñas: intervalo de confianza de una media con la distribución T

No obstante, hay que aclarar que el intervalo de confianza así calculado no sería válido si la muestra fuese muy pequeña (especialmente en muestras con $n < 60$). Si la muestra fuese pequeña y solo se dispone de la desviación estándar muestral, como suele suceder, es imprescindible sustituir el valor de $z$ ($z_{0,025} = 1,96$ para un intervalo de confianza al 95%) por otro mayor que sigue otra distribución llamada T o, a veces, *t de Student*, en vez de seguir la distribución normal. Cuando la muestra es pequeña, la utilización de $n$ (muestral) en vez de σ (sigma o desviación estándar poblacional) para calcular el error estándar supone una inexactitud. Cuanto menor sea la muestra, mayor será la inexactitud. Con muestras inferiores a 120 individuos, el error al usar $z$ es <1%; con muestras de entre 120 y 60 individuos, el error es <2%. En muestras cada vez más pequeñas ($n < 60$), el error va siendo progresivamente mayor.

La distribución *T* es una nueva distribución teórica de probabilidad cuyos valores (que se llaman *t*) se interpretan del mismo modo que los valores $z$ de la distribución normal. La peculiaridad de la distribución *T* es que, para cada error $\alpha$, proporciona un valor de $t$ que es distinto *para cada tamaño de muestra.* En cambio, la distribución normal da siempre el mismo valor $z$ para cada error $\alpha$, sea cual sea el tamaño muestral. Cuando el tamaño de muestra es muy grande, resulta indiferente usar una u otra, ya que entonces se cumple que $t \approx z$.

Esta distribución fue descrita por W. S. Gosset a principios del siglo xx usando como seudónimo «estudiante» *(Student);* este nombre ha perdurado. Al utilizar la *t* para calcular intervalos de confianza para una media, basta saber que los grados de libertad son $n - 1$ (gl = $n - 1$), siendo $n$ el tamaño de la muestra.

Supóngase que en 64 pacientes de alto riesgo coronario, la media de una puntuación global (de 0 a 14) en conformidad con la dieta mediterránea era 8,5 puntos (desviación estándar = 2,0). Se pretende calcular el intervalo de confianza al 95% de la media.

Los datos son:

$\bar{x} = 8,5$

$s = 2,0$

$n = 64$

Se calcula primero el error estándar de la media (EEM):

$$\text{EEM} = \frac{s}{\sqrt{n}} = \frac{2,0}{\sqrt{64}} = 0,25$$

Después, solo hay que restar y sumar ese error estándar multiplicado por la cantidad *t* a la media muestral. Es decir, se aplica la siguiente expresión:

$$\text{IC}(1-\alpha)\text{para } \mu = \bar{x} \pm t_{\alpha/2,\, n-1}\text{EEM}$$

$$\text{IC}(1-\alpha)\text{para } \mu = 8,5 \pm 2(0,25) = 8 \text{ a } 9$$

donde $\mu$ es la media poblacional que se estima, $\bar{x}$ es la media que se ha calculado en la muestra (8,5), EEM es el error estándar de la media (0,25) y $t_{\alpha/2,n-1}$ el valor de la distribución *t* de Student para un error alfa que sea el complementario de la confianza que se desee para el intervalo y unos grados de libertad $n - 1$. Este valor se puede consultar en unas tablas, en Excel o STATA. En este ejemplo, los grados de libertad serán 63 ($n - 1$), ya que había 64 pacientes. Si se mira en las tablas, en Excel o STATA, el valor de *t* es 1,998 para un error alfa de 0,025 en cada cola ($\alpha/2 = 0,025$ y $\alpha = 0,05$) y 63 grados de libertad (gl = 64 – 1).

$$t_{\alpha/2,n-1} = t_{0,025,\,63} = 1,998 \approx 2$$

$$\text{IC } 95\% = \mu \in 8,5 \pm 1,998 \times 0,25 = 8 \text{ a } 9$$

La anchura o amplitud del intervalo de confianza es 1 en este ejemplo, pues va desde 8,0 a 9,0 puntos. En la práctica, en muchas ocasiones en que está indicado su uso, la t de Student para un intervalo de confianza al 95% tendrá un valor muy próximo a 2. Entonces, como en este ejemplo, la amplitud o ancho total del intervalo de confianza multiplicará casi exactamente por 4 el error estándar de la media (0,25 × 4 = 1 = distancia entre 8,0 y 9,0).

Siempre se puede usar la distribución *t*, pero cuando la muestra es muy grande ($n > 100$), utilizar la *z* o la *t* es prácticamente equivalente. En la figura 4.5 se representa una distribución *t* de Student con seis grados de libertad en trazo grueso y la distribución normal en trazo delgado.

La situación de seis grados de libertad corresponde a un tamaño muestral ridículamente pequeño, solo de siete sujetos. Incluso con ese bajo tamaño existe mucha similitud con la distribución normal. Lo destacable es que la diferencia fundamental reside en las colas, que es lo que se utiliza en el cálculo de los intervalos de confianza. Se suele decir que la *t* tiene colas *más pesadas* que la normal. Así, para un mismo error $\alpha$, los valores de *t* siempre serán superiores a *z*, lo cual produce un pequeño ensanchamiento del intervalo de confianza cuando se usa *t* en vez de *z*. Esto resulta coherente con considerar que no solo la *media muestral* es un estimador, sino también la *desviación*

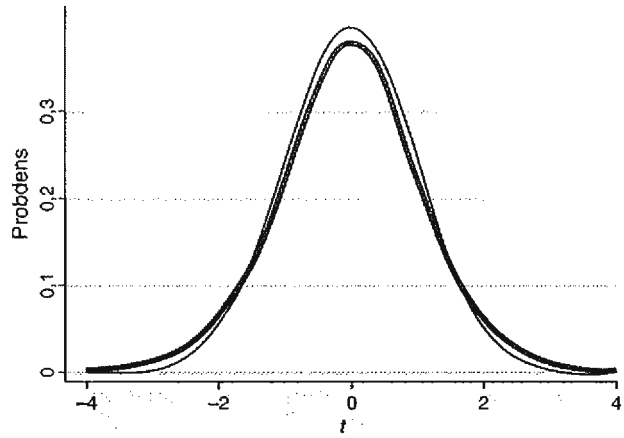



**Figura 4.5** Distribución *t* de Student con seis grados de libertad *(trazo grueso)* y distribución normal *(trazo delgado).*

*estándar* usada para calcular el error de estimación de la media; cuanto menor sea la muestra, en más incertidumbre se incurre al usar la desviación estándar muestral en vez de la poblacional.

### 4.6.3. Supuestos que se deben verificar al calcular el intervalo de confianza a una media

- Si la muestra es pequeña ($n < 30$), debe cumplirse el requisito de normalidad.
- Si la muestra es grande ($n \geq 30$), *no hace falta asumir normalidad.*
- Desde un punto de vista práctico, siempre se puede usar *t*, pero cuando la muestra es muy grande ($n > 100$), utilizar *z* es prácticamente equivalente a usar *t*.

Se pueden aplicar niveles de confianza del 95%, pero también del 90 o del 99%. Un intervalo de confianza del 99% es más creíble que uno del 90%. Sin embargo, el intervalo de confianza al 99% será más ancho que el del 90%. En la figura 4.6 se representan los intervalos de confianza para este ejemplo (media muestral = 8,5, $n = 2$ y $n = 64$) con distintos grados de confianza.

A medida que aumenta la confianza, el intervalo es más ancho e impreciso. Este es el precio que se paga porque sea más creíble. Los intervalos de confianza de la media son simétricos. Por eso, el estimador muestral (media calculada en la muestra o estimación puntual) debe encontrarse siempre a mitad de distancia de los límites.

## 4.7. INTERVALOS DE CONFIANZA CON STATA

La orden para STATA es simple: **ci** o bien **cii.** Si ya están metidos los datos en la base de datos, se usará **ci** y después el nombre de la variable, por ejemplo:

```
. ci DIETA
```

| Variable | Obs | Mean | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|---|
| DIETA | 64 | 8.5 | .25 | 8.000415 | 8.999585 |

Por omisión, devuelve el intervalo de confianza al 95%. Para otra confianza, se usa la opción **level** ***(confianza):***

```
ci DIETA, level(90)
```

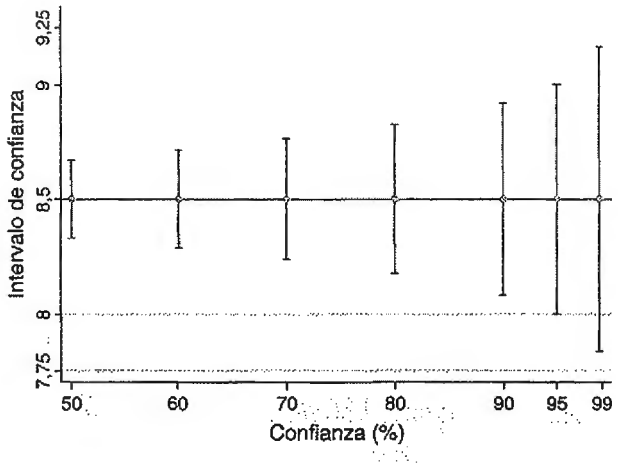



**Figura 4.6** Intervalos de confianza al 50, 60, 70, 80, 90, 95 y 99% para la media (media muestral = 8,5 puntos) de una muestra de 64 sujetos con desviación estándar = 2 puntos. Al aumentar la confianza, también lo hace la amplitud del intervalo.

Si se desea obtener directamente el intervalo de confianza a partir de los estimadores muestrales sin estar usando la base de datos, se dará la orden **cii** seguida de ***n***, media y desviación estándar:

```
. cii 64 8.5 2
```

| Variable | Obs | Mean | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|---|
| | 64 | 8.5 | .25 | 8.000415 | 8.999585 |

Para más detalles, se puede consultar el vídeo titulado: *STATA_3_1: Intervalos confianza* proporción y media en: http://www.unav.es/departamento/preventiva/recursos_bioestadistica

## 4.8. INTERVALOS DE CONFIANZA CON OTROS PROGRAMAS

### 4.8.1. Intervalos de confianza con R/Splus

R/Splus calculará intervalos de confianza para una media si se programa. Por ejemplo, se pueden construir tres funciones, como muestra la tabla 4.3.

Una vez construidas estas tres funciones, basta con escribir el nombre programado, m_lic *(variable)* y m_lsc *(variable)*, para obtener el intervalo:

***Tabla 4.3 Funciones programables para obtener un intervalo de confianza de una media con R/Splus***

| FUNCIÓN | SINTAXIS |
|---|---|
| Error estándar de la media | `eem <- function(x){(var(x)/length(x))^0.5}` |
| Límite superior de confianza (95%) | `m_lsc <- function(x){mean(x) + (qt(0.975, length(x)-1))*eem(x)}` |
| Límite inferior de confianza (95%) | `m_lic <- function(x){mean(x) - (qt(0.975, length(x)-1))*eem(x)}` |

*Téngase en cuenta que qt (0,975,63) proporciona el valor de una t de Student con un error α de 0,025 en cada cola y 63 grados de libertad.*

```
> table(DIETA)
DIETA
 5  6  7  8  9 10 11 12 13 14
 3  5 16  8 14 11  2  2  1  2
> mean(DIETA)
[1] 8.5
> eem <- function(x){(var(x)/length(x))^0.5}
> m_lic <- function(x){mean(x)-(qt(0.975, length(x)-1))*eem(x)}
> m_lsc <- function(x){mean(x)+(qt(0.975, length(x)-1))*eem(x)}
> m_lic(DIETA)
[1] 8.000415
> m_lsc(DIETA)
[1] 8.999585
```

De manera análoga puede programarse el intervalo de confianza para una proporción, introduciendo previamente la fórmula del error estándar de una proporción.

### 4.8.2. Intervalos de confianza de una media con SPSS

En SPSS, desde el menú *Analizar,* se selecciona *Estadísticos descriptivos* y, dentro de ellos, se elige *Explorar.* Aparecerá otro menú en el que se debe situar la variable en la ventana superior *(Lista de dependientes).*

Con sintaxis se pide así:

```
EXAM VAR=dieta

/PLOT NONE

/CINTERVAL 95.
```

La última línea es eliminable si el intervalo es para el 95%. Se puede usar esa línea para modificar la confianza. La salida programada de SPSS ofrece también otros resultados.

**Descriptivos**

| | | | Estadístico | Error típ. |
|---|---|---|---|---|
| DIETA | Media | | 8,5000 | ,25000 |
| | Intervalo de confianza para la media al 95% | Límite inferior | 8,0004 | |
| | | Límite superior | 8,9996 | |
| | Media recortada al 5% | | 8,4097 | |
| | Mediana | | 8,5000 | |
| | Varianza | | 4,000 | |
| | Desv. típ. | | 2,00000 | |
| | Mínimo | | 5,00 | |
| | Máximo | | 14,00 | |
| | Rango | | 9,00 | |
| | Amplitud intercuartil | | 3,00 | |
| | Asimetría | | ,651 | ,299 |
| | Curtosis | | ,564 | ,590 |

***Interpretación:*** *con una confianza del 95% puede decirse que la media poblacional se encuentra en el intervalo encontrado (8,0 a 9,0 puntos), es decir, se confía en que la puntuación media en la población de la que se extrajo esta muestra esté entre 8,0 y 9,0 puntos. Pueden localizarse los límites de confianza (8,0004 y 8,9996) en las filas 2-3. La salida de SPSS presenta también la estimación puntual de la media muestral (8,5), el error estándar de la media (0,25) y otros muchos índices descriptivos.*

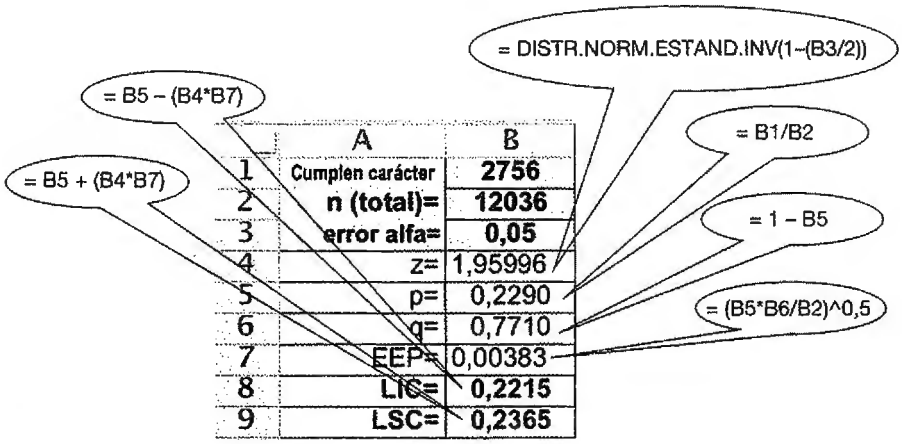



**Figura 4.7** Programación de una hoja de Excel para calcular intervalos de confianza de una proporción.

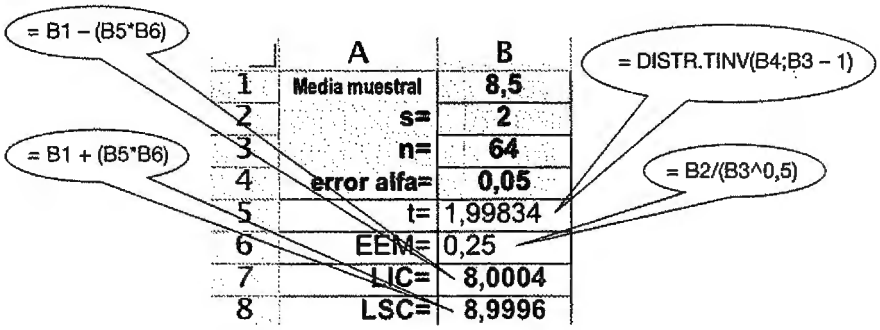



**Figura 4.8** Programación de una hoja de Excel para calcular intervalos de confianza de una media.

### 4.8.3. Intervalos de confianza con Excel

En Excel se puede preparar fácilmente una hoja de cálculo que estime automáticamente intervalos de confianza para una proporción o para una media, siguiendo los pasos que se muestran en las figuras 4.7 y 4.8.

Se ha usado la columna A para introducir los nombres de los elementos necesarios para calcular el intervalo de confianza y la columna B para introducir sus valores. A partir de la casilla B4 en el caso de proporciones y de la B5 en la media, ya no se escriben números, sino que se introduce la fórmula[2] para calcular la $z$ o la $t$ de Student, con las funciones =DISTR.NORM.ESTAND.INV (1 – ($\alpha$/2)) y =DISTR.TINV ($\alpha$; grados de libertad).

En las casillas B7 (v. fig. 4.7) y B6 (v. fig. 4.8) se han introducido las fórmulas del error estándar de la proporción (EEP) y de la media (EEM), respectivamente, para que Excel los calcule automáticamente. Debajo se introducen las fórmulas necesarias para que se calculen los límites de confianza inferior (LIC) y superior (LSC). Así, cuando se cambien los datos de B1 a B3 (v. fig. 4.7) o de B1 a B4 (v. fig. 4.8), Excel recalculará automáticamente todo lo que queda debajo. Cambiando la casilla de $\alpha$ se puede lograr otra confianza; por ejemplo, si se pone $\alpha$ = 0,10, el intervalo tendrá una confianza del 90%. Se puede usar la ayuda para manejar la fórmula de la normal o la $t$.

2 En versiones más antiguas de Excel, la fórmula es DISTR.T.INV en vez de DISTR.TINV.

En el cálculo del intervalo de confianza para la proporción, puede resultar extraño tener que usar uno menos la mitad del error alfa (1 – (B3/2)) para obtener la *z* de la distribución normal. Como se vio en el capítulo anterior, Excel siempre calcula la cola *de la izquierda*. Por eso, si se introdujese 0,05 directamente, Excel produciría una $z = -1{,}645$ en vez de $z = +1{,}96$, que es la que se busca. Para cambiar el signo se procede a restar uno; para repartir el error $\alpha$ en dos colas se hace la división entre dos. No debe olvidarse que, en un intervalo de confianza, el error $\alpha$ (0,05 en el ejemplo) se reparte por mitades entre las dos colas.

Una vez guardadas estas expresiones en la hoja de cálculo, basta con cambiar cualquier valor de entrada para que Excel automáticamente recalcule el resto y proporcione el intervalo de confianza con las nuevas especificaciones. En este sentido, puede decirse que con estas sencillas operaciones se dispondrá de una calculadora que *amigablemente* calcula intervalos de confianza.

## 4.9. LA DISTRIBUCIÓN *T* DE STUDENT EN EXCEL, STATA Y R/SPLUS

### 4.9.1. La distribución *t* en Excel

Al final del libro, como en todo texto de estadística, hay unas tablas de la *t* de Student. No obstante, visto lo anterior, se advertirá de que es más interesante aprovechar las funciones de Excel para sustituir con ventajas a las tradicionales tablas.

Basta con crear tres columnas, una para grados de libertad, otra para error alfa y una tercera en la que se introduce la función =DISTR.TINV(α;grados de libertad).

Como aparece en la figura 4.9, bastará con indicar la siguiente fórmula en C2:
=DISTR.TINV(B2;A2)

Con la sencillez que muestra la figura 4.9, se han construido unas tablas de la *t* que superan en prontitud, exactitud y flexibilidad a la tabla clásica. Solo cambiando los valores de las columnas A y B, Excel dará la respuesta deseada en la columna C.

También puede hacerse la pregunta al revés. Si se encuentra un valor concreto de *t* con unos grados de libertad determinados, ¿cuánto valdrá la probabilidad de la cola? Ahora ya no se trata de error alfa, sino de valor *p*, ya que es algo que se ha encontrado *a posteriori*, como se verá en el apartado 4.12 al hablar de contraste de hipótesis. Se obtendrá la probabilidad (área en las colas) a partir de *t*, en vez de obtener *t* a partir de la probabilidad. Esto se resuelve con otra función[3] que Excel denomina =DISTRT(t;gl;colas).

Se pueden crear cuatro columnas, una para grados de libertad, otra para la *t* encontrada, una tercera para la función =DISTRT(B#;A#;1) que obtiene la probabilidad a una cola, y una cuarta para escribir =DISTRT(B#;A#;2) para la probabilidad a dos colas. En la figura 4.10 se ha supuesto un ejemplo en el que el valor *t* resultante es siempre 2, pero cambian los grados de libertad.

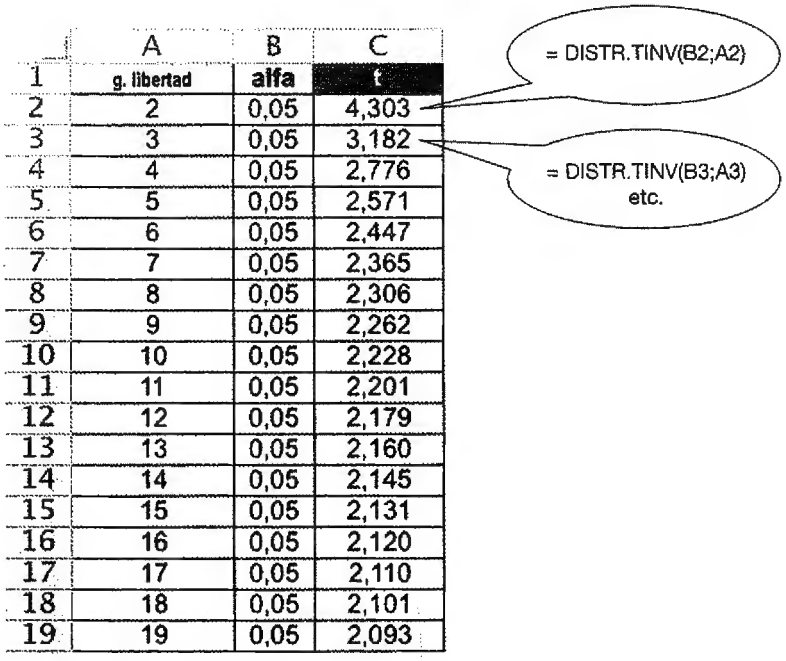
Bastará cambiar 2 por el valor real que se encuentre en el experimento y el programa devolverá automáticamente la probabilidad *(p)* a una o dos colas en las columnas C y D. También se recalcula si se modifican los grados de libertad.

### 4.9.2. La distribución *t* en STATA

Se pueden obtener rápidamente los valores de la *t* en STATA con estas órdenes:

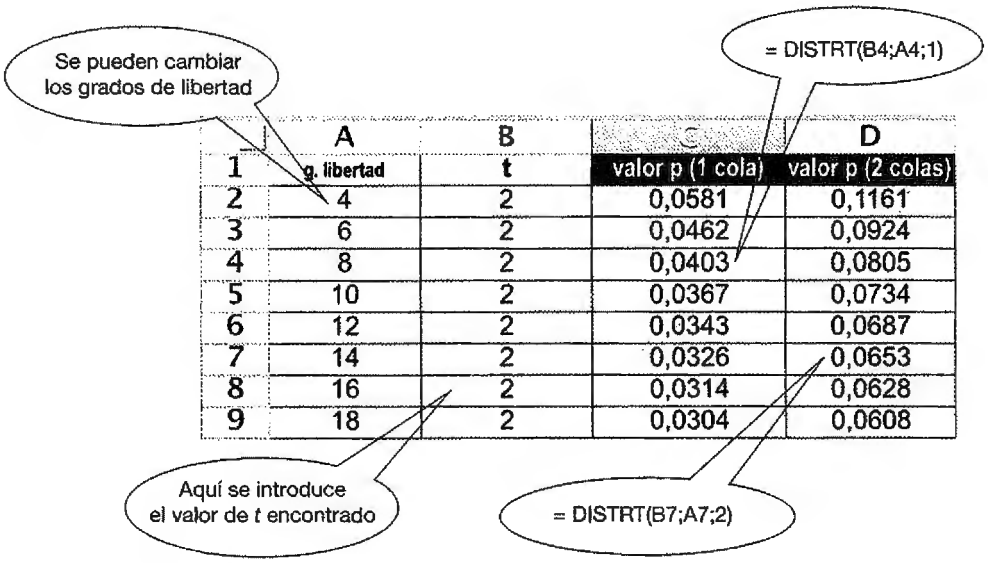
- **invttail*(gl,alfa/2)*:** devuelve el valor de *t* cuando se le da la probabilidad (1 cola).
- **2*ttail*(gl,t)*:** devuelve el valor de la probabilidad (2 colas) cuando se le da *t*.

3 En versiones más antiguas de Excel la fórmula es DISTR.T en vez de DISTRT.




| | A | B | C |
|---|---|---|---|
| 1 | g. libertad | alfa | t |
| 2 | 2 | 0,05 | 4,303 |
| 3 | 3 | 0,05 | 3,182 |
| 4 | 4 | 0,05 | 2,776 |
| 5 | 5 | 0,05 | 2,571 |
| 6 | 6 | 0,05 | 2,447 |
| 7 | 7 | 0,05 | 2,365 |
| 8 | 8 | 0,05 | 2,306 |
| 9 | 9 | 0,05 | 2,262 |
| 10 | 10 | 0,05 | 2,228 |
| 11 | 11 | 0,05 | 2,201 |
| 12 | 12 | 0,05 | 2,179 |
| 13 | 13 | 0,05 | 2,160 |
| 14 | 14 | 0,05 | 2,145 |
| 15 | 15 | 0,05 | 2,131 |
| 16 | 16 | 0,05 | 2,120 |
| 17 | 17 | 0,05 | 2,110 |
| 18 | 18 | 0,05 | 2,101 |
| 19 | 19 | 0,05 | 2,093 |

**Figura 4.9** Programación de tablas de la *t* de Student con Excel. Se proporcionan a Excel los grados de libertad y la probabilidad (ambas colas sumadas), y Excel devolverá el valor de *t*. Esta es la *t* que se debe usar en los intervalos de confianza. El ejemplo de la figura corresponde a intervalos al 95%.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | g. libertad | t | valor p (1 cola) | valor p (2 colas) |
| 2 | 4 | 2 | 0,0581 | 0,1161 |
| 3 | 6 | 2 | 0,0462 | 0,0924 |
| 4 | 8 | 2 | 0,0403 | 0,0805 |
| 5 | 10 | 2 | 0,0367 | 0,0734 |
| 6 | 12 | 2 | 0,0343 | 0,0687 |
| 7 | 14 | 2 | 0,0326 | 0,0653 |
| 8 | 16 | 2 | 0,0314 | 0,0628 |
| 9 | 18 | 2 | 0,0304 | 0,0608 |

**Figura 4.10** Tablas de la *t* de Student con Excel. Se proporcionan a Excel los grados de libertad y el valor de *t*. Entonces, Excel devolverá la probabilidad a una o a dos colas. Aquí es donde debe consultarse la *t* que se obtiene en contrastes de hipótesis.

Ambas requieren ir precedidas de la instrucción **display.**
Ejemplos:

```
. di invttail (63, 0. 025)
1. 9983405

.di 2*ttail (63, 1. 9983405)
. 05
```

### 4.9.3. La distribución t en R/Splus

Para obtener con R/Splus los mismos resultados anteriores se escribirá:

```
> qt(0.025,63,lower.tail=F)
[1] 1.998341
> pt(1.998341,63)
[1] 0.975
> 2*(1-pt(1.998341,63))
[1] 0.04999995
```

La última expresión es la que debe usarse para obtener un valor de probabilidad a dos colas. Para la primera expresión se puede usar también **qt***(0.975,63)*,que equivale a **qt***(0,025,63,***lower. tail = F***)*. Para la segunda expresión se puede utilizar el signo menos **pt***(–1.998341,63)* y se obtendrá $p$ = 0,025 (a una cola).

## 4.10. ESTIMACIÓN DE UNA MEDIANA

Imagínese que se quiere conocer cuál es la mediana de los cigarrillos fumados al día en una base de datos de 25 fumadores que ya se usó en el capítulo 2. Los datos son:

| Cig./día | Frec. | *Acum.* |
|---|---|---|
| 7 | 1 | *1* |
| 8 | 1 | *2* |
| 10 | 2 | *4* |
| 11 | 1 | *5* |
| 15 | 3 | *8* |
| 20 | 11 | *19* |
| 30 | 4 | *23* |
| 40 | 1 | *24* |
| 60 | 1 | *25* |

La tabla 4.4 proporciona los dos números de orden (puestos) cuyos valores serán los límites de confianza al 95% de la mediana. Es decir, el número de orden 5 sería el 5.° valor más bajo, ya que ocupa el puesto 5.° cuando se ordenan todos los valores de menor a mayor. A estos puestos a veces se les llama *rangos.*

Esta tabla sirve para tamaños $n < 30$. Como en el ejemplo hay 25 observaciones, la mediana estará en el valor que ocupe el puesto 13.°, es decir, en 20 cigarrillos/día. Su intervalo de confianza al 95% se mira en la tabla 4.4 y se obtienen los valores situados en los lugares 8.° y 18.°. Es decir, puede afirmarse con una confianza del 95% que la mediana poblacional estará entre 15 y 20 cigarrillos/día. Se expresaría así:

*Mediana = 20 (intervalo de confianza al 95% : 15 – 20).*

*Tabla 4.4 Intervalos de confianza al 95% para una mediana*

| TAMAÑO DE LA MUESTRA | RANGO (PUESTO) DEL LÍMITE INFERIOR DE CONFIANZA AL 95% | RANGO (PUESTO) DEL LÍMITE SUPERIOR DE CONFIANZA AL 95% |
|---|---|---|
| 6 | 1 | 6 |
| 7 | 1 | 7 |
| 8 | 1 | 8 |
| 9 | 2 | 8 |
| 10 | 2 | 9 |
| 11 | 2 | 10 |
| 12 | 3 | 10 |
| 13 | 3 | 11 |
| 14 | 3 | 12 |
| 15 | 4 | 12 |
| 16 | 4 | 13 |
| 17 | 5 | 13 |
| 18 | 5 | 14 |
| 19 | 5 | 15 |
| 20 | 6 | 15 |
| 21 | 6 | 16 |
| 22 | 6 | 17 |
| 23 | 7 | 17 |
| 24 | 7 | 18 |
| 25 | 8 | 18 |
| 26 | 8 | 19 |
| 27 | 8 | 20 |
| 28 | 9 | 20 |
| 29 | 9 | 21 |

Este procedimiento no requiere ninguna suposición ni asumir ninguna condición de aplicación, pero puede proporcionar intervalos tan asimétricos como el de este ejemplo, donde la mediana y el límite superior de su intervalo coinciden. Esto contrasta con la frecuente simetría de intervalos para proporciones y medias. Además, el intervalo de confianza para la mediana puede ser muy amplio y llegar a duplicar la anchura del intervalo de confianza de la media.

En la práctica casi nunca se calculan o se presentan intervalos de confianza para la mediana, aunque debería hacerse más. Especialmente, se debe preferir el intervalo de confianza de la mediana, y no de la media, para variables con distribución asimétrica, que presentan datos truncados o que no siguen una distribución normal (**15**). Se presenta esta situación, por ejemplo, cuando existen valores atípicos *(outliers)*.

Para las muestras de mayor tamaño ($n \geq 30$), se calcula el intervalo de confianza de los rangos para la mediana según la siguiente expresión, que se ha aplicado al ejemplo:

$$\text{IC}\left(\text{n.}^\circ \text{ orden}\right) = \text{n.}^\circ \text{ orden}_{\text{mediana}} \pm t_{\frac{\alpha}{2},n-1} \frac{\sqrt{n}}{2}$$

$$\text{IC}\left(\text{n.}^\circ \text{ orden}\right) = 13.^\circ \pm t_{0,025;24} \frac{\sqrt{25}}{2}$$

$$\text{IC}\left(\text{n.}^\circ \text{ orden}\right) = 13.^\circ \pm (2,064 \times 2,5) = 7,8.^\circ \text{ a } 18,2.^\circ$$

Simplemente se ha sumado y restado una cantidad al puesto que ocupaba la mediana. Tal cantidad vale el producto de $t$ por el error estándar de la mediana (en unidades de número de orden o rango).

Al hacerlo a mano no importa redondear estos dos límites (7,8.° y 18,2.°) a los enteros más próximos (8.° y 18.°), y se elegirán los mismos límites que antes, el valor 8.° y el 18.° de la secuencia ordenada de datos: IC 95% (mediana) = 15 a 20.

La orden **centile** de STATA proporciona el intervalo de confianza de la mediana (o de cualquier percentil). Tiene en cuenta los decimales y hace interpolaciones. Además, usa otro procedimiento (binomial exacto):

```
. centile numcig
```

| Variable | Obs | Percentile | Centile | — Binom. Interp. — [95% Conf. | Interval] |
|---|---|---|---|---|---|
| numcig | 25 | 50 | 20 | 15.52079 | 20 |

## 4.11. INTERVALOS DE CONFIANZA BAYESIANOS

Se pueden estimar intervalos de confianza desde el planteamiento bayesiano (v. apartado 3.8). Para realizarlos se parte de especificar de antemano *(a priori)* cuál sería el rango de valores subjetivamente esperables (fijando su máximo y su mínimo) para el parámetro que se intenta estimar. Se hace así explícita la creencia previa subjetiva de los investigadores y se incluye esa creencia en el cálculo del intervalo.

En el ejemplo de la prevalencia de obesidad podría asumirse que tal creencia previa admitiría que la prevalencia poblacional de obesidad en España, con un 95% de *seguridad,* estará entre el 15 y el 30%. Según la teoría bayesiana, este rango subjetivamente creíble para un parámetro se denominaría un *apriorismo* (*prior,* en inglés). Imagínese, por ejemplo, que se preguntó a un investigador experto en el tema: ¿cuál es la probabilidad de que un intervalo entre 0,15 y 0,30 incluya la verdadera proporción de obesos en España? El investigador contesta que ese intervalo tiene un 95% de probabilidad. Esto no se debe confundir con el intervalo de confianza. Es solo la creencia *subjetiva* de tal investigador.

El método consiste en obtener una media ponderada por los inversos de los cuadrados de los errores estándar (**1**) de la creencia subjetiva previa (0,15 a 0,30) y del resultado muestral. Si se toma la muestra de 12.036 participantes del estudio ENRICA (**21**), que encontró una prevalencia del 22,9% (IC 95%: 22,1 a 23,7%), se procedería así:

$$\text{Seudoerror estándar del apriorismo } (\text{PsEE}_{a\ priori}) = \frac{\text{LS}-\text{LI}}{2\times z} = \frac{0{,}3-0{,}15}{2\times 1{,}96} = 0{,}038$$

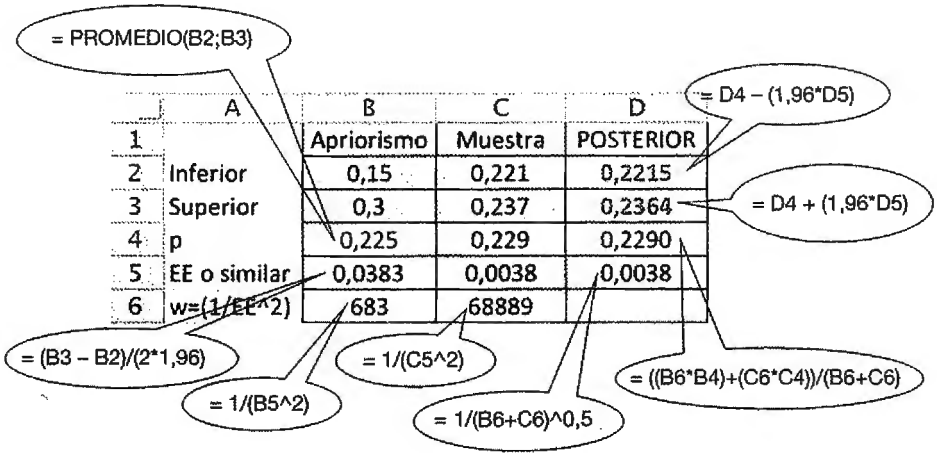
Este cálculo está basado en que la distribución normal tiene el 95% de probabilidad entre ±1,96 desviaciones típicas de la media.

$$\text{Proporción } \textit{a priori } (\text{p}_{\text{prior}}) = \frac{\text{LS}+\text{LI}}{2} = \frac{0{,}3+0{,}15}{2} = 0{,}225$$

$$\text{Ponderación del apriorismo } (\text{w}_{\text{prior}}) = \left(\frac{1}{\text{PsEE}_{a\ priori}}\right)^2 = \left(\frac{1}{0{,}038}\right)^2 = 693$$

$$\text{Error estándar de la proporción muestral (EEP)} = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0{,}229\times 0{,}771}{12.036}} = 0{,}0038$$

$$\text{Ponderación muestral } (\text{w}_{\text{muestra}}) = \left(\frac{1}{\text{EEP}}\right)^2 = \left(\frac{1}{0{,}0038}\right)^2 = 69.252$$




| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Apriorismo | Muestra | POSTERIOR |
| 2 | Inferior | 0,15 | 0,221 | 0,2215 |
| 3 | Superior | 0,3 | 0,237 | 0,2364 |
| 4 | p | 0,225 | 0,229 | 0,2290 |
| 5 | EE o similar | 0,0383 | 0,0038 | 0,0038 |
| 6 | w=(1/EE^2) | 683 | 68889 | |

**Figura 4.11** Excel programado para intervalos de confianza bayesianos.

$$p_{post} = \frac{w_{prior}\, p_{prior} + w_{muestra}\, \hat{p}}{w_{prior} + w_{muestra}} = \frac{693 \times 0,225 + 69.252 \times 0,229}{693 + 69.252} = 0,229$$

Lo anterior viene a ser simplemente una media ponderada, donde la información muestral *pesa* unas 100 veces más (69.252) que la creencia previa (693); por lo tanto, la muestra arrastra todo el cálculo hacia su valor.

$$EEP_{post} = \sqrt{\frac{1}{w_{prior} + w_{muestra}}} = \sqrt{\frac{1}{693 + 69.252}} = 0,0038$$

$$\text{IC 95\% bayesiano} = p_{post} \pm z_{\alpha/2} \times EEP_{post} = 0,229 \pm (1,96 \times 0,0038) = 0,2215 - 0,2364.$$

En este ejemplo, la muestra ha pesado muchísimo más que la creencia previa. El apriorismo (0,15-0,30), por otra parte, no era muy dispar de lo que se halló en la muestra (0,213-0,237). Por eso, el intervalo bayesiano coincide casi por entero con el frecuentista. Además, el apriorismo no era muy fuerte, al ser sus límites algo vagos (anchos). Cuanto más ancho sea el apriorismo, menor peso tendrá. La figura 4.11 muestra una hoja de Excel programada para estos cálculos.

En el otro ejemplo, antes visto, de 64 pacientes, se obtuvo una media de conformidad con dieta mediterránea de 8,5 puntos. Imagínese que este resultado fuese una sorpresa para los investigadores, porque ellos tenían una creencia subjetiva previa *fuerte* y, además, *contradictoria* con el resultado. Imagínese que su *apriorismo* consistiese en tener una seguridad del 95% de que la media poblacional estaría entre 6 y 7, con sigma ($\sigma$) = 1. La solución a este problema[4] sería una media bayesiana posterior = 7,52 (IC 95%: 7,17 a 7,87). Ahora, el peso de la creencia previa es casi el mismo que el de la muestra real, y la estimación bayesiana constituye una solución de compromiso que se encuentra en el terreno intermedio entre la media muestral y la creencia *a priori*. Se han publicado tutoriales sencillos sobre intervalos de confianza bayesianos para otros estimadores como la *odds ratio* o el riesgo relativo (**23**).

4 Se puede descargar el fichero Excel denominado *Intervalo de confianza bayesiano para una media* con la solución en la página web: www.unav.es/departamento/preventiva/recursos_bioestadistica.

## 4.12. CONTRASTE DE HIPÓTESIS

Estimar parámetros mediante *intervalos de confianza* parece lógico y amigable. Paradójicamente, se otorga preferencia a otro enfoque, indirecto y más complejo, llamado *contraste de hipótesis, que viene a ser como la otra cara de la moneda.* Cada vez se plantean más críticas sobre un contraste de hipótesis, especialmente cuando se convierte en un *ritual* mecánico y acrítico (**17,24-28**).

Un contraste de hipótesis consiste en una comparación. Volviendo al ejemplo de la reversión en la arteriosclerosis de la carótida con dieta rica en aceite de oliva virgen, se podría *comparar* el cambio entre antes y después. La reducción (–0,093 mm) sería el *efecto* de esta dieta. En un contraste de hipótesis se compara un *efecto* encontrado en una muestra con su variabilidad aleatoria esperada *(error).* Si el *efecto* en la muestra es muy superior a tal *error,* se concluirá con un apoyo a la existencia del efecto en la población. Si el efecto es similar o inferior al error, se concluye que la muestra no apoya el efecto en la población.

Como primer paso se establecen dos hipótesis opuestas sobre la *población:*

- *Hipótesis nula ($H_0$):* mantiene que el efecto de interés *no* existe (es nulo, es decir vale 0) en la población de la que procede la muestra.
- *Hipótesis alternativa ($H_1$):* mantiene que *existe* algún efecto distinto de 0 en la población de la que procede la muestra. Hay que tener en cuenta que este efecto podría ser de muy diversas magnitudes. Incluso aunque se demuestre que tal efecto existe, podría resultar insignificante en términos prácticos.

A continuación se usan los datos para intentar rechazar la hipótesis nula y optar por la hipótesis alternativa. Se decidirá entre una y otra. Cuando se rechaza $H_0$ se dirá que la comparación resultó estadísticamente *significativa* (tabla 4.5) y se concluirá que los datos apoyaban la hipótesis *alternativa.* Las hipótesis (nula o alternativa) se plantean para la población, no para la muestra. Sin embargo, los datos que se usan en el contraste se obtienen en la muestra.

Lamentablemente, el contraste de hipótesis mal usado puede llevar al *automatismo* y acabar por convertirse en un libro de recetas prefabricadas como sucedáneo del raciocinio. Es imprescindible entenderlo bien para que esto no suceda.

## 4.13. HIPÓTESIS NULA E HIPÓTESIS ALTERNATIVA

Los cuatro pasos de un contraste de hipótesis son:

1. Formular una hipótesis nula ($H_0$) *a priori* y su correspondiente alternativa ($H_1$).
2. Contrastar la hipótesis nula con los resultados hallados en la muestra.
3. Calcular la probabilidad *(valor* p *de significación estadística)* de encontrar unos resultados como los hallados o más distantes aún de $H_0$, si $H_0$ fuese cierta.
4. Decidir si se rechaza o no la hipótesis nula.

A continuación se aplican estos cuatro pasos a un ejemplo simplificado, pero inspirado en datos de hallazgos reales (**29**). Imagínese que en un estudio ocurrieron nueve embarazos imprevistos durante 1 año en 500 mujeres que seguían el método sintotérmico. Se trata de uno de los métodos de regulación de fertilidad basados en el conocimiento natural de la fertilidad *(fertility awareness-based method)* y que capacitan a los esposos para identificar el período del ciclo menstrual en que la mujer es fértil; así, evitan los efectos adversos de anticonceptivos hormonales y otros métodos. El sintotérmico combina varios indicadores: secreción cervical, temperatura basal y gráficas de calendario[5].

5 www.renafer.org/.

La hipótesis nula que se postula (para desmontarla luego) es que los fracasos de este método serían equivalentes a otros métodos (de referencia) más usados y que tienen una eficacia del 96% (índice de fallos = 4%, es decir $\pi = 0{,}04$).

1. *Formular una hipótesis nula a priori (y su correspondiente alternativa).*

   La hipótesis nula ($H_0$) mantendrá que no existen diferencias entre el sintotérmico y el método de referencia en cuanto a la proporción *($\pi$)* de embarazos imprevistos en la población de la que procede la muestra. Por tanto, las únicas diferencias que se encontrarían en la muestra serían debidas al azar (por error de muestreo).

   $H_0$: *la proporción de embarazos imprevistos será igual con uno u otro método.*

   $H_0 \equiv \pi_{\text{sintotérmico}} - \pi_{\text{teórico}} = 0.$

   $H_0 \equiv \pi_{\text{sintotérmico}} = \pi_{\text{teórico}} = 0{,}04.$

   $H_0$ tiene la apariencia de ser *lo contrario* de lo que se desea demostrar.

   La hipótesis alternativa ($H_1$) propondría que el método sintotérmico no es igual, sino que su tasa de fallos es inferior a la proporción de referencia. Esto es lo que los investigadores pretenden demostrar.

2. *Contrastar la hipótesis nula con los resultados muestrales de la investigación.*

   En la muestra se halló una proporción de embarazos de 0,018 (9/500) con el método sintotérmico, mientras que $H_0$ mantenía que esta proporción sería 0,04.

   Si $H_0$ fuese cierta, la diferencia entre ambas proporciones en la población sería exactamente igual a 0, pero he aquí que en la muestra es de –0,022 (0,018 – 0,04 = –0,022). El *efecto* observado, es, por tanto, –0,022.

$$\text{Efecto} = p_{\text{muestral}} - \pi \,|\, H_0 = 0{,}018 - 0{,}04 = -0{,}022.$$

   La barra vertical significa «condicionado a».

   Aunque $H_0$ fuese cierta en la población, prácticamente nunca se hallaría en la muestra una diferencia con respecto a $H_0$ exactamente igual a 0, ya que existe el error de muestreo. La pregunta es: ¿esta diferencia (–0,022) se debe simplemente al azar o se debe a un *«efecto»* real (poblacional) por tener el método sintotérmico en realidad menos fallos que el 4% de referencia como mantiene $H_1$?

3. *Calcular la probabilidad de hallar unos resultados como los hallados o más distantes aún de $H_0$ bajo el supuesto de que $H_0$ sea cierta.*

   Si las diferencias encontradas (efecto) se debiesen solo al azar, ¿cuál sería la probabilidad de hallar este efecto o uno todavía mayor?

   Para responder a esta pregunta suele obtenerse un cociente efecto/error. El error estima la variabilidad esperable por el proceso de muestreo y habitualmente corresponde al error estándar:

$$\frac{\text{Efecto}}{\text{Error}} = \frac{\text{diferencia observado} - \text{esperado} \,|\, H_0}{\text{error estándar}}$$

   Este cociente es el núcleo común general de las ecuaciones que se usarán para muchos contrastes de hipótesis. Este cociente es interpretable como un modo de medir el efecto (–0,022, en el ejemplo) en *unidades* de error estándar. Mide cuántos errores estándar separan lo observado (0,018) de lo esperado (0,04), si $H_0$ fuese verdad.

Aquí se contrasta una proporción, por lo que el error estándar corresponderá a una proporción:

$$\frac{\text{Efecto}}{\text{Error}} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0{,}018-0{,}04}{\sqrt{\frac{0{,}04\times 0{,}96}{500}}} = \frac{-0{,}022}{0{,}0088}$$

($\pi$ = proporción teórica)

Ese cociente vale –2,51, lo cual significa que el efecto se separa de $H_0$ en 2,51 errores estándar, y que se separa hacia abajo (de ahí el signo menos). Como se asume ya de partida, como primer paso, que $H_0$ es cierta, se usarán $\pi$ y $1 - \pi$ para el error estándar en vez de usar $p$ y $q$.

La gran ventaja de este cociente es que se ajustará a una distribución teórica concreta bajo el supuesto de que $H_0$ fuese cierta. En muchos casos, como sucede en este ejemplo, seguirá la *distribución normal* **(15)**:

$$z = \frac{-0{,}022}{0{,}0088} = -2{,}51$$

Basta mirar en la normal la probabilidad que corresponde a esta $z$ para responder al paso 3.° del contraste de hipótesis y obtener *la probabilidad de hallar unos resultados como los hallados o más distantes aún de $H_0$ si $H_0$ fuese cierta.* A esta probabilidad se le llama *valor* p *de significación estadística.*

Se obtendrá tal probabilidad (valor $p$) desde la normal[6] del modo siguiente:

a. STATA

```
display normal(-2.51)
```

b. Excel

=DISTR.NORM.ESTAND(–2,51)

c. R/Splus

```
pnorm(-2.51)
```

Se obtendrá una probabilidad del 0,6%, es decir, $p = 0{,}006$.

$$p\left(dif \geq observada \middle| H_0\right) = 0{,}006$$

Se trata de una probabilidad condicionada. La condición es $H_0$.

En la figura 4.12 se representa esta probabilidad. Se representa la distribución que tendrían las proporciones muestrales (distribución de los estimadores muestrales) si se fuesen obteniendo muchas muestras sucesivas, todas del mismo tamaño ($n$ = 500) de una población en la que $H_0$ se cumpliera ($\pi_{\text{sintomático}} = 0{,}04$). Las unidades de medida son errores estándar.

La probabilidad igual a 0,006 debe interpretarse como el porcentaje de muestras que sería esperable que estuviesen a esta distancia (a –2,51 errores estándar) *o todavía más lejos* de $H_0$, si $H_0$ fuese cierta en la población ($\pi = 0{,}04$). Por tanto, $z$ es la distancia desde nuestro resultado a la hipótesis nula. Pero esa distancia está medida en unidades de error estándar.

6 Se podría usar la binomial también y daría una probabilidad aun inferior, por ejemplo, en STATA: di binomial (500,9,0.04) → devolverá $p$ = 0.00438.

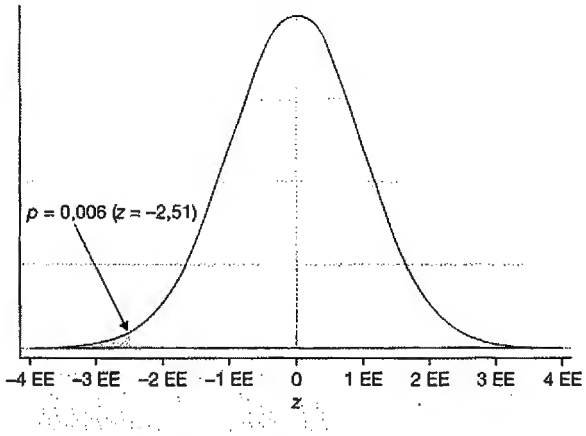



**Figura 4.12** Probabilidad de encontrar una muestra a 2,51 errores estándar o más lejos (por la izquierda) de lo esperado según $H_0$. Para un valor $z = -2{,}51$, la probabilidad a una cola es de 0,006.

Según *el teorema del límite central* (v. apartado 3.11), la distribución normal es la que siguen los estimadores calculados en las muestras. Gracias a esto se puede calcular aproximadamente la proporción de *muestras* de las muchas (con frecuencia infinitas) que se podrían extraer de una población que estarán más allá de una determinada distancia del parámetro poblacional. Al tratarse de estimadores (no de individuos), esta distancia se mide como número de *errores* estándar *(z)*. Por eso, el eje horizontal de la figura 4.6 mide las distancias como número de errores estándar.

Una vez obtenida $p = 0{,}006$, se puede pensar que si $H_0$ fuese cierta (en la población) sería bastante raro (6 de cada 1.000 veces) haber encontrado en la muestra una proporción de 0,018. Este resultado es poco compatible con $H_0$, ya que la probabilidad de haber encontrado la diferencia observada (o una mayor) en la muestra es *muy baja*. Esto conduce a decidir *en contra* de $H_0$. La hipótesis nula podría ser cierta, pero en ese caso habría aparecido una muestra extraordinariamente apartada de lo esperado. Es lógico decidir rechazar $H_0$, porque probablemente $H_0$ no sea cierta. Téngase en cuenta que *no* se ha calculado la probabilidad de que $H_0$ sea verdad, sino la probabilidad de observar esa muestra particular (u otra aún más extrema) *si $H_0$ fuese verdad*. A esta probabilidad se le llama *valor* p *de significación estadística*. Como ya se ha comentado, es una probabilidad condicionada. La condición es $H_0$.

***Interpretación de un valor* p: probabilidad de observar diferencias mayores o iguales a las observadas en la muestra si la hipótesis nula fuera cierta.**

$$\text{Valor } p = p\left(\text{dif} \geq \text{observadas} \middle| H_0\right)$$

**El valor *p* es una probabilidad condicionada. La condición es que $H_0$ sea cierta.**

4. *Decidir si se rechaza o no $H_0$*

Un valor *p* muy bajo significa que sería muy raro encontrar esta muestra si $H_0$ fuese cierta. En cambio, un valor *p* alto significa que, en principio, no hay argumentos en contra de la hipótesis nula y no se podrá rechazar.

**Cuanto *menor* sea el valor *p* de significación estadística, *mayores* argumentos habrá para rechazar la hipótesis nula y apoyar, en cambio, la hipótesis alternativa.**

Habitualmente, aunque es algo arbitrario y no puede dársele una consideración estricta, el límite para considerar bajo un valor *p* se suele poner en $p < 0,05$. Cuando $p < 0,05$, se suele rechazar la hipótesis nula y se dice que el resultado es «estadísticamente significativo». En este caso, habría que quedarse con la hipótesis alternativa ($H_1$). Se concluirá que el método sintotérmico tiene un porcentaje de *fallos significativamente inferior al 4%* (o una eficacia *significativamente* superior al 96%). En cambio, cuando $p \geq 0,10$, no se rechaza la hipótesis nula y se dice que no se alcanzó significación estadística (v. tabla 4.5).

***Tabla 4.5 Decisiones después de conocer el valor p de significación estadística***

| $P < 0,05$ | $P \geq 0,10$ |
|---|---|
| Se rechaza la hipótesis nula | No se puede rechazar la hipótesis nula |
| No parece que el azar lo explique todo | No se puede descartar que el azar lo explique todo |
| El «efecto» es mayor que el «error» | El «efecto» es similar al «error» |
| Hay diferencias estadísticamente significativas | No hay diferencias estadísticamente significativas |
| Existen evidencias a favor de la hipótesis alternativa | No existen evidencias a favor de la hipótesis alternativa |
| Los datos encontrados son poco compatibles con $H_0$ | Los datos encontrados son compatibles con $H_0$ |

**Nota:** *Los límites 0,05 y 0,10 son* arbitrarios, pero comúnmente aceptados.

Cuando *p* está entre 0,05 y 0,10 podría concluirse que se está en *tierra de nadie* y se hablaría de valores *próximos a la significación estadística.* En general, en esta circunstancia es mejor presentar el *valor* p *concreto* que se ha calculado y evitar decisiones dicotómicas radicales (**25,28,30**). Así, cada cual puede juzgar como crea conveniente. Por ejemplo, a la luz de un valor $p = 0,08$, se pensará que hay una probabilidad del 8% de encontrar unos datos tan alejados como estos o más de $H_0$, si $H_0$ fuese cierta; que cada lector juzgue si eso es suficientemente raro o no. Quizá lo más razonable sería aumentar el tamaño muestral, si aún es posible.

La principal y más grave equivocación al interpretar un valor *p* es creer ingenuamente que estima la probabilidad de que $H_0$ sea cierta.

***Principal equivocación:*** **pensar que un valor *p* corresponde a la probabilidad de que la hipótesis nula sea cierta.**

$$\text{valor } p \neq p(H_0)$$

**El valor *p no* es la probabilidad de que $H_0$ sea cierta.**

La probabilidad de que $H_0$ sea cierta *no* se puede calcular con un valor *p*. Es más, hay que asumir que $H_0$ es cierta para poder calcular el valor *p*. El valor *p* es una probabilidad condicionada y su condición es $H_0$.

No se podrá decir nunca, en el ejemplo anterior, que la probabilidad de que el método sintotérmico tenga una tasa de fallos del 4% es baja, del 0,6%. Lo correcto es pensar: si la proporción de fallos del método sintotérmico fuese del 4%, existiría una probabilidad muy baja (0,6%, o $p = 0,006$) de que en una muestra de 500 mujeres se produzcan nueve o menos embarazos.

## 4.14. ERRORES EN EL CONTRASTE DE HIPÓTESIS: ERROR TIPO 1 Y ERROR TIPO 2

La tabla 4.6 combina cuatro posibles decisiones.

Lo deseable es no rechazar $H_0$ cuando es cierta y rechazarla cuando es falsa (y entonces la verdadera es $H_1$). Se pueden cometer dos tipos de errores: el *error tipo 1* consiste en rechazar $H_0$ cuando no debería haberse rechazado, y el *error tipo 2* se comete al quedarse en $H_0$ cuando debería

*Tabla 4.6 Resultados de un contraste de hipótesis*

| | | VERDAD (REALIDAD) | |
|---|---|---|---|
| | | $H_0$ | $H_1$ |
| Decisión | $H_0$ | Acierto<br>Probabilidad = $1 - \alpha$ | Error (tipo 2)<br>Probabilidad = $\beta$ |
| | $H_1$ | Error (tipo 1)<br>Probabilidad = $\alpha$ | Acierto<br>Probabilidad = $1 - \beta$ = potencia |
| Decisión | $H_0$ | No se rechaza la hipótesis nula (el azar puede explicar todas las diferencias observadas en los datos) y es verdad | No se rechaza la hipótesis nula (se dice que no hay diferencias significativas), pero nos equivocamos |
| | $H_1$ | Se rechaza la hipótesis nula (se dice que los resultados son estadísticamente significativos), pero nos equivocamos | Se rechaza la hipótesis nula (se dice que los resultados son estadísticamente significativos) y es verdad |

*Se ha de tener en cuenta que la potencia de contraste ($1 - \beta$) es, en realidad, una función de cada uno de los posibles valores de la hipótesis alternativa.*

haberse rechazado porque la verdad es $H_1$. El error tipo 1 llevaría a decir que existen diferencias significativas, cuando realmente no las hay. Por su parte, en un error tipo 2 se diría que *no* existen diferencias significativas, cuando realmente las hay. Al riesgo de cometer un error tipo 1 se le llama *alfa* ($\alpha$) y al de cometer un error tipo 2, *beta* ($\beta$). El complementario de beta ($1 - \beta$) se denomina *potencia* estadística y corresponde a la probabilidad de encontrar resultados significativos si $H_1$ es cierta (cuadro 4.1; v. tabla 4.6).

A primera vista parecería que el riesgo $\alpha$ coincide con el valor $p$. No es así. El riesgo $\alpha$ es una probabilidad que el investigador fija de antemano, sin ni siquiera conocer los datos. Es el *umbral* o criterio fijado para su decisión y muchas veces será constante de uno a otro experimento (p. ej., $\alpha = 0{,}05$,). Alfa ($\alpha$) establece *a priori* qué riesgo de cometer un error tipo 1 se está dispuesto a admitir. En cambio, un valor $p$ se calcula *a posteriori* y surge de los datos analizados. El valor $p$ variará de un experimento a otro en función de que cambien los datos; $\alpha$ suele ser constante (**1,31**).

El riesgo $\alpha$ también se llama *nivel de significación*. Si el riesgo $\alpha$ establece el criterio de rechazar $H_0$ cada vez que se encuentre un valor $p$ inferior al 5% ($p < 0{,}05$), esto será lo que podrá ocurrir con 5 de cada 100 muestras que pudieran tomarse cuando la hipótesis nula sea cierta, y se dirá *incorrectamente* que hubo diferencias significativas. Tales rechazos incorrectos de $H_0$ son los errores tipo 1. Son graves porque vienen a ser como descubrimientos en falso que pueden traducirse en decisiones desafortunadas. Por eso se suele fijar un riesgo $\alpha$ pequeño.

En cambio, el error $\beta$ es menos grave porque ocurre en un contexto en que no se tomarán decisiones. Para que se cometa un error tipo 2 tiene que haberse concluido que no hubo significación estadística, y lo no significativo *no significa nada* (ni a favor ni en contra, no se decanta ni por la diferencia ni por la igualdad). La no significación obliga a *callarse*. El silencio no suele ser tan grave. De callar uno no suele arrepentirse. Suele ser peor equivocarse al hablar. Por eso, se admitirá un riesgo $\beta$ mayor que el $\alpha$. Además, se puede producir un error tipo 2 (probabilidad = $\beta$) porque el efecto sea pequeño (diferencias reales, pero de poca magnitud), porque el número de sujetos estudiados resulte escaso (muestra pequeña) o por ambas cosas a la vez.

### CUADRO 4.1 LOS OCHO CONCEPTOS CLAVE DEL CONTRASTE DE HIPÓTESIS

- *Hipótesis nula ($H_0$):* hipótesis que se pone a prueba tratando de rechazarla mediante una prueba estadística. Suele referirse a una igualdad o ausencia de asociación. Nunca se acaba *aceptando,* como mucho se afirma que *no se pudo rechazar.*
- *Hipótesis alternativa ($H_1$):* establece que sí existen diferencias. Es lo que se pretende demostrar. Contradice a la hipótesis nula y *se aceptará* (provisionalmente) cuando se pueda rechazar la hipótesis nula (por ser el valor *p* bajo).
- *Valor* p *(significación estadística):* probabilidad de observar las diferencias de la muestra u otras mayores, condicionalmente a que $H_0$ sea cierta.
- *Error tipo 1:* la hipótesis nula es, en realidad, cierta, pero, a pesar de todo, se rechaza (no hay ningún efecto, pero se concluye diciendo que existen diferencias *significativas*).
- *Error tipo 2:* no se rechaza la hipótesis nula cuando es en realidad falsa y se debería haber rechazado ($H_1$ y existe un efecto, pero se concluye que no hay significación estadística).
- *Riesgo* α: probabilidad de cometer un error de tipo 1. Es un umbral de probabilidad establecido *a priori* como regla de decisión, de modo que cuando *p* sea inferior a $\alpha$, se rechazará la hipótesis nula. Un riesgo $\alpha$ del 5% supone aceptar que en 5 de cada 100 muestras que pudieran tomarse cuando $H_0$ sea cierta se concluirá erróneamente que hubo diferencias *significativas*.

$$\alpha = P(\text{decidir } H_1 \mid H_0 \text{ es cierta})$$

- *Riesgo* β: probabilidad de cometer un error tipo 2. Un riesgo $\beta$ del 20% supone aceptar que de cada 100 veces que exista efecto ($H_1$ es cierta), este no se detectará.

$$\beta = P(\text{decidir } H_0 \mid H_1 \text{ es cierta})$$

- *Potencia estadística:* capacidad de una prueba para detectar una diferencia cuando esta existe realmente ($H_1$ es cierta). La potencia es el complementario de $\beta$:

$$\text{Potencia} = 1 - \beta$$
$$\text{Potencia} = p(\text{rechazar } H_0 \mid H_0 \text{ es falsa})$$
$$\text{Potencia} = p(\text{optar por } H_1 \mid H_1 \text{ es cierta})$$

## 4.15. INTERPRETACIÓN DE VALORES *p*: SIGNIFICACIÓN ESTADÍSTICA

Deben entenderse muy bien los valores *p* de significación estadística. Su uso es extremadamente frecuente y, además, es conocido que, lamentablemente, los valores *p* del resultado principal de una investigación pueden ser determinantes en la decisión de publicar o no esa investigación o de poder publicarla en una revista científica más importante (**32**). Algunos investigadores admiten que será inútil intentar publicar un resultado cuya *p* sea >0,10. La pena es que ciertos revisores y editores de revistas médicas colaboran para que esto ocurra.

Un valor *p* inferior a 0,05 se equipara a un resultado *significativo.* Esto no deja de ser una convención y es cuestionable. Esta convención procede de cuando solo había tablas (y no ordenadores) que daban valores de las distribuciones para $p < 0,05$, $p < 0,01$, etc. Se quedó ahí cristalizado para siempre el valor $p < 0,05$ como *árbitro de la verdad.* No conviene darle un valor absoluto. Por ejemplo, están aumentando las propuestas para usar como significativo un límite 10 veces inferior ($\alpha = 0,005$) ya que valores *p* que están sólo ligeramente por debajo de 0,05 corresponden a estudios poco replicables (**33**).

Es preciso también ser moderado y no presentar nunca resultados acompañados de una expresión como $p < 0,000000001$. Basta con indicar: $p < 0,001$. La sobriedad es preferible. Los ordenadores darán salidas del tipo $p = 0,000$, lo cual es matemáticamente incorrecto y no

debe presentarse así ni en las tablas ni en el texto de un manuscrito o comunicación científica. Si aparece $p = 0{,}000$ en el ordenador, se comunicará como $p < 0{,}001$.

Para valores $p$ entre 0,001 y 0,20 se recomienda comunicarlos con tres decimales, por ejemplo, $p = 0{,}137$. Para valores mayores de 0,20 basta con dos decimales, como $p = 0{,}28$. Los valores $p$ menores de 0,001 se presentarán sobriamente como $p < 0{,}001$. Nunca es aconsejable presentar *n.s.* o $p$ = ns (poniendo luego a pie de página o de tabla que n.s. indica *no significativo*). Se debe evitar hablar de un resultado como *positivo* si $p < 0{,}05$ o *negativo* si $p > 0{,}05$.

El valor $p$ se calcula en una *muestra,* pero permite extraer una conclusión acerca de una *población.* En esto consiste la estadística *inferencial.* El valor $p$ depende de tres elementos: el test estadístico, la hipótesis nula y la hipótesis alternativa (**34**). Un valor $p$ es una *probabilidad.* Se corresponde con la probabilidad de que, simplemente por variación aleatoria (error de muestreo), se produzca la diferencia o desigualdad encontrada en una muestra, si no existiese en absoluto desigualdad en la población de la que procede esa muestra. Conviene añadir un matiz importante: un valor $p$ es la probabilidad de que se produzca el resultado encontrado *o uno más desigual todavía.*

Los valores $p$ miden la fuerza de la evidencia estadística en muchos estudios científicos, pero *no* miden la *magnitud* de la asociación. Pero cuanto menor sea un valor $p$, mayor fuerza tienen las evidencias aportadas para apoyar una hipótesis alternativa. Los valores $p$ solo ofrecen un modo de comunicar los resultados de una investigación, y no definen en sí mismos la importancia práctica de esos resultados. La importancia suele darla la magnitud del efecto y esta *magnitud* depende de las diferencias absolutas encontradas, no del valor $p$.

Los valores $p$ vienen a ser también una medición del grado de *sorpresa* ante un resultado. Cuanto menor sea un valor $p$, mayor sería la sorpresa por ese resultado (**34**). Lanzar simultáneamente cinco dados y que en todos salga el uno no deja de producir sorpresa. Se puede calcular la probabilidad de que esto suceda al azar, para cuantificar así la admiración o extrañeza ante tal resultado. La distribución binomial proporciona esta probabilidad, que es $0{,}00013 = (1/6)^5$ y coincide conceptualmente con un valor $p$. Si todo se debiese enteramente al azar ($H_0$), una $p$ tan pequeña solo indicaría que ocurrió algo muy raro o muy poco probable. Pero puede pasar. No siempre que $p$ sea menor que 0,05, será verdad $H_1$. Una $p$ baja puede ser compatible con que $H_0$ sea verdad, porque el azar puede deparar sorpresas. Valores $p$ pequeños indican que un suceso raro ha acontecido por azar o que existe un efecto que crea una desigualdad sistemática.

Para resolver esta disyuntiva es preciso tener en cuenta el diseño del estudio, la consistencia de estos hallazgos con otras investigaciones, la plausibilidad biológica y otros criterios epidemiológicos de causalidad (**1,3,9,35-38**). Aunque ante un valor $p$ muy bajo habitualmente se tiende a asumir que existe un efecto, todo puede haberse debido al azar y la estadística no suele tener nunca la última palabra en este juicio. La estadística bayesiana cuenta más con estos otros factores. El contraste de hipótesis habitual (frecuentista) asume que solo el error aleatorio en uno de sus aspectos (error de muestreo) explicará las discrepancias entre $H_0$ y la muestra (**1**). Este planteamiento olvida los sesgos (errores sistemáticos), otros errores o, simplemente, la posibilidad de que el modelo matemático que se ha asumido sea inadecuado. Además, incluso con un supuesto modelo perfecto, que no existe, y sin sesgos ni otros errores, el 5% de las ocasiones en que el resultado sea *significativo* el supuesto hallazgo será mentira si se asume un riesgo $\alpha$ del 5%. Por eso no puede absolutizarse un valor $p < 0{,}05$ como criterio de *verdad* de $H_1$ (ni *mucho menos* un valor $p > 0{,}05$ como criterio de su falsedad). No hay que olvidar que, aunque la probabilidad de un suceso raro es muy pequeña, pueden ocurrir muchos sucesos raros, y la probabilidad de que suceda al menos uno ya no es tan reducida. De ahí tantas casualidades que se acaban observando diariamente.

## 4.16. SIGNIFICACIÓN ESTADÍSTICA FRENTE A SIGNIFICACIÓN PRÁCTICA

Los estudios se llevan a cabo para detectar diferencias *importantes*, no cualquier diferencia (**39**). Esto hace que deba distinguirse entre significación estadística (valor *p*) y significación *práctica* (magnitud de la asociación). La *magnitud de la asociación* se refiere a la diferencia concreta que existe entre dos medias o entre dos proporciones o, en epidemiología, al valor del riesgo relativo, *odds ratio* (**40**), *hazard ratio* (**41**) o razón de tasas. En general, en medicina se suele admitir que la *diferencia mínimamente importante* (DMI) es la menor diferencia que los pacientes perciben como importante y que se suele traducir en cambios en la actuación diagnóstica o terapéutica de los profesionales sanitarios sobre dichos pacientes (**42**). Podría decirse algo análogo en otros terrenos no médicos de la ciencia. En el ejemplo del método sintotérmico, la importancia *práctica* no vendría dada por la significación estadística ($p = 0{,}006$), sino por la magnitud de la diferencia (el 2,2% *menos* de fallos con este método), y habría que preguntarse cuál es la percepción de una mujer sobre el beneficio de pasar de una eficacia del 96% a otra del 98,8%. Esto se refiere también como significación *clínica* en medicina. Es lo más importante.

Aunque una diferencia sea muy pequeña en cuanto a su magnitud absoluta, siempre que se disponga de gran tamaño muestral podría acabar por ser *estadísticamente significativa*. Aumentar la eficacia en un 0,1% puede producir diferencias significativas con una muestra de muchos miles de personas. Es más que dudoso que una diferencia solo del 0,1% tenga relevancia práctica alguna. Por eso, con frecuencia, en muchos modelos se suele introducir el coste que llevaría consigo un cambio.

En el juicio sobre *significación clínica* interviene la magnitud de la diferencia hallada y otros elementos que no son estadísticos, entre ellos efectos adversos, aceptabilidad del tratamiento, costes y otros asuntos que deben considerarse juiciosamente como posibles pros y contras de las alternativas comparadas.

## 4.17. PRUEBAS A UNA COLA Y PRUEBAS A DOS COLAS

En el ejemplo del método sintotérmico, $H_0$ mantenía que la probabilidad de fallo era igual a la del método de referencia ($H_0 = 0{,}04$), pero en la muestra se encontró una proporción de 0,018. Se calculó la probabilidad de hallar ese resultado o uno más alejado de $H_0$, según el supuesto de que $H_0$ fuese cierta. No obstante, se pueden imaginar otras muestras que diesen resultados más alejados de $H_0$, pero en dirección opuesta. Tal sería el caso, por ejemplo, de una proporción *muestral* del 8% de fallos. Ese posible resultado (0,08) estaría más alejado de $H_0$ ($\pi = 0{,}04$) que el del ejemplo anterior (0,018), pero por el otro lado. Cumple, por tanto, con el criterio de estar *más alejado que lo observado* (aunque se aleje por el lado opuesto).

Lo más común es plantear el contraste de hipótesis con ambas posibilidades. Este tipo de contraste se llama *bilateral*, y las pruebas estadísticas que se utilizan se denominan *de dos colas* (**43**). En cambio, si en el cómputo del valor *p* se incluyen solo las posibilidades más alejadas de lo observado en un solo sentido, los test serían a una cola, como el que se ha calculado antes ($z = -2{,}51$, $p_{1\ cola} = 0{,}006$).

La consecuencia práctica es que, si son verosímiles tanto una posibilidad como la otra, se deberán tener en cuenta ambos valores ($+z$ y $-z$) y después se sumará el área de las dos colas para obtener la probabilidad (valor *p*). Cuando un contraste de hipótesis se basa en la normal o la *t* de Student, el valor *p* a dos colas es *doble* que el valor *p* a una cola.

En el ejemplo del método sintotérmico, el valor *p* a una cola fue $p_{1\ cola} = 0{,}006$; si se plantease a dos colas, el valor *p* sería $p_{2\ colas} = 0{,}012$. En este ejemplo, la prueba resultaría estadísticamente significativa (asumiendo $\alpha = 0{,}05$), tanto a una cola como a dos. Cuando una prueba bilateral es significativa, también lo será una prueba unilateral. Las pruebas a dos colas siempre dan valores de *p* *mayores* (y, por tanto, *menores* posibilidades de alcanzar la significación estadística) que las de una cola. A veces, un investigador tendencioso podría tener la tentación de hacer trampas y, al comprobar que no le resulta significativa una prueba a dos colas, ampararse en que la prueba a una cola sí tiene un valor $p < 0{,}05$. Las pruebas a una cola levantan sospechas y se desaconsejan por principio. En todo caso, la decisión

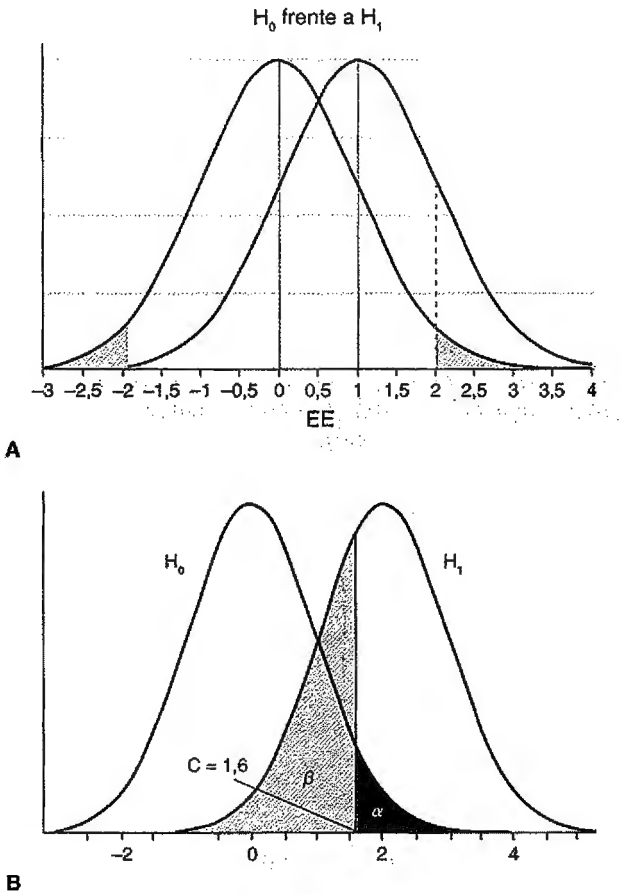



**Figura 4.13** **A.** Planteamiento del contraste de hipótesis. La campana de la izquierda asume que $H_0$ es verdad en la población. Las zonas sombreadas en esa campana corresponden a $\alpha/2$ si se asume que $\alpha = 0{,}05$ y, por tanto, $z = \pm 1{,}96$. La campana de la derecha asume que $H_1$ es verdad en la población y representa cómo se distribuirán las muestras tomadas de una población en que $H_1$ es verdad. El riesgo $\beta$ es la porción del área de esa segunda campana que queda a la izquierda de la línea discontinua. **B.** Ejemplo con hipótesis simples en el que el valor crítico para tomar la decisión (C = 1,6) determina los valores de $\alpha$ y $\beta$.

de usar una prueba a una cola debería tomarse *antes* de iniciar el análisis de los datos, dejándola por escrito en el registro del ensayo y en el plan de análisis estadístico elaborados de antemano. Debe ser una decisión bien fundamentada en el diseño o en el estado de conocimientos. Así, si ya se dispone de una hipótesis previa sólidamente basada en evidencias científicas, se podría aceptar el uso de pruebas a una cola. También se podría justificar cuando la relevancia clínica o biológica de una alternativa en sentido distinto a la prevista no representa ningún hallazgo de interés.

Estas son excepciones. En general, es preferible usar pruebas a dos colas, ya que, si se usa una prueba a una cola, siempre se puede denunciar que si está tan claro que el efecto va a ir en un solo sentido, ¿qué justificación tiene seguir investigando?

En la figura 4.13 se presenta el planteamiento *a priori* de un contraste de hipótesis a dos colas. $H_0$ mantiene que el parámetro vale 0. La campana de la izquierda asume que $H_0$ es verdad en

la población y representa cómo se distribuirán las muestras tomadas de una población donde $H_0$ se cumpla. Seguirán una distribución normal con media = 0. Las zonas sombreadas en esa campana corresponden a $\alpha/2$ a cada lado si se asume que $\alpha = 0{,}05$ ($\alpha/2 = 0{,}025$) y, por tanto, $z = \pm 1{,}96$. Cuando el estimador muestral esté más lejos de 1,96 errores estándar a un lado u otro de la campana, se rechazará $H_0$, pero se habrá cometido un error tipo 1. La zona sombreada es el porcentaje de muestras que proceden de una población donde $H_0$ es verdad y, sin embargo, conducen equivocadamente a rechazar $H_0$. La mitad de esas posibles muestras está en cada una de las dos colas.

$H_1$ (campana de la derecha en la figura 4.13) mantiene que el verdadero parámetro poblacional está un error estándar (EE) por encima de $H_0$. La campana de la derecha asume que $H_1$ es verdad en la población y representa cómo se distribuirán las posibles muestras tomadas de esa población donde $H_1$ es verdad. Algunas de esas muestras caerán muy cerca del 0, lo cual llevará a no poder rechazar $H_0$, aunque se hayan obtenido de una población en la que $H_1$ es cierta. Serán errores de tipo 2. Su probabilidad (riesgo $\beta$) es el área bajo la campana de la derecha que queda a la derecha de la línea discontinua. $\beta$ es siempre a una cola.

Lo ideal es usar pruebas cuyos riesgos $\alpha$ y $\beta$ sean los menores posibles. Esto requiere un cierto compromiso, ya que reducir $\alpha$ supone incrementar $\beta$, y viceversa, como se puede imaginar estudiando la figura 4.13B. Si se desplaza a la derecha la regla de decisión para rechazar $H_0$, se minimizará el riesgo $\alpha$, pero aumentará el $\beta$. Solo con un mayor tamaño muestral se conseguirá que los errores estándar sean menores, y menores serán, por tanto, $\alpha$ y $\beta$, pues se reduce toda la escala horizontal y, por consiguiente, las áreas correspondientes a $\alpha$ y $\beta$. Como se ha visto, interesa que $\alpha$ sea menor que $\beta$ (v. apartado 7.4).

## 4.18. PRUEBAS DE CONTRASTE DE HIPÓTESIS FRENTE A INTERVALOS DE CONFIANZA

La literatura científica está inundada de la expresión *estadísticamente significativo* o *no significativo estadísticamente*. Sin embargo, resulta pobre reducir a esta dicotomía las conclusiones posibles de un estudio. El veredicto con apariencia de objetivo y concluyente resulta, además, falaz. Por eso, se ha recomendado que nunca se presenten en las publicaciones solo valores *p* como resumen de la información estadística de una investigación, sino que *se acompañen siempre de intervalos de confianza*. El nivel de confianza es equivalente al complementario del error $\alpha$, es decir, $(1 - \alpha)$. Un intervalo de confianza no solo cuantifica el hallazgo en una *escala continua* (en vez de dicotómica), sino que indica también la mayor o menor precisión de los resultados (cuanto más estrecho sea el intervalo, tanto más preciso es el resultado) e informa indirectamente de la potencia estadística (los intervalos estrechos denotan mayor potencia).

Al comparar dos grupos, el valor *p* (a dos colas) será inferior a 0,05 (*significativo*) solo cuando el intervalo de confianza al 95% *no* incluya el 0 (o, de modo general, no incluya la hipótesis nula). Si el intervalo de confianza $(1 - \alpha)$ *no incluye al valor correspondiente a la hipótesis nula*, se puede afirmar que el efecto es significativo con un valor *p* a dos colas inferior a $\alpha$.

Por ejemplo, si la diferencia de la nota media de chicos y chicas en una clase es de +0,5 puntos a favor de las chicas y su intervalo de confianza al 95% va desde –0,3 a +1,3 puntos, ese intervalo de confianza está incluyendo la hipótesis nula (diferencia = 0) y permite decir que no hay diferencias significativas entre chicos y chicas. En cambio, si las chicas faltan *menos* horas a clase con diferencia de medias = –3 horas; IC 95%: –4,5 a –1,5), se deduce que los chicos faltan significativamente más horas a clase ($p < 0{,}05$). Pero si se redujese a esto la interpretación de un intervalo de confianza, se estaría incurriendo en el mismo error de la simplificación y dicotomía de los valores *p*. La clara ventaja del intervalo de confianza sobre el valor *p* solo se aprecia cuando se pondera todo el rango del intervalo de confianza en escala continua y en relación con la posible magnitud del efecto.

Cuanto menor sea el tamaño muestral, mayor será el error estándar. Como el intervalo de confianza se determina sumando y restando a lo encontrado un múltiplo de este error estándar, es fácil pensar que, al reducir el tamaño de las muestras, se ensanchan los intervalos de confianza. Las muestras pequeñas tienen baja potencia y originan intervalos de confianza anchos. Un lector inteligente y bien formado concluirá, al ver un intervalo de confianza ancho, que ese estudio tenía poca potencia.

Un intervalo de confianza ancho implica poca potencia estadística.

## 4.19. POTENCIA ESTADÍSTICA

Como se ha dicho (v. cuadro 4.1), la *potencia* de un test es la probabilidad de rechazar correctamente la hipótesis nula, es decir, de rechazar $H_0$ cuando es falsa. Es una probabilidad condicionada y su condición reside en que $H_1$ sea verdad. La potencia estima la probabilidad de concluir con un resultado significativo cuando la hipótesis alternativa es realmente cierta. Un test con mayor potencia es aquel que acierta más cuando $H_1$ es cierta y detecta con más facilidad los efectos que realmente existen.

Se ha de tener en cuenta que la potencia del contraste $(1 - \beta)$ es una función de cada uno de los posibles valores de la hipótesis alternativa. En la práctica, una vez observados los datos se calcula a veces la potencia empírica (u observada), en la que se toma como valor de la hipótesis alternativa el observado en la muestra. No obstante, los cálculos de la potencia estadística una vez que el estudio ya sido realizado *(post hoc)* pueden ser *fútiles*, tanto más si el estudio no resultó significativo y se usan las diferencias halladas como estimación de la magnitud del efecto para calcular la potencia. Esta situación se ha criticado como una tautología o razonamiento circular, porque inevitablemente concluirá que el estudio tenía poca potencia (**44**).

*Potencia estadística:* capacidad de detectar una diferencia cuando esta existe realmente: $p$ (acierto | $H_1$ sea cierta).

Cuando $H_1$ es cierta (existe un efecto), lo deseable es poder rechazar $H_0$ (hallar significación estadística). Una prueba con poca potencia no dará resultados significativos muchas veces en las que $H_1$ sea cierta y, por tanto, se equivocará. Esta situación es frecuente en estudios de escaso tamaño muestral, y aunque se diga no se hallaron diferencias significativas, es posible que sí existan diferencias de cierta magnitud en la población de la que procedía la muestra (cuanto menor tamaño muestral, más probable es el error tipo 2).

La característica que influye más decisivamente en el cálculo de los valores $p$ es el *tamaño de la muestra.* Con muestras de pequeño tamaño ($n < 30$), salvo que haya diferencias enormes entre grupos o entre mediciones, es fácil que los valores $p$ sean superiores a 0,10 y no permitan rechazar la hipótesis nula. Un tamaño muestral inadecuado es generalmente la causa de una baja potencia para detectar efectos clínicamente relevantes (**15,31**). Por lo tanto, cuando un test estadístico resulte no significativo, sería erróneo concluir que se ha *demostrado* la verdad de la hipótesis nula.

Nunca debe concluirse que se ha *demostrado* la igualdad cuando se emplea una muestra pequeña y no se encuentran diferencias significativas.

Hay que ser crítico y tener en cuenta cuál es el tamaño muestral, porque puede haberse cometido fácilmente un error tipo 2 por falta de potencia estadística.

Quien lee un estudio en el que *no se encuentran diferencias estadísticamente significativas* entre dos tratamientos no debe pensar erróneamente que los autores han *demostrado* que ambos tratamientos son iguales o que el nuevo tratamiento no añade ninguna ventaja sobre el antiguo. Peor sería aún que pensasen así los autores. Se deben usar expresiones moderadas del tipo *no se encontraron evidencias para rechazar la hipótesis nula* o *no se pudo rechazar la hipótesis nula.* Nunca se habla de *aceptar* la hipótesis nula. Aunque parezca rebuscado, es mejor proceder así, porque un test estadístico que no sea significativo nunca *demuestra* que la hipótesis nula sea

cierta. Lo *no significativo* no significa nada y simplemente señala que no se debe concluir nada. Bastaría con seleccionar una muestra muy pequeña para *demostrar* así tendenciosamente la igualdad, algo claramente incorrecto e inaceptable. Lo adecuado será concluir que el estudio no fue informativo y no permite extraer conclusiones prácticas. Si así sucede, ¿qué se puede hacer entonces para demostrar la igualdad?

## 4.20. ESTUDIOS QUE DEMUESTRAN EQUIVALENCIA Y ESTUDIOS DE NO INFERIORIDAD

Para demostrar la igualdad hay que cambiar el planteamiento del contraste de hipótesis, en el diseño y en el análisis de datos, y realizar lo que se conoce como *ensayo de equivalencia*. Estos estudios exigen fijar de antemano una diferencia *(d)* que se acepta que no llega a ser relevante desde el punto de vista práctico pues es tan pequeña que será *inferior* a la diferencia mínima clínicamente importante o importante desde el punto de vista práctico (**39,42**). Lamentablemente, no abundan los estudios de equivalencia.

Imagínese que se desea demostrar la igualdad de las notas de los chicos con las de las chicas en un examen de 100 preguntas de tipo test. Lo primero sería delimitar una diferencia tan pequeña que fuese inferior a la mínima diferencia que tiene importancia práctica. Se decide que estar un punto por encima o por debajo sobre un total de 100 puntos es indiferente y carece de relevancia. Por tanto d = ± 1. Después se toma una muestra de chicos y otra de chicas y se calcula el intervalo de confianza para ver la diferencia observada en sus notas. Si *todo el intervalo* de confianza quedase dentro del margen fijado *a priori* (±1 punto), se habrá demostrado la igualdad.

- En la figura 4.14 se presentan cinco posibles resultados de cinco posibles estudios (1 a 5) que trataban de demostrar la igualdad entre dos grupos A (chicas) y B (chicos), con el intervalo de confianza al 95% (IC 95%) para las diferencias entre ellos y ellas. El único estudio que demostrará la igualdad será el 3.
- El 1 y 2 encuentran diferencias significativas (en el 1, las chicas sacaron 4 puntos más, y en el 2, los chicos 3 puntos más). Se sabe que las diferencias son significativas porque el intervalo de confianza excluye una diferencia de 0.

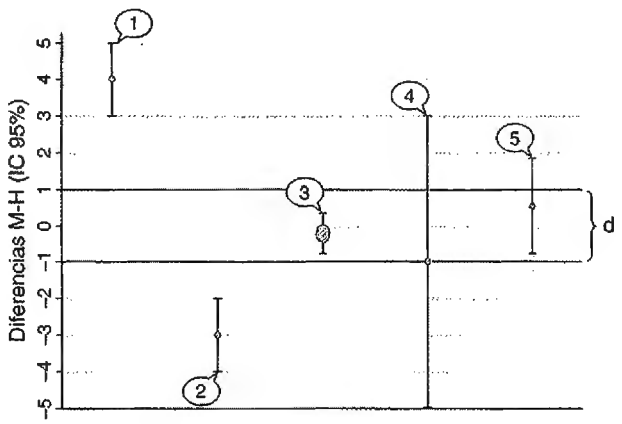



**Figura 4.14** Planteamiento de un ensayo de equivalencia. En un ensayo de equivalencia se fija primero una diferencia (*d* en la figura) que no llegue a ser *clínicamente* significativa, es decir, que sea tan pequeña que no tenga repercusiones importantes desde el punto de vista *práctico*. De los cinco ensayos realizados, el único que demostrará la igualdad será el 3. H, hombres; M, mujeres.

- El 4 no encuentra diferencias significativas entre A y B, pero no demuestra nada (tiene muy poca potencia, pues su IC 95% es amplísimo).
- En el 5, las chicas obtuvieron +0,6 puntos (IC 95%: –0,6 a +1,8), no encuentra diferencias significativas, pero tampoco demuestra la igualdad, ya que su intervalo de confianza al 95% cruza el valor d; por lo tanto, es compatible con una diferencia poblacional superior a d.

Además de ensayos de equivalencia, existen ensayos *de no inferioridad,* que buscan demostrar que un tratamiento es igual o superior a otro. En ese caso, el valor *d* solo tiene un límite, no dos. En el ejemplo anterior se podría hacer un test de no inferioridad de las notas de los chicos con respecto a las chicas (se pretende demostrar que los chicos *por lo menos no tienen notas inferiores).* Entonces, tanto el ensayo 2 (los chicos significativamente mejores) como el 3 demostrarían la *no inferioridad.*

## 4.21. PRUEBAS PARAMÉTRICAS Y NO PARAMÉTRICAS

Las pruebas *paramétricas* se basan en la media, la desviación estándar, etc., (parámetros), y requieren el cumplimiento de unas condiciones de aplicación más estrictas, ya que deben adoptar un modelo matemático correspondiente a una distribución conocida de probabilidad. Exigen, por ejemplo, la normalidad, homogeneidad de varianzas entre grupos u otros requisitos. Cuando las muestras son pequeñas y no se cumplen estos supuestos, o bien se trata de variables ordinales, se deben usar las pruebas *no paramétricas* o libres de distribución. Las pruebas no paramétricas solo tienen en cuenta el puesto *(rango)* o posición relativa que ocupa cada observación en el conjunto de observaciones ordenadas. Suelen utilizar una transformación de los datos originales en estos *rangos* o números de orden. Una vez que se asignan rangos, ya se usan siempre esos rangos, en vez de emplear directamente los valores de las observaciones recogidas. Esto permite realizar pruebas de contraste de hipótesis que no requieren que se asuma ninguna distribución teórica de probabilidad *(distribution-free).*

Los métodos *paramétricos* deben acompañarse de intervalos de confianza y los programas informáticos de estadística proporcionan tales intervalos con facilidad. En cambio, en programas de *software* habituales, las pruebas no paramétricas suelen carecer de intervalos de confianza.

## 4.22. RESUMEN DE LAS INSTRUCCIONES EN STATA, R, SPSS Y EXCEL

| Objetivo | STATA | R | SPSS | Excel |
|---|---|---|---|---|
| Selección aleatoria | **sample**<br>**sample 25, count** | sample(x, round (0.05*length(x)))<br>sample(x,25) | **COMP muest =**<br>**(UNIFORM(1)**<br>**<= .05).**<br>**FILTER BY muest.**<br>**EXE.** | |
| Asignación aleatoria | **gen *g3* = 1+ ///**<br>**floor**<br>**(3*uniform())** | sample(0:3, 100, replace = T) | | |
| Intervalo de confianza | **ci *obes*,**<br>**level(90)**<br>**cii 12036 2756** | | **EXAM VAR = dieta**<br>**/PLOT NONE/**<br>**CINTERVAL 90.** | |
| Distribución *t* para obtener *t* | **display invttail**<br>**(63,0.025)** | **qt(.025,63,**<br>**lower.tail = F)** | | =DISTR.<br>TINV(0,05;63) |
| Distribución *t* para obtener *p* | **display**<br>**ttail(63,2)** | **pt(–2,63)** | | =DISTRT<br>(2;63;1) |

## REFERENCIAS


1. Rothman KJ, Greenland S, Lash T. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
2. Sentís J, Pardell H, Cobo E, Canela J. Bioestadística. 3.ª ed. Barcelona: Masson; 2003.
3. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel; 2008.
4. Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health 2004;58(8):635-41.
5. De Irala J, Alonso A, Martínez-González MA. Conceptos básicos de epidemiología. En: Martínez-González MA, ed. Conceptos de salud pública y estrategias preventivas. Barcelona: Elsevier; 2013. p. 23-32.
6. Motulsky H. Intuitive Biostatistics. 2nd ed. Oxford: Oxford University Press; 2010.
7. Greenhalgh T. How to read a paper. Statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ 1997;315:364-6.
8. Olsen J, Saracci R, Trichopoulos D. Teaching epidemiology. 3rd ed. Oxford: Oxford University Press; 2010.
9. Rothman KJ. Epidemiology: An introduction. 2nd ed. Oxford: Oxford University Press; 2012.
10. Willett WC, Colditz GA. Approaches for conducting large cohort studies. Epidemiol Rev 1998;20:91-9.
11. Stang A, Jockel KH. Studies with low response proportions may be less biased than studies with high response proportions. Am J Epidemiol 2004;159:204-10.
12. Redmond C, Colton T. Biostatistics in clinical trials. New York: John Wiley & Sons Ltd; 2001.
13. Murie-Fernández M, Irimia P, Toledo E, Martínez E, Buil-Cosiales P, Serrano-Martínez M, et al. Carotid intima-media thickness changes with Mediterranean diet: a randomized trial (PREDIMED-Navarra). Atherosclerosis 2011;219:158-62.
14. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann Intern Med 1999;130(12):995-1004.
15. Altman DG. Practical statistics for medical research. Londres: Chapman and Hall; 1991.
16. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 1986;292(6522):746-50.
17. Gardner MJ, Altman DG, editors. Statistics with confidence. London: British Medical Journal Books; 1989.
18. Guyatt G, Jaenschke R, Heddel N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. II. Interpreting study results: confidence intervals. Can Med Assoc J 1995;152:169-73.
19. Martín Andrés A, Luna del Castillo JD. Bioestadística para las Ciencias de la Salud. 5.ª ed. Madrid: Norma-Capitel; 2004.
20. Gutiérrez-Fisac JL, Guallar-Castillón P, León-Muñoz LM, Graciani A, Banegas JR, Rodríguez-Artalejo F. Prevalence of general and abdominal obesity in the adult population of Spain, 2008-2010: the ENRICA study. Obes Rev 2012;13(4):388-92.
21. Clayton D, Hill M. Statistical models in epidemiology. Oxford: Oxford University Press; 1993.

22. Basterra-Gortari FJ, Beunza JJ, Bes-Rastrollo M, Toledo E, García-López M, Martínez-González MA. Increasing trend in the prevalence of morbid obesity in Spain: from 1.8 to 6.1 per thousand in 14 years. Rev Esp Cardiol 2011;64(5):424-6.

23. Martínez-González MA, Seguí-Gómez M, Delgado-Rodríguez M. ¿Cómo mejorar los intervalos de confianza? Med Clin (Barc) 2010;135(1):30-4.

24. Rothman KJ. A show of confidence. N Engl J Med 1978;299(24):1362-3.

25. Gigerenzer G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were afraid to ask. En: Kaplan D, editor. The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks: Sage; 2004. p. 391-408.

26. Cumming G. Replication and p intervals: p values predict the future only vaguely but confidence intervals do much better. Perspect Psychol Sci 2008;3(4):286-300.

27. Marewski JN, Olsson H. Beyond the Null Ritual. Formal Modeling of Psychological Processes. J Psychol 2009;217(1):49-60.

28. Ziliak ST, McCloskey DN. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor: University of Michigan Press; 2008.

29. Frank-Herrmann P, Heil J, Gnoth C, Toledo E, Baur S, Pyper C, et al. The effectiveness of a fertility awareness based method to avoid pregnancy in relation to a couple's sexual behaviour during the fertile time: a prospective longitudinal study. Hum Reprod 2007;22(5):1310-9.

30. Greenhalgh T. How to read a paper. Statistics for the non-statistician. II: "Significant" relations and their pitfalls. BMJ 1997;315(7105):422-5.

31. Rosner B. Fundamentals of Biostatistics. 7th ed. Boston: Brooks/Cole; 2011.

32. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350(9074):326-9.

33. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A. 2013;110(48):19313-7.

34. Ware JH, Mosteller F, Delgado F, Donnelly C, Ingelfinger JA. P Values. En: Bailar JC III, Hoaglin DC, editors. Medical uses of statistics. 3rd ed. New Jersey: John Wiley & Sons Inc; 2009. p. 175-94.

35. Rothman KJ. Causes. Am J Epidemiol 1976;104(6):587-92.

36. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annu Rev Public Health 2013;34(1):61-75.

37. Pearl J. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge: Cambridge University Press; 2009.

38. Hernan MA, Robins JM. Causal Inference. Boca Raton: CRC Press; 2014. (en prensa).

39. Mayo NE. Randomized trials and other parallel comparisons of treatment. En: Bailar JC III, Hoaglin DC, editors. Medical uses of statistics. 3rd ed. New Jersey: John Wiley & Sons Inc; 2009. p. 51-89.

40. Martínez-González MA, De Irala-Estévez J, Guillén-Grima F. ¿Qué es una odds ratio? Med Clin (Barc) 1999;112(11):416-22.

41. Martínez-González MA, Alonso A, López Fidalgo J. ¿Qué es una hazard ratio? Med Clin (Barc) 2008;131(2):65-72.

42. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol 2010;63(1):28-36.

43. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994;309(6949):248.

44. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348-53.

# DATOS CATEGÓRICOS Y PORCENTAJES: COMPARACIÓN DE PROPORCIONES

5

*E. Toledo, J. M. Núñez-Córdoba, M. Á. Martínez-González*

Las decisiones sobre el tipo de análisis estadístico más adecuado para cada diseño de investigación dependen, en primer lugar, de la naturaleza de los datos que se hayan recogido (1). Para describir y resumir la información contenida en variables categóricas se suelen usar *proporciones* o porcentajes. Una proporción presenta un porcentaje como un tanto por uno. Habitualmente se presentan *porcentajes* en la literatura biomédica. Por ejemplo, si se ha recogido el estado civil, lo más adecuado para describirlo sería presentar el porcentaje de solteros, casados, viudos, etc. Para su tratamiento estadístico hay que convertirlos en proporciones.

## 5.1. TEST DE $\chi^2$ DE PEARSON PARA DOS PROPORCIONES

La ji cuadrado ($\chi^2$) de Pearson es una prueba estadística de contraste de hipótesis que se aplica para analizar datos recogidos en forma de número de observaciones en cada categoría: número de éxitos que ha tenido una intervención, porcentaje de pacientes que presentan una característica, proporción de resultados favorables obtenidos en dos grupos de pacientes con tratamientos distintos, etc. En definitiva, sirve para analizar variables *cualitativas o categóricas* y para comparar proporciones (porcentajes). Pero esta prueba tiene una limitación, y es que requiere un número suficiente de observaciones para que sea válida.

Por ejemplo, un estudio investigó si comer helado muy deprisa daba lugar a un dolor de cabeza (cefalea) con más frecuencia que comer helado despacio (2). Para ello se asignó a 145 estudiantes de manera aleatoria para tomar 100 ml de helado en menos de 5 s (aceleradamente) o en más de 30 s (pausadamente), y se registró la aparición de cefalea. Los resultados obtenidos se muestran en la tabla 5.1.

De 73 personas que habían tomado el helado aceleradamente, 20 de ellas (27%) desarrollaron cefalea. En cambio, entre quienes lo habían tomado despacio, solo 9 (13%) la desarrollaron. Este tipo de problemas suele exigir que se aplique la $\chi^2$ de Pearson. Se deben dar los siguientes pasos:

1. *Plantear la hipótesis nula y la hipótesis alternativa del estudio.* La hipótesis nula consistiría en que la proporción de personas que desarrollan cefalea será la misma en los dos grupos, los que lo tomaron aceleradamente y los que lo tomaron con pausa. Es decir, que la cefalea es independiente de la velocidad a la que se consume el helado. La hipótesis alternativa sería que las proporciones de desarrollo de cefalea son distintas entre los acelerados y los pausados. Si $\pi$ es la proporción de los que desarrollan cefalea (a nivel de la población), se formularían así las hipótesis:

$$\text{Hipótesis nula } (H_0) \equiv \pi_{\text{comen aceleradamente}} = \pi_{\text{comen precavidamente}}$$

$$\text{Hipótesis alternativa } (H_1) \equiv \pi_{\text{comen aceleradamente}} \neq \pi_{\text{comen precavidamente}}$$

2. *Construir una tabla 2 × 2 con valores observados como los que aparecen en la tabla 5.1.* Como hay dos variables (velocidad de consumo del helado y cefalea) y cada una tiene dos categorías, resultan cuatro casillas al *cruzar* las categorías de ambas variables. La tabla quedaría con dos filas y dos columnas (además de los totales o *marginales*).

***Tabla 5.1 Resultados obtenidos para la asociación entre la velocidad a la que se come helado y el riesgo de desarrollar cefalea***

| MODO EN QUE COMEN HELADO | APARICIÓN DE CEFALEA | | TOTAL |
|---|---|---|---|
| | SÍ | NO | |
| Aceleradamente | 20 | 53 | 73 |
| Precavidamente | 9 | 63 | 72 |
| Total | 29 | 116 | 145 |

3. *Calcular los valores esperados.* Para ello, se parte de la tabla, pero solo con los valores totales (o *marginales*) de filas y las columnas. Se obtiene el esperado al multiplicar el total de la fila por el total de la columna y dividirlo por el total de los totales:

$$\text{Esperados} = \frac{(\text{total fila}) \times (\text{total columna})}{\text{total de los totales}}$$

Por ejemplo, los esperados para la casilla de los que comen con pausa y desarrollan cefalea serían (tabla 5.2):

$$\text{Esperados}_{\text{casos expuestos}} = \frac{72 \times 29}{145} = 14{,}4$$

Así se puede calcular los esperados para cada casilla, pero es más simple completarlos con sumas y restas, pues se deben mantener fijos los marginales.

4. *Aplicar la ecuación de la ji cuadrado:*

$$\chi^2 = \sum\left(\frac{(\text{obs} - \text{esp})^2}{\text{esp}}\right)$$

Como hay cuatro casillas, tendrá cuatro términos, uno por cada casilla.

$$\chi^2 = \sum\left(\frac{(\text{obs} - \text{esp})^2}{\text{esp}}\right) = \frac{(20 - 14{,}6)^2}{14{,}6} + \frac{(9 - 14{,}4)^2}{14{,}4} + \frac{(53 - 58{,}4)^2}{58{,}4} + \frac{(63 - 57{,}6)^2}{57{,}6}$$
$$= 2{,}00 + 2{,}03 + 0{,}50 + 0{,}51 = 5{,}028$$

5. *Calcular los grados de libertad.* Para una tabla con dos filas y dos columnas (tabla «2 × 2»), una vez fijados los cuatro marginales, en cuanto se introduce un solo valor en alguna de las casillas interiores ya quedan determinados el resto de valores, por lo que, de una forma intuitiva, ya se entiende que tiene un grado de libertad. En general, para una tabla con c columnas y f filas, los grados de libertad son:

$$\text{gl} = (\text{columnas} - 1) \times (\text{filas} - 1)$$

En el ejemplo: gl = (2 – 1) (2 – 1) = 1

***Tabla 5.2 Valores esperados para la asociación entre la velocidad a la que se come helado y el riesgo de desarrollar cefalea***

| ESPERADOS | CEFALEA SÍ | CEFALEA NO | TOTAL |
|---|---|---|---|
| Aceleradamente | 29 – 14,4 = 14,6 | 73 – 14,6 = 58,4 | 73 |
| Precavidamente | 72 × 29/145 = 14,4 | 72 – 14,4 = 57,6 | 72 |
| Total | 29 | 116 | 145 |

6. *Comparar la ji cuadrado encontrada con los valores de tablas (u ordenador) y obtener así el valor* p. Podría compararse la $\chi^2$ encontrada con la recogida en las tablas para un grado de libertad o introducir en STATA o Excel el valor de $\chi^2$ encontrado indicando que tiene un grado de libertad. La distribución de la $\chi^2$ es una distribución de frecuencias que se caracteriza por unir en una sola cola las dos colas de otras distribuciones (normal o *t* de Student). Por ello, los valores *p* de una $\chi^2$ son siempre bilaterales. Los valores tabulados (los que aparecen en las tablas) para la $\chi^2$ con un grado de libertad corresponden a los valores *p* de la fila inferior:

| $\chi^2$ | 2,706 | 3,841 | 5,024 | 6,635 | 7,879 | 10,828 |
|---|---|---|---|---|---|---|
| valor *p* | 0,10 | 0,05 | 0,025 | 0,01 | 0,005 | 0,001 |

Si el valor encontrado supera al de las tablas, el valor *p* será inferior al mostrado; por ejemplo, si $\chi^2 = 5{,}028$, entonces $p < 0{,}025$.

Es preferible, en vez de tablas, usar Excel con la expresión =DISTR.CHI(4;1), que devolverá $p = 0{,}0249$, lo mismo que hará STATA con la orden *di chi2tail(1,4)*.

7. *Conclusión.* Como el valor calculado para la $\chi^2$ (5,028) corresponde a un valor $p < 0{,}05$, se puede afirmar que la diferencia es estadísticamente significativa y hay evidencias para rechazar la hipótesis nula, con un valor $p = 0{,}0249$.

   Si se hubiese encontrado otro valor, por ejemplo 3,5, no se podría decir que la $\chi^2$ era significativa (aunque sí lo sería si se asumiese otro error $\alpha$, por ejemplo, $\alpha = 0{,}10$).

   Como conclusión del ejemplo, puede afirmarse que existe evidencia de una asociación estadísticamente significativa ($p < 0{,}05$) entre comer helado aceleradamente y el riesgo de desarrollar cefalea, ya que sería muy extraño que las diferencias encontradas fuesen simplemente explicadas por el azar.

8. *Verificar las condiciones de aplicación.* Antes de dar por válido el resultado, hay que comprobar los requisitos de aplicación de la $\chi^2$ para la comparación de dos proporciones:
   a. Ambas variables deben ser cualitativas en escala *nominal*. No se deben aplicar cuando la variable dependiente es ordinal.
   b. Ninguno de los valores esperados debe ser menor de 5. En este caso, todos los valores esperados son $\geq 14{,}4$.

La $\chi^2$ puede extenderse a otro ejemplo con más categorías. Por ejemplo, si se desease comparar el nivel de satisfacción (alto/bajo) en pacientes de tres centros de salud diferentes (A, B y C), los datos resultantes se ordenarían en una tabla de contingencia que tendría dos filas (una para cada nivel de satisfacción) y tres columnas (una para cada centro de salud). Resultaría así una tabla $3 \times 2$. Para comparar si existen diferencias significativas en el porcentaje de personas con satisfacción alta entre los tres centros de salud, se emplearía un test de la $\chi^2$. La fórmula sería la misma que para la comparación de dos variables dicotómicas. En este caso, la fórmula de la $\chi^2$ tendrá seis sumandos, uno para cada celda de la tabla $3 \times 2$, y los grados de libertad se obtendrán de (filas – 1) × (columnas – 1) = (2 – 1) × (3 – 1) = 2. En cuanto a los requisitos de aplicación, en tablas que tengan más de cuatro casillas, al menos el 80% de los valores esperados deben ser superiores o iguales a 5.

## 5.2. TEST Z PARA COMPARAR DOS PROPORCIONES

Una alternativa al test de la $\chi^2$ para comparar dos proporciones es usar un test *z*. La formulación de la hipótesis nula y de la hipótesis alternativa será la misma que en el test de la $\chi^2$. El test *z* se fundamenta en el cociente que resulta de dividir un efecto entre un error. En este caso, el *efecto*

será la diferencia entre las dos proporciones, y el *error* será el error estándar de la diferencia de proporciones (EEDP). La varianza de una diferencia es igual a la suma de las varianzas de cada parte de la diferencia. Por tanto, el error estándar de una diferencia será la raíz cuadrada de la suma de varianzas divididas, cada una, por el tamaño de la respectiva muestra ($n_1$, $n_2$), pero se usa $p$, que es la proporción global (considerando conjuntamente los dos grupos como si fuesen uno solo), y *no* las proporciones ($p_1$ y $p_2$) particulares de cada grupo. Así:

$$z = \frac{\text{efecto}}{\text{error}} = \frac{\text{diferencia de proporciones}}{\text{EEDP}} = \frac{p_1 - p_2}{\sqrt{\frac{p \times q}{n_1} + \frac{p \times q}{n_2}}}$$

donde $p_1$ es la proporción de eventos (en el ejemplo, cefaleas) observada en un grupo; $p_2$ es la proporción de eventos en el otro grupo; $p$ es la proporción total (o *marginal*) para ambos grupo juntos; $q$ es el complementario de $p$; $n_1$ es el número de sujetos en el primer grupo, y $n_2$ es el número de sujetos en el otro grupo.

En el ejemplo de la velocidad a la que se consume el helado (v. tabla 5.1):

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p \times q}{n_1} + \frac{p \times q}{n_2}}} = \frac{0{,}274 - 0{,}125}{\sqrt{\frac{0{,}2 \times 0{,}8}{73} + \frac{0{,}2 \times 0{,}8}{72}}} = \frac{0{,}149}{0{,}066} = 2{,}243$$

La proporción global se ha obtenido dividiendo el total de sujetos que desarrollan cefalea entre el total de participantes (29/145).

Este valor de $z$ (2,243) es superior al valor tabulado para un nivel de significación estadística del 5% a dos colas, que es 1,96. Se concluirá, por tanto, que existen diferencias estadísticamente significativas en la proporción de quienes desarrollan cefalea según se tome el helado rápida o lentamente. Por ello, se puede rechazar la hipótesis nula que mantiene que la proporción de sujetos que desarrollan cefalea es igual entre ambos tipos de consumo de helado y se optará por la hipótesis alternativa.

Para obtener el valor $p$ concreto que se corresponde con una $z$ de 2,243 se puede recurrir a Excel con la orden =2*DISTR.NORM.ESTAND(–2,243) o a STATA mediante la orden *di 2*normal(–2.243)*, y se obtendrá un valor $p$ de 0,0249, que coincide con el que antes se había obtenido mediante el test de la $\chi^2$.

El valor de la $z$ obtenido (2,243) coincide, además, con el valor de la raíz cuadrada de la $\chi^2$.

$$z^2 = \chi^2 \text{ (1 gl)}$$
$$2{,}243^2 = 5{,}028$$

Esto se debe a que una distribución $\chi^2$ con un grado de libertad coincide con el valor $z$ de la distribución normal elevado al cuadrado.

## 5.3. INTERVALO DE CONFIANZA DE LA DIFERENCIA DE DOS PROPORCIONES

En el ejemplo del helado existía una diferencia de un 14,9% en el porcentaje de desarrollo de cefalea entre quienes lo tomaban aceleradamente y quienes lo hacían con pausa: $p_1 - p_2 = 0{,}274 - 0{,}125 = 0{,}149$.

El valor $p$ calculado mediante el test de la $\chi^2$ o de la $z$ ha contestado a la pregunta sobre la significación estadística de esta diferencia. No obstante, se puede plantear otra pregunta referente a la *magnitud del efecto*. Esto es importante y se puede formular de muchos modos: ¿en cuánto aumenta el riesgo de cefalea al consumir aceleradamente un helado? Es decir, ¿cómo estimar la verdadera diferencia en el riesgo (o proporción) de cefaleas existente en la población entre los

dos modos de consumo (rápido y lento)?, ¿qué valor tiene esta diferencia de proporciones a nivel poblacional?, ¿cuál es la diferencia real de proporciones en la población de la que procede la muestra? Para contestar a esta pregunta se deberá calcular un intervalo de confianza a la diferencia de proporciones observada en la muestra. Los pasos que han de seguirse serían:

1. *Calcular la diferencia absoluta entre las proporciones de la muestra* $|d|$:

$$|d| = p^1 - p^2$$

$$= 0,274 - 0,125 = 0,149$$

2. *Mirar en las tablas de la normal (o extraer de Excel o STATA) el valor z correspondiente al error alfa (bilateral,* $z_{\alpha/2}$*)* del intervalo de confianza que se quiera calcular. Por ejemplo, para un intervalo de confianza al 95%, $z_{\alpha/2}$ sería 1,96.
3. *Calcular el error estándar de la diferencia de proporciones (EEDP).*

$$\text{EEDP} = \sqrt{\frac{p \times q}{n_1} + \frac{p \times q}{n_2}} = \sqrt{\frac{0,2 \times 0,8}{73} + \frac{0,2 \times 0,8}{72}} = 0,066$$

4. *Calcular el intervalo de confianza de la diferencia de proporciones.* Basta con sumar y restar z veces el EEDP a la diferencia absoluta:

$$\text{IC } 1-\alpha \text{ para } \Delta : |d| \pm z_{\alpha/2} \text{xEEDP}$$

$$\text{IC 95\% para } \Delta : |d| \pm 1,96 \times \text{EEDP} = 0,149 \pm 1,96 \times 0,066 = 0,020 \text{ a } 0,278$$

5. *Conclusión.* Con un 95% de confianza se puede afirmar que la diferencia verdadera en el riesgo de cefalea en la población estará entre el 2,0% y el 27,8%. Consumir rápidamente el helado incrementa el riesgo o proporción de desarrollar cefalea entre un 2,0 y un 27,8 con respecto a quienes lo consumen pausadamente. Si el intervalo de confianza tuviese uno de sus límites negativos, equivaldría a afirmar que comer helado aceleradamente puede tanto aumentar como disminuir el riesgo de desarrollar cefalea; en esa situación se diría que el intervalo de confianza incluía al valor nulo ($\Delta = 0\%$), y esto equivaldría a decir que no existirían diferencias significativas (significativas al 5%, si el intervalo de confianza es al 95%). Pero, en el ejemplo del helado, ambos límites de confianza están en el mismo lado (tanto 0,020 como 0,278 son cifras positivas) y, por tanto, puede afirmarse que con una confianza del 95% todo el efecto apunta hacia un mayor riesgo si el consumo es rápido (equivale a decir que existen diferencias significativas entre quienes consumen helado aceleradamente y quienes lo hacen precavidamente).
6. *Verificar los criterios de aplicación.* Se puede confiar en que se cumplen las condiciones de aplicación si ninguno de los esperados es menor de 5. Esto sucede casi siempre que la muestra de cada grupo es grande ($n_1 > 60$ y $n_2 > 60$). En este ejemplo, ya se había calculado que los esperados eran $\geq 14,4$.

## 5.4. RELACIÓN ENTRE EL INTERVALO DE CONFIANZA Y EL VALOR *P*

La expresión que se acaba de utilizar es:

$$\text{IC } (1-\alpha) \text{ para } \Delta : |d| \pm z_{\alpha/2} \times \text{EEDP}$$

En un contraste de hipótesis por el test de la $\chi^2$ de Pearson, la hipótesis nula puede formularse como:

$$H_0 \equiv \pi_{\text{cosumo acelerado}} = \pi_{\text{consumo pausado}}$$

Pero ya que Δ es la diferencia entre estas proporciones, la hipótesis nula también pudiera formularse así:

$$H_0 \equiv \Delta = \pi_{\text{consumo acelerado}} - \pi_{\text{consumo pausado}} = 0$$

Si Δ vale 0, entonces:

$$|d| = z_{\alpha/2} \times \text{EEDP} \quad \text{y} \quad z_{\alpha/2} = \frac{|d|}{\text{EEDP}}$$

En los apartados 5.1 y 5.2 ya se había calculado que el valor de la $\chi^2$ era 5,028 y el de la $z$, su raíz cuadrada, 2,243.

¿Qué pasaría si el cociente $z$ entre |d| y el EEDP resultara ser exactamente de 1,96? Sucederían dos cosas:

- Que la significación estadística (a dos colas) sería exactamente $p = 0{,}05$.
- Que uno de los límites del intervalo de confianza al 95% estaría exactamente en 0.

Cuando el intervalo de confianza al 95% para la diferencia entre dos proporciones abarca el 0 (lo incluye o contiene entre sus límites), entonces las diferencias no son estadísticamente significativas al 5%.

## 5.5. JI CUADRADO PARA COMPARAR UNA PROPORCIÓN CON UNA REFERENCIA EXTERNA (ESPERADA): CÁLCULO Y SU RELACIÓN CON LA DISTRIBUCIÓN BINOMIAL Y SUS APROXIMACIONES

Esta utilidad de la $\chi^2$ se aplica menos. Se trata de hacer una comparación, dentro de *una sola variable* categórica, para comprobar si la distribución observada se corresponde con una distribución teóricamente esperada.

Se entenderá con un ejemplo. Imagínese que solo hay una variable, sexo, que puede tomar dos categorías (hombre o mujer), y que en una clase con un total de 20 universitarios, el 45% (9) son mujeres y el 55% (11) hombres. Se desea comprobar si esta distribución es consistente con que, en la universidad («población» de la que procede la muestra), el 50% de los alumnos sean hombres y el 50% mujeres ($H_0$). En cambio, $H_1$ a una cola mantendría que realmente existe un mayor número de hombres que de mujeres en toda la universidad. Este problema se podría haber resuelto de manera exacta por la distribución binomial para obtener la $p(k \leq 9)$ (fig. 5.1). Si se emplease un programa informático, habría que introducir en Excel =DISTR.BINOM(9;20;0,5;VERDADERO) o en STATA *di binomial(20,9,0.5)*. En ambos casos se obtendría $p(k \leq 9) = 0{,}412$. Existe una probabilidad del 41,2% de encontrar hasta 9 mujeres (se han sumado las probabilidades desde 0 hasta 9) en una muestra de tamaño 20, suponiendo que $H_0$ (el 50% varones; el 50% mujeres) fuese cierta. Es decir, si se extrajesen muestras de tamaño 20 de una población donde el 50% son varones, en el 41,2% de estas muestras se hallarían 9 o menos mujeres. Este valor corresponde a la $p$ de la cola de la izquierda, pues es exactamente la probabilidad de encontrar lo observado (9 mujeres) o todo lo que esté más lejano de $H_0$ (de 0 a 8 mujeres) en la muestra, si $H_0$ fuese verdadera ($\pi = 0{,}5$).

En el caso de la $\chi^2$, siempre hay que concebir $H_1$ como bilateral (también incluiría como alternativa que en la población hubiese *más mujeres* que hombres). La cola de la derecha correspondería a la probabilidad de encontrar más de 11 mujeres si $H_0$ fuese cierta. Cuando se usa la distribución binomial para hacer un test a dos colas, en la cola correspondiente a la hipótesis alternativa más lejana a lo observado no se incluye la probabilidad de observar un resultado exactamente igual de lejano a $H_0$ que el observado (11 mujeres), sino solo los valores más lejanos de $H_0$ que lo observado (12 o más mujeres). Con Excel (=1-DISTR.BINOM(11;20;0,5;VERDADERO)) o

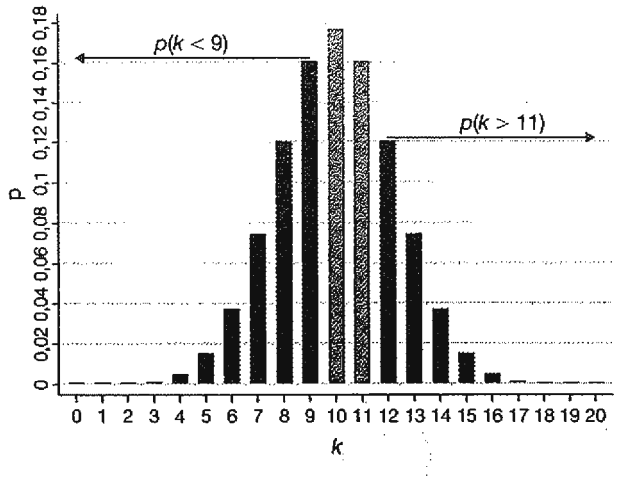



**Figura 5.1** Probabilidad de encontrar $k$ mujeres con la distribución binomial con $n = 20$ y $\pi = 0,5$.

STATA (di binomialtail(20,12,0.5)) se puede obtener este valor ($p(k > 11) = 0,252$) (v. fig. 5.1). El valor $p$ a dos colas se calcularía según la siguiente expresión:

$$\text{Valor } p \text{ (dos colas)} = p(k \leq 9 \mid \pi = 0,5) + p(k > 11 \mid \pi = 0,5) = 0,412 + 0,252 = 0,664$$

No se dispone de evidencias para rechazar la hipótesis nula.

Este ejemplo se puede resolver por la distribución binomial con calculadora, aunque requeriría cierto tiempo:

$$\text{Valor } p \text{ (dos colas)} = 1 - [p(k = 10 \mid \pi = 0,5) + p(k = 11 \mid \pi = 0,5)]$$

En esta última fórmula es interesante fijarse bien en el 1 que hay justamente a la derecha del signo igual. En cualquier caso, es preferible hacer esto con ordenador. La ventaja de usar la distribución binomial es que siempre dará resultados exactos.

Este ejemplo también se puede resolver (aproximadamente) por la $\chi^2$ de Pearson. Los pasos serían:

1. *Plantear la hipótesis de estudio:*

$$H_0 \equiv \pi_{esperada} = p_{observada}$$

$$H_0 \equiv \pi_{esperada} \neq p_{observada}$$

   La proporción esperada (que no tiene por qué ser siempre del 50% para cada categoría en una variable dicotómica) en este ejemplo sí sería que el número de sujetos de cada sexo es el mismo y, por tanto, $\pi_{esperada} = 0,5$.
2. *Calcular los esperados.* Sería una proporción del 50% para cada grupo; por tanto, se esperaría encontrar 10 individuos de cada sexo.
3. *Comprobar que se cumplen los requisitos de aplicación.* La variable ha de ser cualitativa nominal y todos los esperados deben ser >5. En el ejemplo se cumple (10 > 5).
4. *Calcular el valor de* $\chi^2$.

$$\chi^2 = \sum\left(\frac{(\text{Obs} - \text{Esp})^2}{\text{Esp}}\right) = \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} = 0,1 + 0,1 = 0,2$$

5. *Calcular los grados de libertad.* Como hay dos categorías ($k = 2$) y los grados de libertad son ahora $k - 1$, habrá un grado de libertad. El número de grados de libertad se deduce fijando el total de las observaciones (son 20 en total en la muestra) y viendo cuántas categorías se pueden fijar arbitrariamente. En el ejemplo, una vez que se sabe que el total de la muestra son 20 y que hay 11 varones, ya no queda libertad para poner el número de mujeres que se quiera; necesariamente tienen que ser 9 para que el total sea 20. Por eso solo hay un grado de libertad.
6. *Mirar en la tabla de la $\chi^2$* si un valor de 0,2 es significativo para un grado de libertad. Para un grado de libertad, los valores de $\chi^2$ (v. tablas al final del libro) son:

| | | | |
|---|---|---|---|
| $\chi^2$ | 2,706 | 3,841 | 5,024 |
| $p$ | 0,10 | 0,05 | 0,025 |

Como el valor hallado (0,2) es inferior a 2,706, se puede afirmar que la prueba de $\chi^2$ ha resultado no significativa ($p > 0,10$).

Para obtener el valor $p$ concreto se puede consultar Excel (=DISTR.CHI(0,2;1)) o STATA (*di chi2tail(1,0.2)*), donde se obtendrá un valor $p = 0,655$.

7. *Conclusión.* A la vista del valor encontrado en el ejemplo para $\chi^2$ y de los valores de la tabla, hay que tomar la decisión sobre el rechazo o no rechazo de la hipótesis nula. En este ejemplo ($p > 0,10$), no hay evidencia para rechazar la hipótesis nula. En conclusión, no se puede rechazar que la muestra proceda de una población donde un 50% son mujeres y el otro 50% son hombres.

En este ejemplo se cumple que $n \times \pi > 5$, por lo que se podría resolver este ejemplo también mediante la aproximación a la normal, como se vio en el capítulo 3. Para calcular el valor de $z$ en el caso de una variable cuantitativa, se disponía de la fórmula:

$$z = \frac{x - \mu}{\sigma^2}$$

Si se remplaza $\mu$ por $n\pi$ y $\sigma^2$ por $n\pi(1 - \pi)$, la expresión anterior de $z$ para aproximarse la distribución binomial mediante la normal para una variable cualitativa nominal era:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

Sustituyendo los datos con los del ejemplo:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{9 - 20 \times 0,5}{\sqrt{20 \times 0,5 \times (1-0,5)}} = \frac{-1}{2,24} = -0,446$$

Si se consulta el valor $p$ correspondiente a $z = -0,446$ a dos colas, bien en Excel (=2*DISTR.NORM.ESTAND(–0,446)) o bien en STATA (*di 2*normal*(–0.446)), se obtendrá un valor $p = 0,655$, el mismo que con el test de la $\chi^2$ para una proporción.

Como se ha comentado anteriormente, la $\chi^2$ con un grado de libertad coincide con $z$ al cuadrado:$(-0.446)^2 = 0,2$.

## 5.6. TEST EXACTO DE FISHER

En un ejemplo ficticio, se realiza un estudio para prevenir las caídas en ancianos. Un grupo de 19 ancianos se asigna a una determinada intervención que trata de evitar las caídas, y el otro grupo de 11 sirve como control. Se obtienen los datos que se muestran en la tabla 5.3.

*Tabla 5.3 Resultados obtenidos en el estudio para prevenir caídas en ancianos*

| GRUPO | NO SE CAEN | SE CAEN | TOTAL |
|---|---|---|---|
| Intervención | 14 | 5 | 19 |
| Control | 5 | 6 | 11 |
| Total | 19 | 11 | 30 |

Si se plantease resolver este ejemplo con el test de la $\chi^2$ se vería que no se cumple el requisito de aplicación de que todos los esperados sean >5, ya que en la celda de controles que se caen habría 11 × 11/30 = 4,03 esperados. El test exacto de Fisher contrasta la asociación entre dos variables cualitativas nominales (tablas 2 × 2, como la $\chi^2$). La ventaja que tiene es que *no* exige cumplir este requisito de aplicación. Puede emplearse con independencia del número de esperados que haya en las celdas.

Para entender el test exacto de Fisher hay que plantearse en primer lugar lo siguiente: ¿cuáles son las distintas posibilidades de que se caigan 5 de los 19 que forman el grupo de intervención? La primera caída que se produce en el grupo de intervención podría haber afectado a cada uno de los 19 que formaban el grupo; por consiguiente, hay 19 formas diferentes de que se produzca la primera caída en ese grupo. Cuando ya hay uno que ha caído, la segunda caída podría haber afectado a 18, luego hay 18 formas diferentes de que se produzca, y así sucesivamente. Por tanto, hay 19 × 18 × 17 × 16 × 15 formas posibles de que se produzcan las 5 caídas en el grupo de intervención. Ese producto da como resultado 1.395.360 formas de producirse las 5 caídas. En muchas de estas posibilidades caerían las mismas cinco personas, pero en un orden distinto. Por ejemplo, una posibilidad sería que las caídas afectaran a los individuos 1.°, 2.°, 3.°, 4.° y 5.°, exactamente en este orden. Pero, si no importase el orden en que se han producido las caídas, sino solo interesase saber quiénes son los 5 que se caen de entre los 19 candidatos a caerse, habrá que dividir el producto antes obtenido entre las distintas permutaciones de 5 individuos, que valen 5!

$$\text{Posibilidades} = \frac{19\times18\times17\times16\times15}{5\times4\times3\times2\times1} = 11.628$$

La forma matemática de expresar lo anterior corresponde a las combinaciones de 19 elementos tomados de 5 en 5, y se expresa como:

$$\binom{19}{5} = \frac{19!}{(19-5)!5!}$$

En general, puede afirmarse que:

Combinaciones de $n$ elementos tomados de k en $k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$

En el ejemplo, las posibilidades de que se produzcan 5 caídas entre 19 individuos del grupo de intervención son:

$$\text{Posibilidades}_{\text{intervención}} = \binom{n}{k} = \frac{n!}{(n-k)!k!} = \binom{19}{5} = \frac{19!}{14!5!} = 11.628$$

Para aplicar esto a nuestro problema, que buscaba responder la pregunta de si hay diferencias entre la proporción de los que se han caído en el grupo de intervención y en el grupo control, habrá que preguntarse también de cuántas formas se pueden producir 11 caídas en un total de 30 individuos. La respuesta es:

$$\text{Posibilidades}_{\text{totales}} = \binom{n}{k} = \frac{n!}{(n-k)!k!} = \binom{30}{11} = \frac{30!}{19!11!} = 54.627.300$$

Hay, por tanto, unos 54 millones y medio de formas de que, entre 30 personas, se caigan 11. Pues bien, de esos 54 millones y medio solo algunas posibilidades coinciden con el hecho de que, de los 11 que se caen, 6 sean de un grupo compuesto por 11 individuos y 5 de otro grupo compuesto por 19 individuos. Una de estas últimas posibilidades ya se había estimado (las posibles formas de caerse 5 personas de un grupo de 19 eran 11.628). Las otras (que se caigan 6 en un grupo de 11) serán:

$$\text{Posibilidades}_{\text{control}} = \binom{11}{6} = \frac{11!}{5!\,6!} = 462$$

Por tanto, las posibilidades de que ocurran las cosas como aparecían en nuestra tabla serán 462 veces 11.628.

$$\text{Posibilidades}_{\text{observadas}} = \text{posibilidades}_{\text{control}} \times \text{posibilidades}_{\text{intervención}} = 462 \times 11.628 = 5.372.136$$

No llegan a 5,4 millones. Como había un total de 54,6 millones de formas de que se produjera, la probabilidad de nuestra tabla es aproximadamente del 10%; es decir, los datos observados representan un ≈10% de las posibilidades totales en que pueden ocurrir las cosas.

$$P_{\text{TABLA}} = \frac{\text{posibilidades}_{\text{observadas}}}{\text{posibilidades}_{\text{totales}}} = \frac{\binom{19}{5}\binom{11}{6}}{\binom{30}{11}} = \frac{5.372.136}{54.627.300} = 0,0983$$

Una vez visto esto, los pasos que han de seguirse para realizar el test exacto de Fisher son:

1. Plantear la hipótesis nula y la hipótesis alternativa. Son equivalentes al test de la $\chi^2$ de Pearson:

$$H_0 \equiv \pi_{\text{intervención}} = \pi_{\text{control}}$$

$$H_1 \equiv \pi_{\text{intervención}} \neq \pi_{\text{control}}$$

2. *Calcular las combinaciones totales.* Con los totales («marginales») de la tabla se calculan las combinaciones posibles del total ($n_{\text{TOTAL}}$) de elementos tomados en grupos de $k_{\text{TOTAL}}$ en $k_{\text{TOTAL}}$.

$$\text{Posibilidades}_{\text{totales}} = \binom{n_{\text{TOTAL}}}{k_{\text{TOTAL}}}$$

3. *Calcular las combinaciones observadas en cada grupo.* Con el total de cada grupo de la tabla se calculan las combinaciones posibles de los elementos de ese grupo ($n_i$) tomados en grupos de $k_i$ en $k_i$. Esto se hace para los dos grupos.

$$\textit{Posibilidades}_{\textit{grupo}_i} = \binom{n_i}{k_i}$$

4. *Dividir el producto de las combinaciones observadas por las combinaciones totales.* Se divide el producto de las combinaciones posibles en ambos grupos, por las combinaciones totales posibles obtenidas a partir de los marginales.

$$p(\text{tabla}) = \frac{\text{posib}_{\text{observadas}}}{\text{posib}_{\text{totales}}} = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_{\text{TOTAL}}}{k_{\text{TOTAL}}}} = \frac{\binom{19}{5}\binom{11}{6}}{\binom{30}{11}} = \frac{\left(\frac{19!}{14!5!}\right)\left(\frac{11!}{5!6!}\right)}{\left(\frac{30!}{19!11!}\right)} = \frac{5.372.136}{54.627.300} = 0,0983$$

*Tabla 5.4 Resultados anotados obtenidos en el estudio para prevenir caídas en ancianos*

| GRUPO | NO SE CAEN | SE CAEN | | TOTAL | |
|---|---|---|---|---|---|
| Intervención | 14 | 5 | $k_1$ | 19 | $n_1$ |
| Control | 5 | 6 | $k_2$ | 11 | $n_2$ |
| Total | 19 | 11 | $k_{TOTAL}$ | 30 | $n_{TOTAL}$ |

5. *Repetir el proceso para todas las tablas posibles más alejadas de la hipótesis nula que la observada.* Hay que considerar que el valor *p* no es la probabilidad de *una tabla,* sino también de todos los sucesos más extremos que podían haber ocurrido. Hay que considerar también las tablas más extremas que la tabla 5.3, respetando los marginales, como se muestra en la tabla 5.4. Véase también la tabla 5.5. Ya no hay más posibilidades hacia ese lado.
6. *Sumar las probabilidades de todas esas tablas más las de la tabla observada.* Al final, el valor de *p* para la comparación de proporciones será la suma de los valores *p* de todas las posibles tablas iguales o más alejadas de la hipótesis nula que la encontrada. Esto dará el valor *p* a una cola.

$$P_{\text{Fisher una cola}} = 0,0983 + 0,0234 + \ldots + 0,0000000183 = 0,1248.$$

Si se desea a dos colas, hay que repetir el proceso para todas las posibles tablas en la otra cola a partir de la tabla que tenga una probabilidad igual o inferior a la observada.

*Tabla 5.5 Tablas más extremas a la observada en el ejemplo de la prevención de caídas en ancianos*

| | NO SE CAEN | SE CAEN | TOTAL | CÁLCULO | VALOR *P* |
|---|---|---|---|---|---|
| **Intervención** | 15 | 4 | 19 | $\frac{\binom{19}{4}\binom{11}{7}}{\binom{30}{11}}$ | 0,0234 |
| **Control** | 4 | 7 | 11 | | |
| **Total** | 19 | 11 | 30 | | |

| | NO SE CAEN | SE CAEN | TOTAL | CÁLCULO | VALOR *P* |
|---|---|---|---|---|---|
| **Intervención** | 16 | 3 | 19 | $\frac{\binom{19}{3}\binom{11}{8}}{\binom{30}{11}}$ | 0,00293 |
| **Control** | 3 | 8 | 11 | | |
| **Total** | 19 | 11 | 30 | | |

| | NO SE CAEN | SE CAEN | TOTAL | CÁLCULO | VALOR *P* |
|---|---|---|---|---|---|
| **Intervención** | 17 | 2 | 19 | $\frac{\binom{19}{2}\binom{11}{9}}{\binom{30}{11}}$ | 0,000172 |
| **Control** | 2 | 9 | 11 | | |
| **Total** | 19 | 11 | 30 | | |

| | NO SE CAEN | SE CAEN | TOTAL | CÁLCULO | VALOR *P* |
|---|---|---|---|---|---|
| **Intervención** | 18 | 1 | 19 | $\frac{\binom{19}{1}\binom{11}{10}}{\binom{30}{11}}$ | 0,00000383 |
| **Control** | 1 | 10 | 11 | | |
| **Total** | 19 | 11 | 30 | | |

| | NO CAEN | SE CAEN | TOTAL | CÁLCULO | VALOR *P* |
|---|---|---|---|---|---|
| **Intervención** | 19 | 0 | 19 | $\frac{\binom{19}{0}\binom{11}{11}}{\binom{30}{11}}$ | 0,0000000183 |
| **Control** | 0 | 11 | 11 | | |
| **Total** | 19 | 11 | 30 | | |

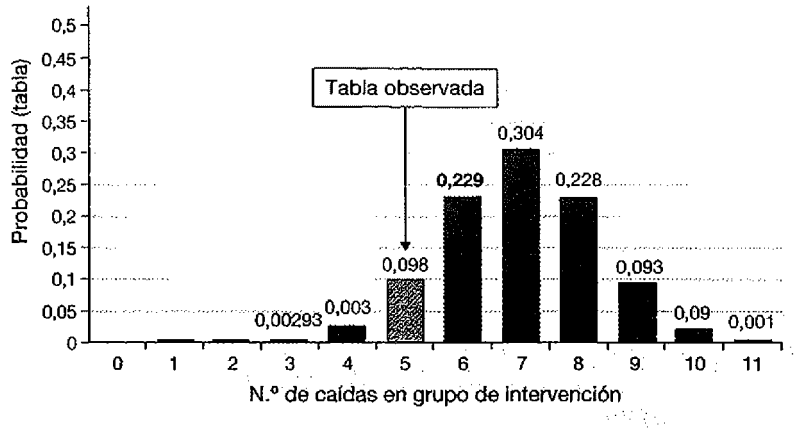



**Figura 5.2** Test exacto de Fisher. Probabilidades de encontrar cada posible tabla para el ejemplo de caídas en ancianos. Se deben sumar todas las posibilidades más extremas, además de la tabla observada para calcular el valor *p* de significación estadística.

Para hacerla a dos colas, habría que calcular la *p* a cada una de las posibles tablas en el otro extremo de la distribución, empezando a partir de la que tenga una probabilidad igual o menor que la de la tabla observada (0,0983) (fig. 5.2). En el ejemplo, consistiría en sumar la probabilidad de encontrar nueve o más caídas en el grupo de intervención.

## 5.7. TEST DE MCNEMAR PARA DATOS EMPAREJADOS

Para introducir el test de McNemar, hay que hablar previamente del concepto de *emparejamiento.* Hasta ahora, las muestras que se iban a comparar eran *independientes,* cada sujeto se observaba una sola vez y cada observación en un grupo no guardaba una relación especial con ninguna observación particular del otro grupo.

Se dice que hay *emparejamiento* cuando:

1. Se realizan mediciones repetidas en un mismo individuo (*autoemparejamiento* o *de medidas repetidas*).
2. Se comparan entre sí —de dos en dos— parejas naturales de individuos, por ejemplo, gemelos *(emparejamiento natural).*
3. Se busca un control para cada paciente, de modo que el control tenga iguales características (edad, sexo, etc.) que el caso *(emparejamiento artificial).* Esto se hace en epidemiología para los estudios de casos y controles emparejados.

Los diseños de investigación de este tipo son más *eficientes,* porque, con menos individuos, se puede alcanzar mayor potencia estadística. Ello se debe a que un individuo es más parecido a sí mismo que al de cualquier otro grupo control, por lo que se reduce la variabilidad aleatoria y una pequeña diferencia entre el antes y el después acaba por ser significativa. Permiten extraer gran cantidad de información de un pequeño número de observaciones, pero no siempre son factibles.

Los estudios emparejados requieren un tratamiento estadístico especial.

El emparejamiento implica diferencias en la estructura de la base de datos, como se muestra en la tabla 5.6.

*Tabla 5.6 Estructura de una base de datos en diseños independientes y en diseños emparejados*

**DATOS INDEPENDIENTES**

| GRUPO | SUCESO |
|---|---|
| 1 | 1 |
| 1 | 0 |
| 1 | 1 |
| 1 | 0 |
| 2 | 1 |
| 2 | 1 |
| 2 | 1 |
| 2 | 0 |

**DATOS EMPAREJADOS POR AUTOEMPAREJAMIENTO**

| INDIVIDUO | PRIMERA VEZ | SEGUNDA VEZ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |
| 7 | 0 | 1 |
| 8 | 1 | 0 |

**DATOS EMPAREJADOS POR EMPAREJAMIENTO NATURAL**

| PAREJA | GEMELO 1 | GEMELO 2 |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |
| 7 | 0 | 1 |
| 8 | 1 | 0 |

**CASOS Y CONTROLES EMPAREJADOS**

| PAREJA | CASO | CONTROL |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 1 | 1 |
| 7 | 0 | 1 |
| 8 | 1 | 0 |

*En los ejemplos de datos independientes y de datos emparejados por autoemparejamiento o emparejamiento natural, 0 indica que no se ha producido el suceso que se está estudiando y 1 que sí se ha producido. En el ejemplo de un estudio de casos y controles emparejados, en las variables «caso» y «control», el 1 significa «expuesto» y el 0 significa «no expuesto».*

Las tablas 2 × 2 son distintas en grupos independientes y en sujetos emparejados o medidas repetidas (v. tabla 5.6). Si se ha asignado el código 1 a que ocurra el suceso (o a que el caso o el control estén expuestos, en estudios de casos y controles) y el código 0 a que no ocurra el suceso (o a que el caso o el control estén expuestos, en estudios de casos y controles), cuando se trata de grupos emparejados, lo que se sitúa dentro de la tabla no son los valores individuales, sino las *parejas* de valores. Si son medidas repetidas, cada casilla sería la pareja de mediciones repetidas para cada sujeto. Por ejemplo, imagínese que se estudia a 20 pacientes con cólico nefrítico recidivante. Todos

***Tabla 5.7 Resultados obtenidos al estudiar las respuestas a metamizol y ketorolaco en el tratamiento del cólico nefrítico recidivante***

| PACIENTE | METAMIZOL | KETOROLACO |
|---|---|---|
| 1 | 1 | 1 |
| 2 | **0** | **1** |
| 3 | 1 | 1 |
| 4 | 0 | 0 |
| 5 | **1** | **0** |
| 6 | 1 | 1 |
| 7 | 1 | 1 |
| 8 | 1 | 1 |
| 9 | **0** | **1** |
| 10 | **0** | **1** |
| 11 | **0** | **1** |
| 12 | 1 | 1 |
| 13 | 1 | 1 |
| 14 | 0 | 0 |
| 15 | 1 | 1 |
| 16 | 1 | 1 |
| 17 | **0** | **1** |
| 18 | 1 | 1 |
| 19 | **0** | **1** |
| 20 | 1 | 1 |

*El 1 indica respuesta al tratamiento y el 0 ausencia de respuesta al tratamiento.*
*Se han destacado en negrita los resultados discordantes con uno y otro tratamiento.*

ellos han sido tratados en un episodio de cólico con metamizol y en otro episodio con ketorolaco. Ambos fármacos son analgésicos que se usan para controlar el dolor en el cólico nefrítico. Se investiga qué tratamiento es más eficaz y se compara la proporción de pacientes que respondieron a metamizol, pero no a ketorolaco, o viceversa (tablas 5.7 y 5.8). En la tabla 5.7 puede apreciarse que cada paciente proporciona dos observaciones.

***Tabla 5.8 Tabla 2 × 2 que resume los resultados obtenidos en la Tabla 5.7 al estudiar las respuestas a metamizol y ketorolaco en el tratamiento del cólico nefrítico recidivante***

| | ÉXITO DEL METAMIZOL | | FRACASO DEL METAMIZOL | |
|---|---|---|---|---|
| Éxito del ketorolaco | 11 | a | **6** | **b** |
| Fracaso del ketorolaco | 1 | c | 2 | d |

*Se han destacado en negrita los resultados discordantes con uno y otro tratamiento.*

La primera pareja de observaciones corresponde a un paciente que respondió bien a ambos tratamientos. En cambio, el segundo paciente no respondió a metamizol, pero sí a ketorolaco. Hay cuatro patrones de posibles parejas (1 + 1; 1 + 0; 0 + 1; 0 + 0), que se organizan en un tabla como la 5.8.

Podría pensarse erróneamente, al ver la tabla 5.8, que aquí sería aplicable la $\chi^2$ de Pearson o el test exacto de Fisher, pero hay una diferencia fundamental entre este ejemplo y los anteriores: ya *no hay dos grupos* de pacientes. Se trata de *un solo grupo de pacientes* que tuvieron dos cólicos nefríticos, y hay *dos mediciones repetidas para cada uno.* Cada paciente es su propio control. Es un diseño emparejado y, por tanto, deberá utilizarse un test especial para datos

emparejados: la $\chi^2$ de McNemar ($\chi^2_{McNemar}$). Los pasos que han de aplicarse para realizar este test serían:

1. *Formular la hipótesis nula.* Mantendría que la proporción de resultados positivos entre ambos tratamientos es igual, y cualquier diferencia observada se debe simplemente al azar. La hipótesis alternativa es que sí hay diferencias.

$$H_0 = \pi_{metamizol} = \pi_{ketorolaco}$$

$$H_1 = \pi_{metamizol} \neq \pi_{ketorolaco}$$

2. *Construir una tabla de datos emparejados.* Para calcular la $\chi^2_{McNemar}$ hay que ordenar los datos como se muestra en la tabla 5.8. Aunque hay 40 resultados, la suma de las cuatro casillas de la tabla da un total de 20, ya que se trata de 20 parejas de valores. Un sujeto que respondió bien a ambos fármacos ha requerido ser observado dos veces; sin embargo, solo aporta una unidad a la casilla «a».
3. *Calcular la ji cuadrado de McNemar* según la siguiente expresión:

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|6-1|-1)^2}{6+1} = \frac{16}{7} = 2,286$$

STATA no resta el valor 1 en el numerador y la $\chi^2$ valdría 25/7 = 3,57.

4. *Hallar los grados de libertad.* Se procede como en la $\chi^2$ de Pearson:

$$\text{Grados de libertad} = (\text{columnas} - 1) \times (\text{filas} - 1)$$

En este problema hay un solo grado de libertad: $(2 - 1) \times (2 - 1) = 1$.

5. *Mirar en las tablas de la ji cuadrado (o en Excel o STATA) la significación estadística* (*valor* p). Para un grado de libertad, los valores críticos de $\chi^2$ son:

| | | | |
|---|---|---|---|
| $\chi^2$ | 2,706 | 3,841 | 5,024 |
| $p$ | 0,10 | 0,05 | 0,025 |

Como el valor que se ha hallado es 2,286 y la $\chi^2$ con un grado de libertad no es significativa a $p = 0,10$ hasta que no llega a valer 2,706, se puede afirmar que la prueba de $\chi^2$ ha resultado no significativa ($p > 0,10$).

6. *Conclusión.* Para un valor de $\chi^2$ de 2,286, la $p$ será mayor de 0,10, es decir, no significativa. Por tanto, no hay evidencias para rechazar la hipótesis nula. En conclusión, no se puede rechazar que la muestra proceda de una población donde la eficacia de ambos analgésicos es la misma.

## 5.8. TEST DE TENDENCIA LINEAL PARA CATEGORÍAS ORDENABLES LÓGICAMENTE (VARIABLES ORDINALES)

Cuando la comparación de proporciones se hace entre categorías o grupos que son susceptibles de seguir un orden en algún sentido, entonces no se debe usar el test de la $\chi^2$ de Pearson, sino el de la $\chi^2$ de tendencia lineal. En la tabla 5.9 se compara la proporción de fumadores entre cuatro grupos definidos por el máximo nivel de estudios alcanzado (variable cualitativa ordinal).

Se aprecia que la proporción de fumadores varía en función del nivel de estudios. Si se aplicase una $\chi^2$ de Pearson, no se estaría contrastando como hipótesis alternativa un descenso *progresivo* en la proporción de fumadores a medida que es mayor el nivel educativo, sino que el test simplemente

*Tabla 5.9 Número de fumadores según el máximo nivel de estudios alcanzado*

| ESTUDIOS | FUMADOR ACTUAL | | TOTAL |
|---|---|---|---|
| | NO | SÍ | |
| <Primarios | 7 | 13 | 20 |
| Primarios | 30 | 20 | 50 |
| Secundarios | 20 | 10 | 30 |
| Universitarios | 75 | 25 | 100 |
| Total | 132 | 68 | 200 |

respondería a la pregunta de si las proporciones son homogéneas o no, sin considerar el sentido o dirección en que crece o disminuye la proporción.

Para calcular la $\chi^2$ de tendencia lineal *a mano,* debe aplicarse un procedimiento que se estudiará más adelante. En definitiva, se trata de calcular coeficiente de correlación de Pearson (v. capítulo 10) representado por la letra *r,* a continuación elevarlo al cuadrado (entonces se llama $R^2$) y multiplicarlo por $n - 1$:

$$\chi^2_{t.lineal} = R^2(n-1) = (-0,24)^2(200-1) = 11,5$$

La $\chi^2$ de tendencia lineal *siempre* tiene un grado de libertad. En este ejemplo, resulta significativa ($p$ = 0,001). Puede rechazarse la hipótesis nula y afirmarse que existe una reducción *progresiva* en la proporción de fumadores a medida que el nivel de estudios es mayor. El signo negativo de $r$ ($r$ = –0,24) indica que una variable (tabaquismo) se reduce a medida que otra sube (estudios).

Otra alternativa para resolver este mismo problema consiste en usar el test de Mann-Whitney (v. capítulo 6), que proporciona un resultado similar ($z$ = –3,3), aunque esto no siempre tiene por qué ser así.

## 5.9. *ODDS RATIO* EN TABLAS 2 × 2

En el capítulo 3 se explicó el concepto de *odds,* que se calculaba dividiendo el número de sujetos que presentaban una determinada característica entre el número de sujetos que no la presentaban. Una *odds* indica la frecuencia con la que se da un fenómeno. Volviendo a otro ejemplo anterior (tabla 5.10), se puede calcular la *odds* de cefalea entre quienes toman helado aceleradamente y quienes lo hacen con pausa. Así, la *odds* de sufrir cefaleas entre quienes toman helado aceleradamente sería $odds_{cefaleas|aceleradamente}$ = 20/53 = 0,377 y entre quienes lo hacen con pausa sería $odds_{cefaleas|pausa}$ = 9/63 = 0,143. Si ahora se quisiese comparar cuántas veces es más frecuente la cefalea en unos que en otros, simplemente se dividirá una *odds* entre otra. Este cociente entre dos *odds* se conoce como *odds ratio* (OR) (3-5). De forma más general, si se considera la cefalea como el evento de interés y tomar helado aceleradamente o no como exposición, se podría formular la OR como:

$$OR = \frac{odds_{evento|expuestos}}{odds_{evento|no\ expuestos}}$$

*Tabla 5.10 Resultados anotados obtenidos para la asociación entre la velocidad a la que se come helado y el riesgo de desarrollar cefalea*

| MODO EN QUE COMEN HELADO | APARICIÓN DE CEFALEA | | | | TOTAL |
|---|---|---|---|---|---|
| | SÍ | | NO | | |
| Aceleradamente | 20 | a | 53 | c | 73 |
| Precavidamente | 9 | b | 63 | d | 72 |
| Total | 29 | 116 | 145 | | |

Así, la *odds ratio* da una idea de cuántas veces es más frecuente el evento entre los expuestos que entre los no expuestos. En el ejemplo, la *odds ratio* de cefaleas sería 0,377/0,143 = 2,64. Esto se interpretaría como que la *odds* de desarrollar cefaleas después de haber comido helado aceleradamente es 2,64 veces superior a la *odds* de desarrollar cefaleas si se consumiese con pausa.

Al ser un cociente, los posibles valores que puede tomar la *odds ratio* oscilan entre 0 e infinito. Cuando la frecuencia del evento sea igual en expuestos y en no expuestos, es decir, cuando no exista asociación alguna (independencia) entre la exposición y el evento, la *odds ratio* tendrá un valor de 1. Si la frecuencia con que se da el evento de interés es mayor en expuestos que en no expuestos, la *odds ratio* tendrá un valor superior a 1 y se interpretará como que la exposición es un *factor de riesgo* para el desenlace que se está estudiando. Por el contrario, cuando la frecuencia con que se dé el evento de interés sea menor en expuestos que en no expuestos, la *odds ratio* tendrá un valor inferior a 1 y se interpretará que la exposición es un *factor protector* para el desenlace en cuestión.

La *odds ratio* no es una razón de proporciones, sino de *odds*. La razón de proporciones (o riesgo relativo, RR) no sería 2,64, sino que sería RR = (20/73)/(9/72) = 2,19. Una ventaja de la *odds ratio* es que es simétrica, da igual intercambiar una variable por otra. No sucede así con el RR, ya que la razón de proporciones de comer helado aceleradamente según se tenga o no cefalea sería RR = (20/39)/(53/116) = 1,12. Puede comprobarse, en cambio, que hacer este intercambio no modificará la OR. La OR se separará siempre más (por abajo o por arriba) del valor nulo (OR = 1) que el RR. Cuanto más frecuente sea el fenómeno, más distancia habrá entre OR y RR.

## 5.10. ERROR ESTÁNDAR E INTERVALO DE CONFIANZA DE LA *ODDS RATIO*

La *odds ratio* calculada en el apartado anterior es un estimador puntual. Como para todo estimador, resultaría interesante poder calcular un rango creíble de datos en el que se esperaría que estuviese contenida la *odds ratio* para la asociación estudiada en la población de la que procede la muestra; es decir, calcular su intervalo de confianza.

Los pasos que han de seguirse para calcular un intervalo de confianza para una *odds ratio* son:

1. *Calcular la* odds ratio.

$$OR = \frac{odds_{evento|expuestos}}{odds_{evento|no\ expuestos}}$$

Sustituyendo los valores de la tabla 5.10, se obtendría:

$$OR = \frac{a/c}{b/d}, \text{ y reordenando } OR = \frac{a \times d}{c \times b}$$

Por eso, a veces se denomina razón de productos cruzados. En el ejemplo concreto:

$$OR = \frac{20/53}{9/63} = 2{,}64$$

2. *Transformar logarítmicamente la estimación puntual de la* odds ratio. La *odds ratio* tiene una escala que abarca de 0 a infinito, pero el nulo es el 1, lo que hace que su distribución sea asimétrica. La transformación logarítmica de la *odds ratio* (lnOR) la hace simétrica, con igual distancia hacia abajo y hacia arriba, y la aproxima a una distribución aproximadamente normal (3). En el ejemplo, lnOR = ln(2,64) = 0,971.

3. *Calcular el error estándar del logaritmo neperiano de la* odds ratio. El error estándar es la raíz cuadrada de la suma de los recíprocos de las cuatro frecuencias de la tabla 2 × 2.

$$EE_{lnOR} = \frac{1}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} = \frac{1}{\sqrt{\frac{1}{20}+\frac{1}{9}+\frac{1}{53}+\frac{1}{63}}} = 0,443$$

4. *Buscar en las tablas el valor de* $z_{\alpha/2}$ *correspondiente al nivel de confianza del IC.* Habitualmente, se calculan intervalos de confianza al 95%, por lo que el valor de $z_{\alpha/2}$ correspondiente es 1,96.
5. *Calcular el intervalo de confianza para el logaritmo neperiano de la* odds ratio. Se hará según la fórmula:

$IC\ (1-\alpha)\ \ln OR = \ln OR \pm z_{\alpha/2} \times EE_{lnOR}$, que, si es al 95%, será:

$$IC\ 95\%\ \ln OR = \ln OR \pm 1,96 \times EE_{lnOR}$$

$$IC\ 95\%\ \ln OR = 0,971 \pm 1,96 \times 0,443 = 0,104 \text{ a } 1,839$$

6. *Revertir la transformación logarítmica.* Esto se hará tomando antilogaritmos de los límites del intervalo de confianza calculado para el logaritmo neperiano de la *odds ratio*:

$$\exp(0,104) = 1,110 \text{ y } \exp(1,839) = 6,289$$

7. *Interpretar el IC obtenido para la* odds ratio. Según los datos obtenidos, la *odds* de desarrollar cefalea es 2,64 veces superior en quienes comen helado aceleradamente que en quienes lo hacen precavidamente. Con una confianza del 95%, se puede afirmar que la verdadera *odds ratio* para esta asociación en la población de la que procede la muestra estará comprendida entre 1,110 y 6,289.

## 5.11. OTRAS MEDIDAS EN TABLAS CATEGÓRICAS (TAU DE KENDALL, GAMMA DE GOODMAN Y KRUSKAL)

La tau *(τ)* de Kendall y la gamma *(γ)* de Goodman y Kruskal se emplean para tablas categóricas cuando las dos variables que se están comparando son variables cualitativas ordinales **(6)**.

Se han definido diferentes tipos de coeficiente tau de Kendall:

- $\tau_a$: no corrige por empates.
- $\tau_b$: corrige por empates y es un buen estimador cuando el número de categorías es el mismo en ambas variables ordinales.
- $\tau_c$: corrige por empates y es recomendable cuando las dos variables ordinales tienen un número diferente de categorías.

Las tres oscilan entre −1 y +1. Un valor de +1 indica una asociación positiva perfecta entre ambas características, es decir, a medida que aumenta una, también lo hace la otra. Un valor de −1 indica una asociación negativa o inversa perfecta entre ambas características, es decir, a medida que aumenta una, la otra disminuye. Cuando no existe asociación, tau valdrá 0.

El índice gamma de Goodman y Kruskal no corrige por empates ni por un número asimétrico de categorías de ambas variables ordinales. También oscila entre −1 y +1, y un valor de 0 indica la ausencia de asociación.

## 5.12. TEST PARA PROPORCIONES Y TABLAS CATEGÓRICAS CON STATA

### 5.12.1. Cálculo de la $\chi^2$ en STATA (caso de una sola variable)

En el apartado 5.5 se ha contrastado si hallar nueve mujeres en una muestra de 20 universitarios es compatible con que esa muestra provenga de una población con una proporción de mujeres del 50%. Esto mismo se podría resolver con STATA con la opción:

**Statistics → Summaries, tables, and tests → Classical tests of hypothesis → One-sample proportion test**

e indicando a continuación la variable para la cual se quiere realizar el contraste de hipótesis (sexo en este caso) y la proporción esperada *(Hypothesized proportion),* que será 0,5. Así, se obtiene la siguiente salida:

```
. prtest sexo == 0.5

One-sample test of proportion                    sexo: Number of obs =       20
------------------------------------------------------------------------------
    Variable |      Mean   Std. Err.                      [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |       .45    .111243                      .2319678    .6680322
------------------------------------------------------------------------------
    p = proportion(sexo)                                          z =  -0.4472
Ho: p = 0.5

     Ha: p < 0.5                 Ha: p != 0.5                 Ha: p > 0.5
 Pr(Z < z) = 0.3274         Pr(|Z| > |z|) = 0.6547          Pr(Z > z) = 0.6726
```

STATA calcula el valor de $z$ en vez de la $\chi^2$ de Pearson. Como en este ejemplo la $\chi^2$ tendría un único grado de libertad, bastaría con elevar la $z$ al cuadrado para obtener el valor de $\chi^2$. STATA ofrece varios valores $p$, considerando tres hipótesis alternativas y que la proporción de mujeres observada es 0,45. La cola de la izquierda es $p = 0{,}3274$ para $H_1$: $p < 0{,}5$. La que aparece en medio ($p = 0{,}6547$) es el test a dos colas que se ha hecho antes. La que figura a la derecha ($p = 0{,}6726$ para $H_1$: $p > 0{,}5$) sería el test a una cola. Se debe elegir habitualmente el test a dos colas ($p = 0{,}6547$).

### 5.12.2. Cálculo de la $\chi^2$ en STATA (caso de dos variables)

Para resolver el ejemplo de la tabla 5.1 con STATA, se deberá escoger la siguiente opción:

**Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association**

Esta selección conduce a otro menú, en el que se indicará que la variable helado va en las filas y la variable cefalea en las columnas, y se solicitará al programa que calcule *Pearson's chi-squared,* seleccionando la casilla correspondiente. Así, se obtendrá:

```
. tabulate helado cefalea, chi2

           |   cefalea
    helado |        no         sí |     Total
-----------+----------------------+----------
precavidam |        63          9 |        72
acelerada  |        53         20 |        73
-----------+----------------------+----------
     Total |       116         29 |       145

          Pearson chi2(1) =   5.0278   Pr = 0.025
```

Se concluirá que hay diferencias estadísticamente significativas ($p < 0{,}05$) entre quienes comen helado aceleradamente y quienes lo hacen con pausa.

### 5.12.3. Cálculo del test exacto de Fisher en STATA

Para resolver el ejemplo expuesto en el apartado 5.7 con STATA, se deberá escoger la siguiente opción:

**Statistics → Summaries, tables, and tests → Tables → Two-way tables with measures of association**

En el menú que aparece se deberá indicar en filas la variable *grupo*, y en columnas la variable *caída*, y solicitar al programa que calcule *Fisher's exact test* seleccionando la casilla correspondiente. Así, se obtendrá:

```
. tabulate grupo caída, exact
```

| grupo | caída no se cae | se cae | Total |
|---|---|---|---|
| control | 5 | 6 | 11 |
| intervención | 14 | 5 | 19 |
| Total | 19 | 11 | 30 |

```
          Fisher's exact =                 0.238
1-sided Fisher's exact =                   0.125
```

Preferiblemente se deberá escoger el valor *p* a dos colas. Se concluirá que no hay evidencia suficiente para rechazar la hipótesis nula que mantiene que el porcentaje de personas que se caen es igual en el grupo control y en el grupo de intervención.

### 5.12.4. Cálculo del test de McNemar en STATA

Para datos emparejados, como los del ejemplo del cólico nefrítico (v. tabla 5.8), con STATA, se procederá así en los menús:

- **Statistics → Epidemiology and related → Tables for epidemiologists → Matched case-control studies**

Se situará en cada una de las dos casillas la variable que contiene el 0 o el 1 para cada tratamiento. Se obtendrá:

```
. mcc metamiz ketorol
```

| Cases | Controls Exposed | Unexposed | Total |
|---|---|---|---|
| Exposed | 11 | 1 | 12 |
| Unexposed | 6 | 2 | 8 |
| Total | 17 | 3 | 20 |

```
McNemar's chi2(1) =      3.57   Prob > chi2 = 0.0588
Exact McNemar significance probability      = 0.1250

Proportion with factor
        Cases              .6
        Controls           .85     [95% Conf. Interval]
        ---------------------------------------------
        difference         -.25    -.5349914   .0349914
        ratio            .7058824   .4909712   1.014866
        rel. diff.        -1.666667 -4.489337   1.156004

        odds ratio       .1666667    .0036234  1.373736   (exact)
```

STATA ha calculado la $\chi^2$ sin restar el valor 1 a la diferencia entre b y c antes de elevarla al cuadrado $(b-c)^2/(b+c) = (6-1)^2/(6+1) = 3{,}57$. Esta aproximación solo es válida con muestras grandes, pero no en este caso. Cuando las muestras son pequeñas ($b < 3$ o bien $c < 30$), solo es válido el procedimiento exacto indicado debajo, que, en este ejemplo, tiene un valor $p = 0{,}125$.

La *odds ratio* en este tipo de diseños es simplemente el cociente de parejas discordantes: OR = c/b = 1/6 = 0,167.

## 5.13. REPRESENTACIÓN DE INTERVALOS DE CONFIANZA PARA PROPORCIONES Y *ODDS RATIO* CON STATA

En el ensayo PREDIMED, se quiso comparar si había diferencias en la incidencia de diabetes tipo 2 entre participantes asignados a una intervención dirigida a incrementar su adhesión al patrón de dieta mediterránea (expuestos) y entre los asignados al grupo control a los que se aconsejó seguir una dieta baja en grasa. Los resultados obtenidos en uno de los centros del ensayo (7) se resumen en la tabla 5.11.

La proporción de participantes que desarrollan diabetes tipo 2 entre los asignados a la dieta mediterránea es de 30/284 = 0,1056, y entre los asignados al grupo control 24/134 = 0,1791. Si se calculase el error estándar para cada una de estas proporciones, se obtendría:

$$\sqrt{\frac{0{,}1056\times(1-0{,}1056)}{284}}=0{,}0182$$

para el grupo de dieta mediterránea y:

$$\sqrt{\frac{0{,}1791\times(1-0{,}1791)}{134}}=0{,}0331$$

para el grupo control. Además, se podría comparar si hay diferencias en la incidencia de diabetes entre ambos grupos calculando la *odds ratio* de desarrollar diabetes entre los asignados a uno u otro grupo. Con estos datos, se obtendría una OR = 0,541, con un intervalo de confianza al 95% que iría de 0,302 a 0,968. Si se quisiera representar estos datos gráficamente, se podrían emplear las siguientes órdenes en STATA:

```
clear
input ///
grupo   diab    n
0       24      134
1       30      284
end
gen p = diab/n
gen EEP =(p*(1-p)/n)^.5
gen masEEP=p+EEP
gen menosEEP=p-EEP
twoway (bar p grupo, bcolor(sand)) ///
(rcap menosEEP masEEP grupo) ///
, ylabel(0(.1).3, angle(horizontal)) ///
xlabel(0 "Control" 1 "D. Mediterránea") ///
xtitle("Grupo") ytitle("Riesgo de diabetes") ///
legend(order(1 "proporción" 2 "+/- EE"))
```

Así, se obtendría la figura 5.3.

***Tabla 5.11 Resultados del ensayo PREDIMED para valorar el efecto de la dieta mediterránea sobre la prevención de diabetes tipo 2***

| | DESARROLLO DE DIABETES TIPO 2 | | TOTAL |
|---|---|---|---|
| | SÍ | NO | |
| Dieta mediterránea | 30 | 254 | 284 |
| Control | 24 | 110 | 134 |
| Total | 54 | 364 | 418 |

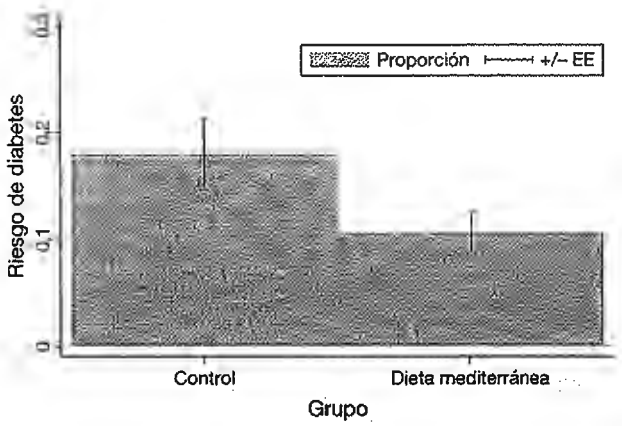



**Figura 5.3** Proporción (y error estándar) de participantes que desarrollan diabetes tipo 2 en el ensayo PREDIMED.

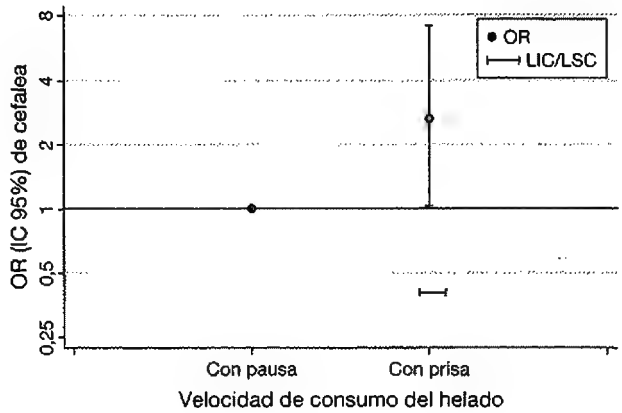



**Figura 5.4** *Odds ratio* de desarrollar cefalea según la velocidad de consumo de helado.

En cambio, en la figura 5.4 se recoge cómo representar una *odds ratio* con sus intervalos de confianza aplicándolo al ejemplo de la velocidad de consumo de helado y el riesgo de cefalea. Es conveniente representar la *odds ratio* en escala logarítmica.

Las órdenes en STATA para conseguir esta figura serían:

```
clear
input ///
grupo   cefal   n
0       0       63
0       1       9
1       0       53
1       1       20
end
expand n
cc grupo cefal
g OR=1
g LSC=1
g LIC=1
replace OR=r(or) if grupo==1
replace LIC=r(lb_or) if grupo==1
replace LSC=r(ub_or) if grupo==1
twoway (scatter OR grupo, scale(log)) ///
(rcap LIC LSC grupo),  ysca(range(0.25 8) log) ///
yline(1) ylabel(.25 .5 1 2 4 8)  ///
xlabel(-1 " " 0 "Con pausa" 1 "Con prisa" 2 " ") ///
xtitle("Velocidad consumo helado") ///
ytitle("OR (95% CI) de cefalea")
```

Los nombres de variables precedidos de una *r* y que contienen un paréntesis, por ejemplo *r(or)*, etc., corresponden a variables *internas* que almacena STATA cuando se da la orden *cc*.

## 5.14. TEST PARA PROPORCIONES CON OTROS PROGRAMAS

### 5.14.1. Cálculo de la $\chi^2$ en SPSS (caso de una sola variable)

Para comparar una proporción observada frente a una esperada y poder resolver el ejemplo del apartado 5.5, hay que seleccionar:

**Analizar → Pruebas no paramétricas → Cuadros de diálogo antiguos → Chi-cuadrado**

Aparecerá un cuadro de diálogo, donde se debe seleccionar primero la variable cuya distribución en grupos o categorías se desea contrastar con lo esperado. Esta variable aparece en el recuadro de la izquierda. Al hacer doble clic sobre ella, pasará al recuadro central. A continuación, basta con pulsar en el botón «Aceptar». Por defecto aparecen como valores esperados los correspondientes a que todas las categorías sean iguales, pero esto se puede cambiar.

Se obtendrá:

**sexo**

| | N observado | N esperado | Residual |
|---|---|---|---|
| varón | 11 | 10,0 | 1,0 |
| mujer | 9 | 10,0 | -1,0 |
| Total | 20 | | |

**Estadísticos de contraste**

| | sexo |
|---|---|
| Chi-cuadrado | ,200[a] |
| gl | 1 |
| Sig. asintót. | ,655 |

a. 0 casillas (0,0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 10,0.

En la primera tabla, la primera columna son las categorías de la variable *sexo*. Después aparecen los valores observados (N observados) y esperados (N esperados) para cada categoría, así como la diferencia entre ambos. A esta diferencia se le llama a veces «residual», porque es lo que quedaría sin explicar (el «residuo», lo «residual») si la hipótesis nula fuese cierta. SPSS también da la suma del total de los observados, que debe coincidir con el total de individuos de la muestra. Finalmente, proporciona el valor de la $\chi^2$, los grados de libertad (gl) y el valor de la *p* o significación estadística asintótica (Sig. asintót.). Al final, el programa informa de que se cumplen las condiciones de aplicación, ya que en ninguna casilla el esperado es menor de 5.

### 5.14.2. Cálculo de la $\chi^2$ con SPSS: la comparación de dos proporciones

Para resolver el ejemplo de la tabla 5.1 con SPSS, se debe seleccionar:

**Analizar → Estadísticos descriptivos → Tablas de contingencia**

En la ventana que aparece, se arrastra la variable *helado* al recuadro de filas y la variable *cefalea* al recuadro de columnas. A continuación, en ese mismo menú se pulsa el botón superior derecho, que lleva por título «Estadísticos», y aparecerá la nueva ventana que contiene como primera opción

«Chi cuadrado». Se debe pulsar el pequeño recuadro situado a la izquierda de «Chi cuadrado». Por último se pinchan los botones «Continuar» y «Aceptar». La salida que se obtiene incluye los siguientes resultados:

**Tabla de contingencia helado * cefalea**

Recuento

| | | cefalea | | |
|---|---|---|---|---|
| | | no | sí | Total |
| helado | precavidamente | 63 | 9 | 72 |
| | aceleradamente | 53 | 20 | 73 |
| Total | | 116 | 29 | 145 |

**Pruebas de chi-cuadrado**

| | Valor | gl | Sig. asintótica (bilateral) | Sig. exacta (bilateral) | Sig. exacta (unilateral) |
|---|---|---|---|---|---|
| Chi-cuadrado de Pearson | 5,028[a] | 1 | ,025 | | |
| Corrección por continuidad[b] | 4,140 | 1 | ,042 | | |
| Razón de verosimilitudes | 5,135 | 1 | ,023 | | |
| Estadístico exacto de Fisher | | | | ,037 | ,020 |
| Asociación lineal por lineal | 4,993 | 1 | ,025 | | |
| N de casos válidos | 145 | | | | |

a. 0 casillas (0,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 14,40.

b. Calculado sólo para una tabla de 2x2.

Dentro de la tabla 2 × 2 aparecen los valores observados para cada combinación. Aparecen los marginales, suma por filas, columnas y total. La primera $\chi^2$ que aparece es la de Pearson, que es la que se ha calculado antes (5,028). Siempre que se pide a SPSS una $\chi^2$ desde este menú para una tabla 2 × 2, ofrece también el test exacto de Fisher, que se verá a continuación. El programa también da el mínimo de los valores esperados (14,40). Si algún valor esperado en una tabla 2 × 2 es inferior a 5, no sirve la $\chi^2$ de Pearson y hay que recurrir al test exacto de Fisher.

### 5.14.3. Cálculo de la $\chi^2$ con R/Splus

R/Splus no es especialmente adecuado para calcular este test. Es mejor recurrir a Excel, introduciendo las ecuaciones en cada casilla. En R/Splus, se debe aplicar el test a un nuevo objeto que es la **tabla** construida cruzando las dos variables (**t_trans**, en el ejemplo):

```
INSTRUCCIONES A «R» para hacer un test χ²
> mi<-c(rep(0,36),rep(1,48))
> trans<-c(rep(0,23),rep(1,13),rep(0,20),rep(1,28))
> t_trans<-table(mi,trans)
> t_trans
   trans
   mi  0  1
   0  23  13
   1  20  28
> chisq.test(t_trans)
```

Esto conduciría a obtener los siguientes resultados:

```
Pearson's Chi-squared test with Yates' continuity correction

data: t_trans
X-squared = 3.225, df = 1, p-value = 0.07252
```

Por defecto, R solo obtiene la $\chi^2$con la corrección de Yates.

### 5.14.4. Programación en Excel de una calculadora para $\chi^2$

Si bien la opción de calcular una $\chi^2$ de Pearson no está implementada directamente en Excel, se ha programado una hoja de cálculo que la realiza y que está descargable en http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

### 5.14.5. Cálculo del test exacto de Fisher en SPSS

Para resolver el ejemplo del apartado 5.7 con SPSS, se debe seleccionar:

**Analizar → Estadísticos descriptivos → Tablas de contingencia**

En la ventana que aparece, se arrastra la variable *grupo* al recuadro de filas y la variable *caída* al recuadro de columnas. A continuación, en ese mismo menú se pulsa el botón superior derecho, que lleva por título «Estadísticos», y aparecerá la nueva ventana que contiene, como primera opción, «Chi cuadrado». Se debe pulsar el pequeño recuadro situado a la izquierda de «Chi cuadrado». Por último, se pulsan los botones «Continuar» y «Aceptar». La salida que se obtiene incluye los siguientes resultados:

**Tabla de contingencia grupo * caída**

Recuento

| | | caída | | |
|---|---|---|---|---|
| | | no se cae | se cae | Total |
| grupo | control | 5 | 6 | 11 |
| | intervención | 14 | 5 | 19 |
| Total | | 19 | 11 | 30 |

**Pruebas de chi-cuadrado**

| | Valor | gl | Sig. asintótica (bilateral) | Sig. exacta (bilateral) | Sig. exacta (unilateral) |
|---|---|---|---|---|---|
| Chi-cuadrado de Pearson | 2,391[a] | 1 | ,122 | | |
| Corrección por continuidad[b] | 1,330 | 1 | ,249 | | |
| Razón de verosimilitudes | 2,371 | 1 | ,124 | | |
| Estadístico exacto de Fisher | | | | ,238 | ,125 |
| Asociación lineal por lineal | 2,311 | 1 | ,128 | | |
| N de casos válidos | 30 | | | | |

a. 1 casillas (25,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 4,03.

b. Calculado sólo para una tabla de 2x2.

Dentro de la tabla 2 × 2 aparecen los valores observados para cada combinación. Se muestran los marginales y la suma por filas, columnas y total.

SPSS ha calculado el test de Fisher a una y dos colas. Hay que tener en cuenta que la $\chi^2$ por definición es siempre a dos colas. Cuando se pida a SPSS una $\chi^2$ para comparar dos proporciones, siempre facilitará también el test exacto de Fisher, que es válido en general en este tipo de tablas. Según se cumplan o no los requisitos de aplicación, se podrá usar o no el test de la $\chi^2$ de Pearson.

En este ejemplo, se concluirá que no hay evidencia suficiente para rechazar la hipótesis nula que mantiene que la proporción de sujetos que se caen es igual en ambos grupos, ya que el valor *p* a dos colas es estrictamente mayor que 0,05.

### 5.14.6. Cálculo del test de McNemar en SPSS

Se puede hacer de dos modos. Uno de ellos consiste en seleccionar:

**Analizar → Estadísticos descriptivos → Tablas de contingencia**

En el menú que aparece se pulsa el botón superior derecho, que lleva por título «Estadísticos», y aparecerá la nueva ventana, que contiene en la parte inferior derecha la opción McNemar. Después ya solo habrá que pulsar continuar y aceptar.

El otro modo de realizar el test de McNemar se basa en seleccionar:

**Analizar → Pruebas no paramétricas → Cuadros de diálogo antiguos → 2 muestras relacionadas...**

donde se abrirá un menú en el que se deberá introducir en el primer par para contrastar la variable *metamizol* como Variable 1, y la variable *ketorolaco* como Variable 2.

Las salidas que se obtienen por ambos procedimientos son prácticamente iguales y se muestran a continuación:

**Tabla de contingencia metamizol * ketorolaco**

Recuento

| | | ketorolaco | | Total |
|---|---|---|---|---|
| | | 0 | 1 | |
| metamizol | 0 | 2 | 6 | 8 |
| | 1 | 1 | 11 | 12 |
| Total | | 3 | 17 | 20 |

**Pruebas de chi-cuadrado**

| | Valor | Sig. exacta (bilateral) |
|---|---|---|
| Prueba de McNemar | | ,125[a] |
| N de casos válidos | 20 | |

a. Utilizada la distribución binomial

SPSS presenta la tabla de datos emparejados, indica que hay 20 casos (con medidas repetidas) y calcula un valor *p* de significación estadística basado en la distribución binomial exacta a dos colas y que es, por tanto, exacto, lo mismo que hace STATA. En este caso, el valor *p* obtenido es 0,125.

## 5.15. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| $\chi^2$ de Pearson para una variable | prtest *var* = =*num* | NPAR TESTS<br>/CHISQUARE=*var*<br>/EXPECTED=*n_categ1 n_categ2*<br>/MISSING ANALYSIS |
| $\chi^2$ de Pearson para dos proporciones | *tabulate var1 var2, chi2* | CROSSTABS<br>/TABLES=var1 BY var2<br>/FORMAT=AVALUE TABLES<br>/STATISTICS=CHISQ<br>/CELLS=COUNT<br>/COUNT ROUND CELL |
| Test exacto de Fisher | *tabulate var1 var2, exact* | CROSSTABS<br>/TABLES=var1 BY var2<br>/FORMAT=AVALUE TABLES<br>/STATISTICS=CHISQ<br>/CELLS=COUNT<br>/COUNT ROUND CELL |

## REFERENCIAS


1. Greenhalgh T. Statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ 1997;315(7104):364-6.
2. Kaczorowski M, Kaczorowski J. Ice cream evoked headaches. Ice cream evoked headaches (ICE-H) study: randomised trial of accelerated versus cautious ice cream eating regimen. BMJ 2002;325(7378):1445-6.
3. Bland JM, Altman DG. Statistics notes. The odds ratio. BMJ 2000;320(7247):1468.
4. Martínez-González MA, De Irala-Estévez J, Guillén-Grima F. ¿Qué es una odds ratio? Med Clin (Barc) 1999;112(11):416-22.
5. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Editorial Ariel; 2008.
6. Indrayan A. Medical Biostatistics. 3rd ed. Boca Raton: Chapman &Hall/CRC biostatistics series; 2013.
7. Salas-Salvadó J, Bulló M, Babio N, Martínez-González MA, Ibarrola-Jurado N, Basora J, et al. PREDIMED Study Investigators. et al. Reduction in the incidence of type 2 diabetes with the Mediterranean diet: results of the PREDIMED-Reus nutrition intervention randomized trial. Diabetes Care 2011;34(1):14-9.

# COMPARACIÓN DE MEDIAS ENTRE DOS GRUPOS

6

*E. Toledo, C. López del Burgo, C. Sayón-Orea, M. Á. Martínez-González*

## 6.1. TEST DE LA *T* DE STUDENT PARA DOS MUESTRAS INDEPENDIENTES

Este test compara dos *medias* en muestras independientes. Se aplica cuando la variable comparada es cuantitativa y solo hay dos grupos. Es un test muy utilizado. Se entenderá con un ejemplo. Se compara el efecto de dos dietas para saber si la pérdida de peso (en kilogramos) difiere en función de si se siguió una u otra dieta. Habrá dos grupos (dieta 1 y dieta 2) y se comparará el cambio medio de peso entre ambos grupos. Supóngase que las reducciones de peso a 3 meses fuesen las que aparecen en la figura 6.1. En este tipo de problemas habrá siempre dos variables: una es cuantitativa (en el ejemplo, la pérdida de peso) y la otra es cualitativa *dicotómica,* pues solo admite dos categorías (en el ejemplo, la dieta, que solo puede ser dieta 1 o dieta 2). A la variable cuantitativa (supuesto *efecto*) se le llama variable *dependiente* y al factor dicotómico *(exposición)* se le denomina variable *independiente.* Se trata de comprobar si la variable cuantitativa (cambio de peso, en el ejemplo) *depende* de la dicotómica (seguir una u otra dieta).

En el ejemplo (v. fig. 6.1), se comparará el peso medio perdido (18,58 kg) en los 19 participantes que siguieron la dieta 1 con respecto a la media de pérdida de peso (17,55 kg) de los 11 que siguieron la dieta 2. El procedimiento estadístico más habitual para comparar estas medias consistirá en usar la *t de Student para dos muestras independientes con varianzas homogéneas.* Para realizar esto, se darán los siguientes pasos:

1. *Formular la hipótesis nula y la hipótesis alternativa.* Así, si se llama $\mu$ a la media de kilogramos adelgazados en la población:
   a. Hipótesis nula ($H_0$) $\equiv \mu_{dieta1} = \mu_{dieta2}$. (Ambas medias son iguales en la población.)
   b. Hipótesis alternativa ($H_1$) $\equiv \mu_{dieta1} \neq \mu_{dieta2}$. (Las medias difieren entre ambas dietas.)
2. *Verificar que se cumplen los requisitos de aplicación.* Estos son:
   a. Normalidad en la distribución de la variable o $n > 30$ en cada grupo. Si no se pudiese asumir la normalidad, se intentará una transformación de los datos en sus logaritmos y se repetirá la comprobación de la normalidad con la variable transformada. Cuando hay asimetría positiva (caso frecuente en medicina), suele mejorar la aproximación a la normal al hacer la transformación logarítmica (v. apartado 6.5). Pero, si tampoco entonces se aproxima a la normalidad, se deberá aplicar una prueba no paramétrica, en este caso la *U* de Mann-Whitney (v. apartado 6.7). En caso de que alguno de los grupos tenga menos de 10 observaciones, es mejor usar directamente la *U* de Mann-Whitney (**1-3**).
   b. Homogeneidad de varianzas (v. más adelante). Si no se cumpliese, debe usarse el test de Welch.
3. *Estimación de la varianza conjunta*, también llamada varianza ponderada ($s_p^2$). Esta varianza tiene en cuenta que la muestra total está dividida en dos grupos y se calcula como una media ponderada de las varianzas de cada grupo. Los pesos de esa ponderación son los grados de libertad de cada grupo:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

«Exposición» = factor ***independiente*** dicotómico que establece o indica los grupos.

«Efecto» = variable ***dependiente*** numérica que se compara.

| DIETA 1 | | | DIETA 2 | | |
|---|---|---|---|---|---|
| **ID** | **Dieta *(gr)*** | **Pérdida de peso (kg) *(cambpes)*** | **ID** | **Dieta *(gr)*** | **Pérdida de peso (kg) *(cambpes)*** |
| 1 | 1 | 19 | 20 | 2 | 23 |
| 2 | 1 | 28 | 21 | 2 | 23 |
| 3 | 1 | 17 | 22 | 2 | 5 |
| 4 | 1 | 15 | 23 | 2 | 14 |
| 5 | 1 | 16 | 24 | 2 | 22 |
| 6 | 1 | 21 | 25 | 2 | 20 |
| 7 | 1 | 27 | 26 | 2 | 4 |
| 8 | 1 | 18 | 27 | 2 | 12 |
| 9 | 1 | 15 | 28 | 2 | 18 |
| 10 | 1 | 10 | 29 | 2 | 33 |
| 11 | 1 | 10 | 30 | 2 | 19 |
| 12 | 1 | 2 | | | |
| 13 | 1 | 12 | | | |
| 14 | 1 | 13 | | | |
| 15 | 1 | 19 | | | |
| 16 | 1 | 9 | | | |
| 17 | 1 | 29 | | | |
| 18 | 1 | 35 | | | |
| 19 | 1 | 38 | | | |
| Suma | | 353 | Suma | | 193 |
| Media | | 18,58 | Media | | 17,55 |
| *s* | | 9,26 | s | | 8,43 |
| *n* | | 19 | n | | 11 |

**Figura 6.1** Pérdida de peso a 3 meses en participantes que siguen dos dietas distintas.

En el ejemplo:

$$s_p^2 = \frac{18 \times 85{,}8 + 10 \times 71{,}1}{18 + 10} = 80{,}55$$

La desviación típica ponderada $s_p$ será la raíz cuadrada de la varianza ponderada:

$$s_p = \sqrt{s_p^2}$$

$$s_p = 80{,}55 \,\hat{}\, 0{,}5 = 8{,}975$$

Los grados de libertad son el denominador de la varianza ponderada:

$$\text{gl} = (n_1 - 1) + (n_2 - 1) = N - 2$$

$$\text{gl} = (19 - 1) + (11 - 1) = 18 + 10 = 28$$

4. *Calcular la diferencia entre las dos medias.* El cálculo de la diferencia de medias se realiza mediante una simple sustracción o resta: $\bar{x}_1 - \bar{x}_2$.

   En el ejemplo: 18,58 – 17,55 = 1,03.

5. *Calcular el error estándar de la diferencia de medias (EEDM).* El cálculo del error estándar de la diferencia de medias se consigue mediante la siguiente expresión:

$$\text{EEDM} = s_p\sqrt{(1/n_1)+(1/n_2)}$$

   En el ejemplo:

$$\text{EEDM} = 8{,}975\sqrt{1/19+1/11} = 3{,}4$$

6. *Calcular el valor de la* t *de Student.* Se calcula un cociente entre un *efecto* y un *error:* la diferencia entre las dos medias (efecto) se divide entre el error estándar de la diferencia de medias (EEDM). En vez de utilizar la distribución normal, se usa una *t* de Student porque se desconocen *dos* parámetros poblaciones (no solo la diferencia poblacional de medias $\mu_1$-$\mu_2$, sino también la varianza poblacional, $\sigma^2$). Se toman prestados los índices estadísticos muestrales (medias muestrales y $s^2$) y la distribución *t* de Student proporciona una corrección por el tamaño de muestra, de ahí que haya que tener siempre en cuenta sus grados de libertad: existe una distribución *t* para cada número de grados de libertad. A medida que el tamaño de muestra se hace mayor, la *t* de Student se parece más a la normal. Cuando $n > 100$, es prácticamente igual usar una distribución normal.

   La fórmula de la *t* de Student en caso de que las varianzas sean homogéneas es la siguiente:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\text{EEDM}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

   En el ejemplo:

$$t_{19+11-2} = \frac{1{,}03}{3{,}4} = 0{,}30$$

7. *Comparar con las tablas de la* t *de Student.* Una vez obtenido el valor de *t*, se debe comparar con el de las tablas para N – 2 grados de libertad, siendo $N = n_1 + n_2$, es decir, la suma del número de individuos de cada una de las dos muestras. Si el valor de *t* encontrado en el experimento es superior al de las tablas, podrá rechazarse la hipótesis nula y se demostraría que sí hay diferencias significativas entre ambas medias. Si la *t* encontrada es inferior a la de las tablas, no se rechazará la hipótesis nula (no habrá diferencias significativas), pero se correrá el riesgo de equivocarse, y ese riesgo equivaldrá al error beta (cuanto menor sea el tamaño de muestra, mayor es el error beta).

   En el ejemplo, el número de grados de libertad es gl = 28. Buscando en la tabla, se halla que una $t_{28}$ ha de valer al menos 2,048 para ser significativa al 5% (dos colas). Como el valor para *t* obtenido era de 0,30, no hay evidencias para rechazar la hipótesis nula de que la media del peso perdido era la misma con las dos dietas.

   En STATA se puede pedir lo siguiente:

```
display 2*ttail(28,0.30)
```

   y se obtendrá: $p = 0{,}766$.

Al resolverlo usando Excel, el valor *p* (a dos colas) se obtendrá con: =DISTR.T(0,30;28;2), que devuelve un valor *p* de 0,766.

El valor *p* asociado a cada posible resultado de *t* indica la probabilidad de encontrar las diferencias observadas o unas todavía mayores si no hubiese diferencia poblacional entre ambas dietas ($H_0$).

8. *Conclusión.* Se concluye que no hay diferencias significativas en el peso perdido entre los dos grupos. Por lo tanto, no se puede rechazar la hipótesis nula que mantiene que no hay diferencias en la pérdida de peso entre ambas dietas.

Este test exige asumir, además de la normalidad, la homogeneidad de varianzas («homoscedasticidad»). La normalidad se debe comprobar, como se indicó en el capítulo 3, siempre que $n_1$ o $n_2$ sean menores que 30. La homoscedasticidad requiere comprobar que las varianzas de ambos grupos son iguales (es decir, homogéneas).

## 6.2. TEST PARA COMPARAR VARIANZAS

Mediante la prueba F para la homogeneidad de varianzas se comprobará que no hay diferencias significativas entre las varianzas. Para ello se calcula el cociente entre las varianzas de ambos grupos. Este cociente entre varianzas se conoce como F. Un ejemplo es:

$$F_{18,10} = \frac{s^2_{max}}{s^2_{min}} = \frac{9,26^2}{8,43^2} = 1,21$$

Este cociente F se obtiene simplemente al dividir la varianza mayor entre la menor. Cuanto mayor sea F, más desiguales (menos homogéneas) serán las dos varianzas. F tiene dos tipos de grados de libertad: los de la varianza del numerador y los de la varianza del denominador. Aquí son 18 (19 – 1) y 10 (11 – 1), ya que los grados de libertad son $n_i - 1$.

Se buscará en las tablas de la F de Snedecor el valor crítico (para $p = 0,05$), pero siempre es más práctico recurrir a STATA o Excel. Si se usan las tablas, se comparará el valor obtenido (1,21) con el valor crítico para estos grados de libertad, y así se sabrá si hay diferencias significativas entre ambas varianzas. Si la F calculada es superior al valor que aparece en las tablas como límite de la significación estadística, o si el valor *p* que proporciona Excel o STATA es $p < 0,05$, se concluye que las varianzas *no* son homogéneas entre sí y no será válida la *t* calculada antes. Si el cociente F es inferior al valor crítico de las tablas o Excel o STATA dan un valor $p > 0,05$, podrá asumirse que las varianzas son homogéneas y se podrá usar con tranquilidad el test que se acaba de exponer.

En el ejemplo, las varianzas eran homogéneas, ya que el valor de F encontrado (1,21) es inferior al necesario para que $p = 0,05$.

En Excel:

=DISTR.F(1,21;18;10) devuelve un valor $p = 0,392$.

En STATA:

```
di Ftail(18,10,1.21)
.39027169
```

Hay otros test alternativos para comprobar que las varianzas son homogéneas: test de Bartlett, test de Levene y otros. Si el valor *p* correspondiente a estos test es inferior a 0,05, entonces se asume

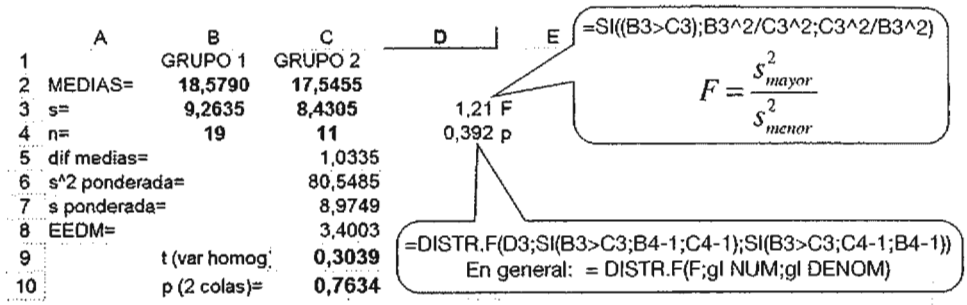



**Figura 6.2** Comprobación con Excel de la homogeneidad de varianzas en un test *t* para comparar dos medias independientes.

que las varianzas son distintas. En STATA se puede pedir un test de comparación de varianzas con la orden **sdtest:**

```
. sdtest cambpes, by(gr)

Variance ratio test
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 1 | 19 | 18.57895 | 2.125198 | 9.263523 | 14.11407 | 23.04382 |
| 2 | 11 | 17.54545 | 2.541881 | 8.430464 | 11.88179 | 23.20912 |
| combined | 30 | 18.2 | 1.612737 | 8.833322 | 14.90158 | 21.49842 |

```
    ratio = sd(1) / sd(2)                                 f =   1.2074
Ho: ratio = 1                             degrees of freedom =   18, 10

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
 Pr(F < f) = 0.6083          2*Pr(F > f) = 0.7834           Pr(F > f) = 0.3917
```

El cociente F aparece en este listado de STATA con letra minúscula (f = 1,2074).

Puede apreciarse que el valor *p* de significación estadística para el test F de homogeneidad de varianzas es *p* = 0,3917; por lo tanto, no hay diferencias significativas entre ambas varianzas y puede asumirse que son homogéneas. Si las varianzas fuesen distintas (heteroscedasticidad), se debe emplear otra variedad del test *t*, en la que se modifican el error estándar y los grados de libertad. Esta *t* modificada se denomina *aproximación de Welch.*

La comprobación de la homogeneidad de varianzas en Excel se hace como se indica en la figura 6.2.

SPSS usa automáticamente el test de Levene para valorar la heteroscedasticidad (la hipótesis nula es la homoscedasticidad o igualdad de varianzas) siempre que se le pide una *t* de Student. De este test solo interesa el valor *p*, ya que la F será distinta de la calculada a mano. Incluso puede suceder que también haya disparidad en los valores *p* obtenidos con uno y otro método. En caso de duda, siempre es más recomendable usar el test para varianzas heterogéneas en cuanto haya sugerencia de que las varianzas pueden ser heterogéneas.

## 6.3. TEST *T* PARA DOS MEDIAS INDEPENDIENTES CON VARIANZAS HETEROGÉNEAS (TEST DE WELCH)

Este test es más robusto que el de varianzas homogéneas y es preferible por muchas propiedades, pero requiere hacer dos modificaciones:

1. En el denominador de la *t* de Student, en vez de usar una única varianza ponderada, se deben usar las varianzas de cada grupo separadamente para calcular el error estándar.

2. Los grados de libertad (gl*) ya no son N – 2, sino que deben calcularse usando una fórmula más compleja, tal y como se presenta a continuación:

$$t_{gl^*} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad gl^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Obsérvese que los grados de libertad anteriormente definidos gl* podrían tener decimales.

Se verá ahora un ejemplo con varianzas heterogéneas. En un estudio de casos y controles que valoró si la exposición a cromo se relacionaba con el riesgo de infarto de miocardio (4), se recoge la comparación entre el índice de masa corporal (IMC) de los casos y los controles.

| | Casos | Controles |
|---|---|---|
| IMC medio | 26,5 | 25,9 |
| *s* | 3,9 | 3,4 |
| *n* | 684 | 724 |

El test F resultaría significativo:

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2} = \frac{3,9^2}{3,4^2} = \frac{15,21}{11,56} = 1,32$$

F = 1,32 (con 683 y 723 grados de libertad) resulta significativo, pues, si se le pide a Excel la significación con =DISTR.F(1,32;683;723), devolverá un valor $p$ = 0,0001. Esto indica que las varianzas son significativamente diferentes, es decir, heterogéneas.

Debe calcularse primero el valor para la *t*:

$$t_{gl^*} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{26,5 - 25,9}{\sqrt{\frac{3,9^2}{684} + \frac{3,4^2}{724}}} = \frac{0,6}{0,1955} = 3,07$$

Después se calculan los grados de libertad corregidos (gl*):

$$gl^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{15,21}{684} + \frac{11,56}{724}\right)^2}{\frac{\left(\frac{15,21}{684}\right)^2}{683} + \frac{\left(\frac{11,56}{724}\right)^2}{723}} = 1.355,7$$

Usando Excel obtenemos el valor $p$ a dos colas correspondiente:

=DISTR.T(3,07;1355,7;2), que devuelve $p$ = 0,002, indicando que hay diferencias estadísticamente significativas entre el IMC medio de los casos y los controles.

En resumen, para comparar las medias de dos grupos independientes, una vez comprobado que se puede asumir la normalidad (o que $n \geq 30$ en los dos grupos), debe procederse como indica el algoritmo de la figura 6.3.

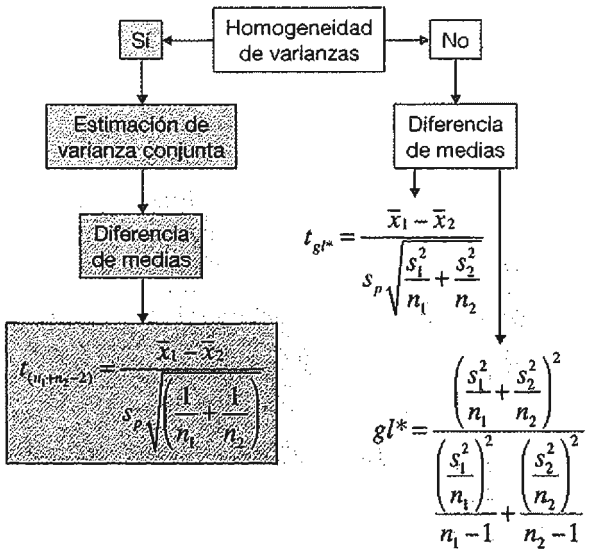



**Figura 6.3** Comparación de dos medias con la *t* de Student.

No se podrá usar la t si se desea comparar más de dos muestras. Habrá que emplear el análisis de la varianza o ANOVA (v. capítulo 8).

## 6.4. INTERVALO DE CONFIANZA PARA LA DIFERENCIA DE MEDIAS

Lo anterior resuelve el contraste de hipótesis, pero se llega a conclusiones algo limitadas: simplemente, que *no existen diferencias significativas* en el primer caso (dos dietas) y que sí existen en el segundo (casos y controles de infarto). Pero faltará algo imprescindible: estimar la *magnitud* de la diferencia entre ambos grupos. Esto se resuelve calculando unos límites de confianza a la diferencia de medias. Si las varianzas son homogéneas, se usará una ecuación parecida al intervalo de confianza para una media, pero, en vez de utilizar la media muestral, se utilizará la diferencia de medias y, en vez del error estándar de la media, se usará el error estándar de la diferencia de medias (EEDM):

$$\text{IC } 95\% = \text{dif. medias} \pm t_{0,025,28}\ (\text{EEDM})$$

En el primer ejemplo del cambio de peso con dos dietas, resultará:

$$\text{IC } 95\% = 1,034 \pm 2,0484 \times 3,4 = -5,93 \text{ a } 8,00$$

Obsérvese que el intervalo de confianza incluye lo postulado por la hipótesis nula (diferencia = 0), lo cual es coherente con la ausencia de significación estadística.

¿De dónde sale el valor $t_{0,025,28}$ = 2,0484? Como se desea un intervalo de confianza al 95%, el error alfa será del 5% (0,05) a dos colas, es decir, con alfa = 0,025 en cada cola, por lo cual se representa $t_{0,025}$. Como los grados de libertad de la comparación son N – 2, *t* tendrá 28 grados de libertad ($t_{28}$). Esto se puede obtener de las tablas de la *t* de Student, o bien de Excel, con la expresión:

=DISTR.T.INV(0,05;28), que devolverá 2,0484.

O bien de STATA, con la expresión:

```
display invttail(28,0.025)
```

Se interpretaría así: «Con la dieta 1 se perdió una media de peso que era 1,03 kg superior a la pérdida de peso obtenida con la dieta 2. Con una confianza del 95%, puede decirse que la diferencia entre las dietas 1 y 2 oscilaba entre 8 kg más y 5,9 kg menos para la dieta 1.» Se confía al 95% en que la verdadera diferencia poblacional esté entre estos dos valores.

En el caso de varianzas heterogéneas, el EEDM no utilizará la desviación estándar ponderada, ya que no pueden combinarse ambas varianzas (por ser distintas), sino que será (en el ejemplo del estudio de casos y controles de infarto):

$$\text{EEDM} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3{,}9^2}{684} + \frac{3{,}4^2}{724}} = 0{,}1955 \text{ (con gl* = 1355,7)}$$

$$\text{IC 95\%} = \text{dif. medias} \pm t_{0{,}025,1355{,}7}\,(\text{EEDM}) = (26{,}5 - 25{,}9) \pm 1{,}96(0{,}1955) = 0{,}22 \text{ a } 0{,}98$$

Obsérvese que ahora *no* es compatible con la hipótesis nula (hay diferencias estadísticamente significativas).

## 6.5. TRANSFORMACIÓN LOGARÍTMICA DE LA VARIABLE DEPENDIENTE EN UN TEST *T*

No se podrá usar la *t* si la muestra es pequeña y no se puede asumir la normalidad. Puede intentarse entonces una transformación de la variable dependiente para conseguir así una mayor aproximación a la normalidad. También es recomendable (pero no imprescindible) probar la transformación logarítmica con muestras grandes, pues en ocasiones así se mejora la adaptación a la normal. No obstante, esto complica la interpretación de los resultados.

En el siguiente ejemplo se desea comprobar si la calidad de vida (medida de 0 a 100) de los pacientes intervenidos quirúrgicamente en un servicio depende de si la cirugía ha sido de larga estancia o de alta precoz. Los datos son los que aparecen en la tabla 6.1.

Se dispone de 12 observaciones en cada grupo. Como $n_i < 30$, es imprescindible comprobar la normalidad *en cada grupo*. El grupo de alta precoz presenta asimetría positiva y no supera el test de normalidad, pero su transformación logarítmica sí supera el test de normalidad, como puede comprobarse en la figura 6.4.

***Tabla 6.1 Calidad de vida (escala 0 a 100) según tipo de cirugía (muestras independientes)***

| ALTA PRECOZ | LARGA ESTANCIA |
|---|---|
| 19 | 66 |
| 43 | 28 |
| 24 | 32 |
| 86 | 25 |
| 40 | 26 |
| 43 | 20 |
| 31 | 23 |
| 40 | 17 |
| 24 | 16 |
| 12 | 10 |
| 40 | 17 |
| 24 | 16 |

```
. bys gr: ladder QoL
```

-> gr = larga

| Transformation | formula | chi2(2) | P(chi2) |
|---|---|---|---|
| cubic | QoL^3 | 20.97 | 0.000 |
| square | QoL^2 | 19.39 | 0.000 |
| identity | QoL | 14.20 | 0.001 |
| square root | sqrt(QoL) | 9.48 | 0.009 |
| log | log(QoL) | 4.47 | 0.107 |
| 1/(square root) | 1/sqrt(QoL) | 1.52 | 0.468 |
| inverse | 1/QoL | 4.56 | 0.102 |
| 1/square | 1/(QoL^2) | 12.43 | 0.002 |
| 1/cubic | 1/(QoL^3) | 17.46 | 0.000 |

-> gr = precoz

| Transformation | formula | chi2(2) | P(chi2) |
|---|---|---|---|
| cubic | QoL^3 | 20.35 | 0.000 |
| square | QoL^2 | 17.41 | 0.000 |
| identity | QoL | 9.59 | 0.008 |
| square root | sqrt(QoL) | 4.46 | 0.108 |
| log | log(QoL) | 0.83 | 0.659 |
| 1/(square root) | 1/sqrt(QoL) | 3.64 | 0.162 |
| inverse | 1/QoL | 7.75 | 0.021 |
| 1/square | 1/(QoL^2) | 15.06 | 0.001 |
| 1/cubic | 1/(QoL^3) | 18.87 | 0.000 |

**Figura 6.4** Comprobación de la normalidad en STATA con la orden **ladder.** Se debe comprobar para cada grupo por separado. Los datos corresponden al ejemplo de la tabla 6.1.

La transformación permite realizar el test de la *t* de Student, pero hay que realizar todas las operaciones con la variable transformada logarítmicamente según aparece en la tabla 6.2:

$$s_p^2 = \frac{(11)(0,47^2)+(11)(0,5^2)}{(12+12-2)} = 0,24; \Rightarrow s_p = \sqrt{0,24} = 0,49$$

$$t_{22} = \frac{3,45-3,09}{0,49\sqrt{\left(\frac{1}{12}+\frac{1}{12}\right)}} = \frac{0,36}{0,20} = 1,8$$

***Tabla 6.2 Resumen de los datos de la tabla 6.1 tras su transformación logarítmica usando el logaritmo neperiano como escala de la calidad de vida***

| ALTA PRECOZ | LARGA ESTANCIA |
|---|---|
| Media = 3,45 | Media = 3,09 |
| Desv. est. = 0,50 | Desv. est. = 0,47 |
| N = 12 | N = 12 |

El valor de *t* a dos colas no resulta estadísticamente significativo ($p$ = 0,08). Está en *«tierra de nadie»*. No se puede rechazar la hipótesis nula porque faltan evidencias para comparar la calidad de vida según el alta sea precoz o no. Este es un problema de las muestras pequeñas.

Puede estimarse también la diferencia de medias mediante intervalos de confianza:

$$\text{IC } 95\% \ (\ln[\mu_1] - \ln[\mu_2]) = (3{,}45 - 3{,}09) \pm 2{,}074(0{,}2) = -0{,}05 \text{ a } +0{,}77$$

Estos son los límites de confianza en *escala logarítmica*. Para interpretarlos se deben transformar a escala lineal. Debe tenerse en cuenta que, cuando se hace una media con logaritmos neperianos como se ha hecho aquí, el antilogaritmo de esa cantidad ya no es una media aritmética, sino la *media geométrica* (MG): $MG = e^{\sum \ln(x_i)/n}$

Es decir, la media aritmética de los logaritmos es el logaritmo de la media geométrica.

$$\frac{\sum \ln(x_i)}{n} = \ln(MG)$$

Como la diferencia entre el logaritmo de dos cantidades equivale al logaritmo del cociente de ambas cantidades [ln(A) – ln(B) = ln(A/B)], cuando se toman antilogaritmos de la diferencia del logaritmo de las medias geométricas se cumplirá que la diferencia entre los dos logaritmos es igual al logaritmo del cociente:

$$\ln(MG_1) - \ln(MG_2) = \ln\left(MG_1 / MG_2\right)$$

En el ejemplo, la diferencia entre las medias de los dos grupos (transformadas logarítmicamente) era 0,5:

$$\ln(MG_1) - \ln(MG_2) = 3{,}45 - 3{,}09 = \ln\left(MG_1 / MG_2\right) = 0{,}36$$

Si el ln ($MG_1/MG_2$) = 0,36, el cociente entre las dos medias geométricas será igual al antilogaritmo de 0,36. Por lo tanto, $MG_1/MG_2 = e^{0{,}36} = 1{,}43$, y los límites de confianza al 95% para el *cociente de medias geométricas* serán los que se estarán estimando:

$$\text{IC } 95\% \left(\frac{\mu_{gcom1}}{\mu_{gcom2}}\right) = e^{-0.05} \text{ a } e^{+0.77} = 0{,}95 \text{ a } 2{,}16$$

***Interpretación:*** *la media geométrica de la calidad de vida es 1,43 veces superior en el grupo 1 (alta precoz). Hay una confianza del 95% de que la media geométrica de la calidad de vida se multiplique por un factor que estará entre 0,95 y 2,16 veces en los pacientes con alta precoz con respecto a las estancias prolongadas. Como se aprecia, el intervalo de confianza incluye el valor 1, que sería un cociente unitario (igualdad entre ambos grupos).*

Otras transformaciones (1/*x*, raíz cuadrada, etc.) son también difíciles de interpretar al revertirlas tras el intervalo de confianza. Ante esta dificultad, una recomendación pragmática es pasar a usar métodos no paramétricos (*U* de Mann-Whitney) cuando no se consigue la normalidad con la transformación logarítmica, lo que sucede muchas veces. También es útil y válida la aproximación pragmática de realizar los cálculos por ambos métodos (con y sin transformación; por métodos paramétricos y no paramétricos) y solo preferir el que usa transformación o el no paramétrico cuando los resultados difieran (1). Con mucha frecuencia, sobre todo con muestras mayores que esta, a pesar de pequeñas transgresiones de los supuestos, los resultados serán bastante similares con uno y otro método. Esto suele confirmar la validez de la aproximación utilizada, da tranquilidad y corrobora las conclusiones. En cambio, nunca será correcto realizar diversas aproximaciones con el objetivo tendencioso de buscar aquel método que proporcione los resultados deseados por

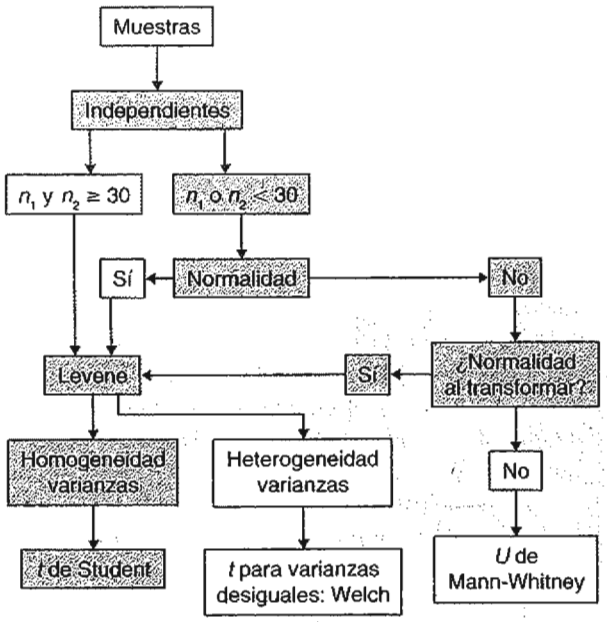



**Figura 6.5** Algoritmo de decisiones en una comparación de medias independientes.

el *investigador.* Sería contrario a la ética y afortunadamente tampoco suele ser posible si se sigue lo aquí establecido.

La figura 6.5 presenta las alternativas cuando se desea comparar una variable cuantitativa en dos grupos independientes. Se ha sombreado el recorrido concreto seguido en el ejemplo de la calidad de vida, que requirió transformación logarítmica para conseguir la adaptación a la normal, se pudo asumir la homogeneidad de varianzas y se terminó por emplear una prueba *t* para varianzas homogéneas.

Además de los problemas de normalidad y homogeneidad de varianzas, debe tenerse en cuenta que cuando uno de los grupos tiene *menos de 10 observaciones,* o cuando la variable que se compara (variable dependiente) sigue una escala *ordinal,* entonces se debe elegir directamente el test de la *U* de Mann-Whitney, obviando los pasos de comprobación de la normalidad o la transformación de la variable.

## 6.6. TEST DE LA *T* DE STUDENT PARA COMPARAR UNA MEDIA CON UN VALOR DE REFERENCIA

El problema más simple que se puede resolver con el test de la *t* de Student es el referido a una sola muestra, tendría $n - 1$ grados de libertad y es un caso poco frecuente.

Puede presentarse con el siguiente ejemplo. Se ha determinado el colesterol sérico a 25 personas. En esta *muestra,* la media resultó ser 208 mg/dl y la desviación típica fue de 20 mg/dl. ¿Podría asumirse que la media *poblacional* del colesterol es 200 mg/dl? El planteamiento es:

Hipótesis nula ($H_0$) $\equiv \mu = 200$ (la media poblacional vale 200).

Hipótesis alternativa ($H_1$) $\equiv \mu \neq 200$ (la media poblacional es diferente de 200).

Se soluciona el problema calculando una cantidad «t», con una expresión muy parecida a la que se utiliza para calcular *z* usando la normal. Simplemente se trata de sustituir la desviación

estándar *(s)* por el error estándar de la media (EEM = $s / \sqrt{n}$), ya que aquí no se trata de hallar la probabilidad de que *un sujeto* con 208 mg/dl provenga de una población con media $\mu$ de 200, sino de encontrar la probabilidad de que la *media de una* muestra ($\bar{x}$ = 208), provenga de una población con media $\mu$ de 200. El error estándar es a la muestra lo que la desviación estándar al individuo. De hecho, en el caso extremo de que la muestra tuviese tamaño 1, el error estándar y la desviación estándar coincidirían. En el ejemplo, la *t* valdrá:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{208 - 200}{20 / \sqrt{25}} = 2$$

Esta cantidad *t* se dice que tiene 24 grados de libertad, porque hay 25 individuos en una sola muestra y los grados de libertad son $n - 1$. Se mirará en las tablas de la *t* de Student (con 24 grados de libertad) cuál es el valor de *t* que dejaría a cada lado un error alfa del 5% (a dos colas). El valor que aparece en las tablas es $t_{g.l.=24,\alpha/2=0,025}$ = 2,064. Como el valor de la tabla (2,064) es superior al valor encontrado (*t* = 2,00), no se puede rechazar $H_0$ con un valor *p* a dos colas inferior al 5%. No obstante, sí se podría afirmar que el valor *p* es inferior al 10%, ya que, según las tablas, $t_{g.l.=24,\alpha/2=0,05}$ = 1,711 y lo encontrado para *t* (*t* = 2,00) es superior a 1,711. Si le pidiésemos a Excel un valor concreto de *t*, nos daría un valor *p* = 0,057.

Siempre es preferible usar STATA o Excel. En STATA se debe multiplicar por dos el valor de la cola solicitada:

```
. display 2*ttail(24,2)
.05693985
```

En Excel se debe introducir:

=DISTR.T(2,0;24;2)

La última cifra se refiere a que se solicita el valor *p* a dos colas.

En la figura 6.6 aparece una calculadora programada en Excel para resolver este tipo de problemas.

La interpretación más correcta es que si la media *poblacional* ($\mu$) realmente fuese de 200 mg/dl, habría una probabilidad del 5,7% de hallar una media muestral de 208 o todavía más alejada de 200. Este alejamiento incluye también la posibilidad de que la media muestral sea ≤192 mg/dl (la otra cola).

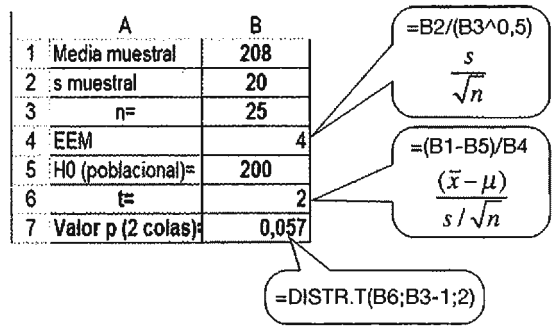



**Figura 6.6** Calculadora programada en Excel para resolver problemas de *t* de Student para una sola muestra (comparación de media muestral con valor externo de referencia).

## 6.7. TEST DE LA *U* DE MANN-WHITNEY

Es un procedimiento no paramétrico que sustituye a la *t* para comparar las medias de dos grupos independientes (5). Como requiere ordenar los valores antes de hacer el test, no compara realmente las dos medias, sino las dos *medianas*. Se debe usar en vez de la *t* si:

- alguna de las dos muestras contiene menos de 30 observaciones y no se puede asumir la normalidad (ni trasformando la variable), o
- se trata de una variable *ordinal* en vez de ser realmente cuantitativa, o
- la muestra es muy pequeña (<10 observaciones en alguno de los dos grupos).

Si se da alguna de estas circunstancias, el test indicado es el de Mann-Whitney y no la *t*. La ventaja de Mann-Whitney es que está libre de requisitos, supuestos y distribuciones. Sus inconvenientes son:

- Se pierde potencia (aproximadamente para una comparación que resulta significativa con 95 individuos en el test *t*, se necesitan 100 individuos para que sea significativa con Mann-Whitney).
- En principio, no proporciona intervalos de confianza, aunque podrían hacerse (6).

Hay dos modos de realizarlo, uno se debe a Wilcoxon y otro a Mann y Whitney. Por eso hay textos (o *software*) que le llaman test de Wilcoxon. Es mejor llamarle test de Mann-Whitney para evitar confusiones con el otro test de Wilcoxon emparejado (v. más adelante). Además, Mann y Whitney describieron una aproximación más intuitiva al calcular un índice estadístico *U*, que presenta dos ventajas:

1. Es más fácil de interpretar.
2. A no ser que haya muchas categorías en la variable dependiente, es más fácil de calcular la *U* que la alternativa *W* que propuso Wilcoxon.

Se aplicará este test al mismo ejemplo de la figura 6.1 de las dos dietas para perder peso. La *U* se obtiene al ir comparando cada individuo de un grupo con cada individuo del otro para contabilizar el número de veces que alguien de un grupo presenta un valor superior a alguien del otro (5). Para facilitar las comparaciones, se debe comenzar ordenando en cada grupo de menor a mayor la variable comparada (tabla 6.3).

A continuación, cada participante del grupo 1 se enfrenta con cada participante del grupo 2. Así, por ejemplo, el último participante (n.° 19) con la dieta 1 que perdió 38 kg de peso ha tenido una pérdida de peso que supera a los 11 sujeros con dieta 2, por eso se anotaría 11 *victorias* en las comparaciones por parejas. Su tanteo se representa en la columna de la izquierda *Dieta 1 mejor*. El siguiente participante, que adelgazó 35 kg, también aventaja a los 11 del grupo 2 y se anota 11 victorias, y así sucesivamente. Se contabilizan así las «victorias» de cada sujeto cuando se le enfrenta a cada uno del otro grupo, obteniéndose los siguientes resultados:

- Suma de *dieta 1 mejor* = 0 + 2 + … + 11 = 100.
- Suma de dieta 2 mejor = 1 + 1 + … + 17 = 105.

Habrá 100 victorias del grupo 1 y 105 del grupo 2. Pero hay que tener en cuenta los empates. Para ello, hay que preguntarse: ¿cuántas comparaciones 1 a 1 (por parejas) se pueden hacer en este ejemplo? Como hay 19 sujetos en el grupo con dieta 1 y 11 sujetos en el que siguió la dieta 2, se podrán hacer 209 comparaciones posibles (19 × 11 = 209). Sin embargo, al sumar las victorias de cada grupo, el resultado es 205 (100 + 105) y no 209. Sucede así porque hay cuatro empates (efectivamente: 19, 19, 18 y 12, *en cursiva*). Como había un total de 209 comparaciones posibles, el resultado ha sido: en 105 de ellas ha ganado el grupo 2, en cuatro ha habido empates, y en las

***Tabla 6.3 Los mismos datos de la figura 6.1 (pérdidas de peso) ordenados de menor a mayor para cada uno de los dos grupos***

| DIETA 1 MEJOR | DIETA 1 | DIETA 2 | DIETA 2 MEJOR |
|---|---|---|---|
| 0 | 2 | 4 | 1 |
| 2 | 9 | 5 | 1 |
| 2 | 10 | *12* | 4 |
| 2 | 10 | 14 | 6 |
| 2 | *12* | *18* | 10 |
| 3 | 13 | *19* | 11 |
| 4 | 15 | 20 | 13 |
| 4 | 15 | 22 | 14 |
| 4 | 16 | 23 | 14 |
| 4 | 17 | 23 | 14 |
| 4 | *18* | 33 | 17 |
| 5 | *19* | | |
| 5 | *19* | | |
| 7 | 21 | | |
| 10 | 27 | | |
| 10 | 28 | | |
| 10 | 29 | | |
| 11 | 35 | | |
| 11 | 38 | | |

*Los datos en negrita indican el número de pacientes que siguieron cada una de las dos dietas, para distinguirlos de la primera y cuarta columnas, que señalan los casos en los que las dietas del segundo grupo son mejoradas por un caso concreto de las dietas del primer grupo y viceversa. Los valores en cursiva representan los empates en la valoración anterior.*

100 restantes ha ganado el grupo 1. Como es lógico, los cuatro empates se reparten equitativamente: dos al que tiene 105, que pasa a tener 107, y dos al que tiene 100, que pasa a tener 102. Estas sumas finales (incluyendo empates) corresponden a lo que miden unas cantidades que se llaman $U_1$ y $U_2$. $U_1$ equivale al número de comparaciones en las que alguien del grupo 1 perdió más peso que alguien del grupo 2. Para resolver el problema basta con calcular solo una de ellas, $U_1$ o $U_2$. Sabiendo, por ejemplo, que $U_2$ vale 107 y que el total de comparaciones posibles es $19 \times 11 = 209$, $U_1$ forzosamente tiene que valer 102 (fig. 6.7).

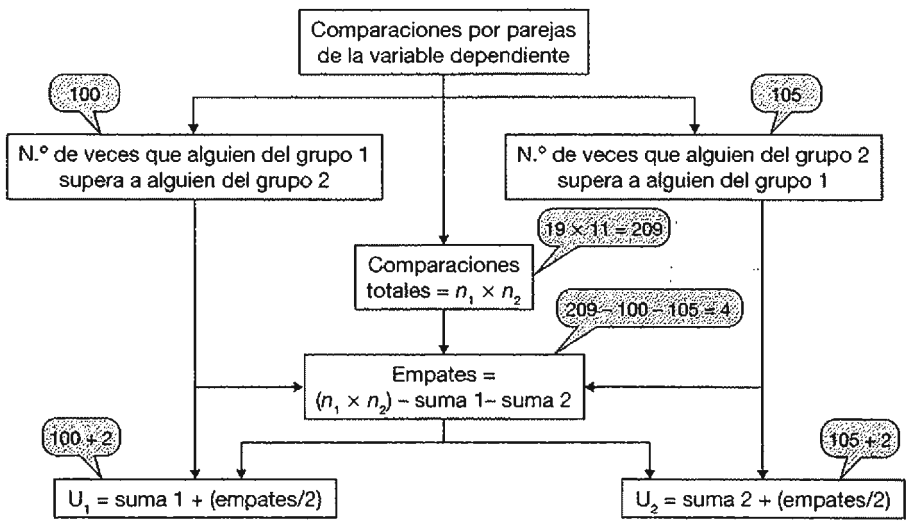



**Figura 6.7** Modo de calcular el test de Mann-Whitney.

Una vez hallada cualquiera de las cantidades *U*, por ejemplo $U_2$, se puede calcular un valor *z* que permite hacer un contraste de hipótesis:

$$z = \frac{U_2 - (n_1 n_2 / 2)}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} = \frac{107 - (19 \times 11/2)}{\sqrt{19 \times 11(19 + 11 + 1)/12}} = \frac{2,5}{23,2} = 0,108$$

Para una $z = 0,108$, según la distribución normal, la cola de la derecha vale 0,457. Como es preferible hacer test a dos colas, el valor *p* bilateral correspondiente a la *z* encontrada es 0,914, claramente no significativo, y se concluirá diciendo que no hay evidencia para rechazar la hipótesis nula de igualdad en el cambio de peso entre las dos dietas.

Se han calculado además dos cantidades $U_1$ y $U_2$. *¿Qué interpretación tiene, por ejemplo, $U_2$?*

Además de tener muchas y muy buenas interpretaciones musicales, *$U_2$ en el test de Mann-Whitney* tiene una interpretación directa.

Se halló $U_2 = 107$. Se sabe que el total de comparaciones posibles es 209 (19 × 11). Pues bien, puede decirse que $U_2 = 107$ se interpreta como que en el 51,2% de las veces el grupo con dieta 2 superó al grupo con dieta 1 en cuanto a la pérdida de peso, dado que 107/209 = 0,512. En el 51,2% de las comparaciones la victoria fue para el grupo 2, y podría interpretarse como que existe una probabilidad de 0,512 de que una nueva observación procedente del grupo con dieta 2 sea superior a una nueva observación que proceda del otro grupo. Alternativamente, podrá decirse que hay solo una probabilidad de 0,488 de adelgazar más con la dieta 1 que con la 2.

### 6.7.1. Test de Mann-Whitney con datos agrupados

Se compara un grupo de casos ($n_1 = 171$) con un grupo independiente de controles ($n_2 = 171$) en cuanto a una exposición medida en escala ordinal (I al IV), y los datos son:

| Exposición | Casos | Controles |
|---|---|---|
| IV | 21 | 70 |
| III | 34 | 33 |
| II | 61 | 44 |
| I | 55 | 24 |

Se comparará la exposición entre los dos grupos (casos y controles) *agrupadamente*. Así, los 21 casos con exposición = IV están más expuestos que los 101 controles con exposiciones I, II y III (24 + 44 + 33), con lo que alcanzan 21 × 101 = 2.121 *victorias* de los casos frente a los controles. Con este mismo método se hallarán los resultados de la tabla 6.4.

Habrá 5.897 comparaciones con *victoria* de los casos (2.121 + 2.312 + 1.464) y 16.748 *victorias* de controles. Hay que adjudicar los empates, que se calculan así:

$$\text{Empates} = (171 \times 171) - 5.897 - 16.748 = 29.241 - 5.897 - 16.748 = 6.596$$

Una vez repartidos equitativamente, las cantidades $U_1$ y $U_2$ serán:

$$U_1 = 5.897 + (6.596/2) = 9.195 \text{ y } U_2 = 16.748 + (6.596/2) = 20.040$$

*Tabla 6.4* **U *de Mann-Whitney con datos agrupados***

| EXPOSICIÓN | CASOS | CASOS > CONTROLES | CONTROLES | CONTROLES > CASOS |
|---|---|---|---|---|
| IV | 21 | 21 × 101 = 2.121 | 70 | 70 × 150 = 10.500 |
| III | 34 | 34 × 68 = 2.312 | 33 | 33 × 116 = 3.828 |
| II | 61 | 61 × 24 = 1.464 | 44 | 44 × 55 = 2.420 |
| I | 55 | | 24 | |

El valor $z$ será:

$$z = \frac{20.040 - (171 \times 171/2)}{\sqrt{171 \times 171(171+171+1)/12}} = 5{,}93, \text{ claramente significativo } (p < 0{,}0001).$$

La $z$ así calculada requerirá una pequeña corrección por empates que simplemente nosotros repartimos equitativamente. Es siempre recomendable realizar el test de la $U$ de Mann-Whitney con ordenador, pues el ordenador proporciona una cantidad $z$ que difiere un poco de la calculada por nosotros.

La interpretación directa de $U_2$ = 20.040 aquí encontrado se derivará de que representa el 68,5% del total de comparaciones posibles (20.040/(171 × 171) = 0,685). Existe, por tanto, una probabilidad de 0,685 de que al elegir al azar un caso y un control, el control esté más expuesto que el caso.

Este problema habría podido resolverse también por la ji cuadrado de tendencia lineal, obteniendo una respuesta similar[1].

## 6.8. TEST DE LA *T* DE STUDENT PARA DATOS EMPAREJADOS (MUESTRAS RELACIONADAS)

En los test antes vistos se comparaban las medias de dos grupos *independientes*. Hay otra situación, caracterizada porque los grupos no son independientes, sino que cada observación de un grupo está relacionada específicamente con otra observación del otro grupo; es decir, hay *parejas* de valores relacionados. Se trata entonces de un diseño *emparejado*. Puede deberse a mediciones repetidas en un mismo individuo *(autoemparejamiento)*, a parejas *naturales* de individuos (p. ej., gemelos) o a un *emparejamiento artificial* donde se busca un control para cada caso de enfermedad, de tal modo que el control sea, por ejemplo, de la misma edad y sexo que el caso.

Desde un punto de vista práctico, se advertirá que los datos son emparejados cuando en la base de datos no existe una columna para la variable grupo y otra para la variable dependiente. Ahora, como se aprecia en la figura 6.8, habrá dos columnas para la variable dependiente (antes-después o caso-control).

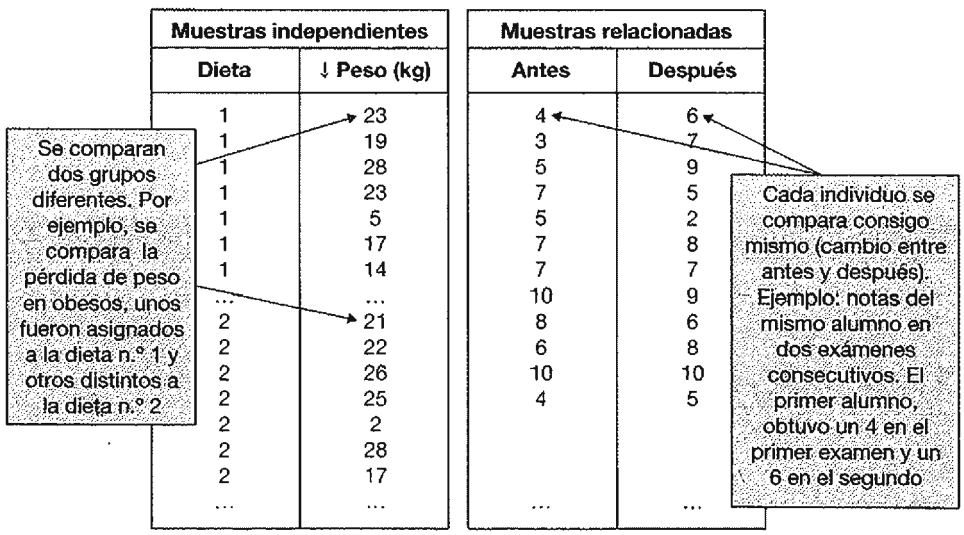



**Figura 6.8** Comparación de medias independientes o relacionadas (emparejadas).

1 El coeficiente de correlación de Pearson será –0,3355, el coeficiente de determinación ($R^2$) será 0,1126 y $\chi^2$ = 341 × 0,1126 = 38,4 para el test de tendencia lineal, con $p < 0{,}0001$.

*Tabla 6.5 Comparación de medias emparejadas (medidas repetidas)*

| ID | ANTES | DESPUÉS | DIF |
|---|---|---|---|
| 1 | 10 | 14 | 4 |
| 2 | 10 | 12 | 2 |
| 3 | 12 | 12 | 0 |
| 4 | 13 | 12 | –1 |
| 5 | 14 | 8 | –6 |
| 6 | 15 | 12 | –3 |
| 7 | 16 | 9 | –7 |
| 8 | 16 | 10 | –6 |
| 9 | 16 | 16 | 0 |
| 10 | 16 | 12 | –4 |
| 11 | 16 | 11 | –5 |
| 12 | 16 | 8 | –8 |
| 13 | 17 | 15 | –2 |
| 14 | 19 | 11 | –8 |
| 15 | 20 | 16 | –4 |
| | Media | | –3,2 |
| | DE | | 3,65 |
| | EEM | | 0,94 |

*dif = diferencia de medias antes – después; id = número de participantes.*

En los diseños de medidas repetidas (autoemparejamiento) no se valora la variabilidad *entre* individuos (interindividual), sino *dentro* de un mismo individuo (intraindividual). Algo análogo podría aplicarse de algún modo a la pareja en diseños *emparejados*. El tratamiento estadístico es distinto, porque la variabilidad aleatoria intraindividual (o intrapareja) es menor que la interindividual.

El siguiente ejemplo (tabla 6.5) corresponde a muestras relacionadas o emparejadas. Imagínese que se realiza un estudio sobre la rapidez de reflejos de un grupo de 15 individuos antes y media hora después de tomar un botellín de cerveza. En una escala continua de 0 a 20 se mide su capacidad de reacción a diversos estímulos: cuanto más alto es el valor de la escala, mayor es la rapidez de reacción.

Para comprobar si tras consumir alcohol se ha reducido la rapidez de reflejos, se deberá calcular, en primer lugar, la variable diferencia *(dif)* entre antes y después del consumo de alcohol para cada individuo, que aparece en la última columna de la tabla 6.5. El planteamiento es el siguiente:

Hipótesis nula: $(H_0) \equiv \mu_{antes} = \mu_{después}$

Hipótesis alternativa: $(H_1) \equiv \mu_{antes} \neq \mu_{después}$

Hipótesis nula: $(H_0) \equiv \overline{dif} = 0$

Hipótesis alternativa: $(H_1) \equiv \overline{dif} \neq 0$

Se calcula el error estándar de la media de la diferencia y resulta ser 0,94, como aparece en la tabla 6.5, ya que la desviación estándar de la variable *dif* es 3,65:

$$EE_{dif} = \frac{s_{dif}}{\sqrt{n}} = \frac{3,65}{\sqrt{15}} = 0,94$$

La *t* de Student emparejada es simplemente el cociente entre la media de la variable *dif* y su error estándar, y tendrá $n - 1$ grados de libertad (gl).

$$t_{n-1} = \frac{\overline{x}_{dif}}{\left(\frac{s_{dif}}{\sqrt{n}}\right)} = \frac{-3,2}{0,94} = -3,4$$

Los gl serán 15 – 1 = 14. Se comparará el valor encontrado ($t = -3,4$) con el valor tabulado de la *t* de Student, teniendo en cuenta que los grados de libertad son 14. Un valor de 3,4 para 14 grados

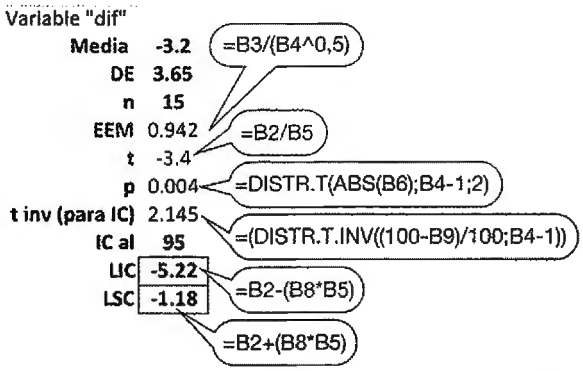



**Figura 6.9** Programación en Excel de una calculadora para resolver el test *t* de muestras emparejadas.

de libertad es significativo a dos colas con una $p < 0{,}01$. Por tanto, puede concluirse que ha existido un descenso significativo ($p < 0{,}01$) de la capacidad de reflejos después del consumo de alcohol.

*Cuando el tamaño muestral sea inferior a 30, se debe comprobar si esta nueva variable (dif) sigue una distribución normal.* Si no fuera así, habría que intentar una transformación para conseguir la distribución normal; si tampoco se consigue de este modo, habría que recurrir a una prueba no paramétrica (test de Wilcoxon; v. siguiente epígrafe).

Debe acompañarse siempre el valor *p* del intervalo de confianza para la variable diferencia *(dif)*. Se calcula exactamente igual que para la estimación descriptiva de una media. Es preciso sumar y restar a la media de *dif* su error estándar multiplicado por un valor de *t*. Si se desea un intervalo de confianza al 95% para una muestra de 15 individuos, la *t* para el intervalo de confianza deberá tener 14 grados de libertad y el error alfa a dos colas del 5% deberá valer:

$$t_{\alpha/2} = 0{,}05, \text{gl} = 14 = 2{,}145$$

que se puede obtener de Excel con la expresión:

= DISTR.T.INV(0,05;14), o de STATA con:

```
. display invttail(14,0.025)
```

Es importante destacar que, para calcular el intervalo de confianza, la *t* que se usa es la de las tablas ($t = 2{,}145$) y no la hallada en la comparación ($t = 3{,}4$). El intervalo de confianza valdrá:

$$\text{IC } 95\% = -3{,}2 \pm 2{,}145\left(\frac{3{,}649}{\sqrt{15}}\right) = -3{,}2 \pm 2{,}02 = -5{,}221 \text{ a } -1{,}179$$

La interpretación indicaría que el descenso medio fue de 3,2 puntos en la escala. Con una confianza del 95%, este descenso estará entre 1,18 y 5,22 puntos.

En Excel puede programarse un test *t* para muestras relacionadas o emparejadas, disponiendo solo del tamaño muestral, la media y la desviación estándar de la variable *dif*, como se presenta en la figura 6.9 con los datos del ejemplo anterior.

## 6.9. TEST DE WILCOXON PARA DATOS EMPAREJADOS

Es la alternativa no paramétrica de la *t* emparejada. Debe usarse si:

- Los datos para comparar son ordinales.
- Los datos son cuantitativos, pero la muestra es pequeña ($n < 30$) y, además, no sigue una distribución normal en la variable diferencia entre las dos mediciones emparejadas (ni siquiera tras realizar transformaciones). Se usará como ejemplo la práctica deportiva en la juventud de nueve parejas de gemelos donde, después de un seguimiento de 20 años, uno de los gemelos ha muerto y el otro sobrevive. Se

*Tabla 6.6 Ejemplo de test de Wilcoxon*

| PAREJA | GEMELO VIVO | GEMELO MUERTO | DIF | DIF (ABS) | N.° DE ORDEN SIN CORREGIR | RANGOS | R+ | R– |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | | | | |
| 2 | 1 | 0 | +1 | 1 | 1 | 4 | 4 | |
| 3 | 2 | 1 | +1 | 1 | 2 | 4 | 4 | |
| 4 | 2 | 1 | +1 | 1 | 3 | 4 | 4 | |
| 5 | 1 | 0 | +1 | 1 | 4 | 4 | 4 | |
| 6 | 2 | 1 | +1 | 1 | 5 | 4 | 4 | |
| 7 | 1 | 2 | −1 | 1 | 6 | 4 | | 4 |
| 8 | 1 | 2 | −1 | 1 | 7 | 4 | | 4 |
| 9 | 3 | 1 | +2 | 2 | 8 | 8 | 8 | |
| | | | | | Sumas | 36 | 28 | 8 |

*Los números en negrita corresponden a la pareja de gemelos donde la diferencia fue negativa, por lo que van a formar parte de la columna de R–.*
*dif = gemelo vivo – gemelo muerto; dif (abs) = |dif|; n.° de orden = secuencia de |dif| sin corregir los empates; rangos = secuencia de |dif| una vez corregidos los empates; R+ = rangos positivos; R– = rangos negativos.*

ha codificado la práctica deportiva en escala *ordinal*, y se asigna un 0 a los sedentarios, un 1 a los que realizaban esporádicamente deporte en el tiempo libre, un 2 a los que lo realizaban regularmente, y un 3 a los que hacían deporte competitivamente y estaban sometidos a entrenamiento (tabla 6.6).

En primer lugar, se *hallarán las diferencias entre cada par de individuos relacionados*, lo que equivale a lo realizado en el test *t* emparejado. No obstante, en este caso se despreciarán las diferencias que sean iguales a 0. En el ejemplo, solo la primera diferencia vale 0. El tamaño de muestra ($n$), para todos los efectos a partir de ahora, es solo el número de parejas cuya diferencia es distinta de 0. En nuestro caso son ocho parejas ($n = 8$).

A continuación se ordenan las diferencias (en valor absoluto) de menor a mayor. Es decir, no se tiene en cuenta el signo (+ o –) de las diferencias para ordenarlas. Después, se asignan rangos a cada dato (usando su valor absoluto), pero se debe aplicar la corrección por empate; es decir, a cada uno se le asigna un número de orden y a los que estén empatados se les asigna el rango medio de los que tengan el mismo valor absoluto de la diferencia. Por ejemplo, si están empatados los puestos 1, 2 y 3, se les asignará el 2 a todos ellos. En el ejemplo están empatados los puestos 1 al 7, y se les ha asignado el valor 4 a todos. A continuación se suman los rangos de las diferencias negativas, por un lado, y los de las diferencias positivas, por otro, y se calculan por separado las sumas y medias de rangos para los valores positivos y negativos. En el ejemplo, la suma de los rangos con signo positivo es 28 y la de los rangos que tienen signo negativo es 8. La suma total de rangos es 36. Se comprueba que debe ser así, ya que:

$$\text{Suma de rangos} = n(n+1)/2 = 8(9)/2 = 72/2 = 36 = 28+8$$

Si el *tamaño de muestra efectivo fuese ≥ 25, se calculará un test «z»*. Para ello se elige una de las dos sumas de rangos, por ejemplo R+ = 28, y se aplica la siguiente expresión:

$$z = \frac{\sum(\text{R}+) - (n(n+1)/4)}{\sqrt{n(n+1)(2n+1)/24}}$$

Como ya se ha dicho, el valor $n$ es el correspondiente al número de parejas cuya diferencia entre los datos emparejados (después-antes) es distinta de 0. En el ejemplo serán ocho parejas.

Si fuese verdad la hipótesis nula (*dif* = 0), la suma de rangos positivos sería igual a la suma de rangos negativos, y ambas deberían valer la mitad de la suma total de rangos. Como la suma total de rangos es $n(n + 1)/2$, lo esperado bajo $H_0$ será:

$$\Sigma(\text{R}+) = \Sigma(\text{R}-) = n(n+1)/4 \text{ (si } H_0 \text{ fuese cierta)}$$

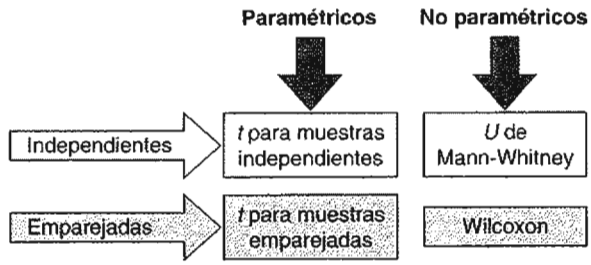



**Figura 6.10** Comparaciones de dos medias.

Así, se entiende que el numerador de $z$ sea la diferencia entre lo observado y lo esperado (si $H_0$ fuese cierta). El denominador viene a ser análogo al error estándar.

Aplicando esta expresión, el valor $z$ para el ejemplo será:

$$z = \frac{28-(8\times 9)/4}{\sqrt{8(9)(17)/24}} = \frac{28-18}{\sqrt{3\times 17}} = \frac{10}{7{,}14} = 1{,}4$$

que no permitirá rechazar $H_0$, ya que la significación estadística sería $p = 0{,}16$. En STATA:

```
. di 2*normal(-1.4)
.16151332
```

Con muestras pequeñas (<25 parejas), este test debe hacerse con ordenador.

Ahora ya se pueden valorar las principales alternativas para realizar comparaciones de dos medias (figs. 6.10 y 6.11).

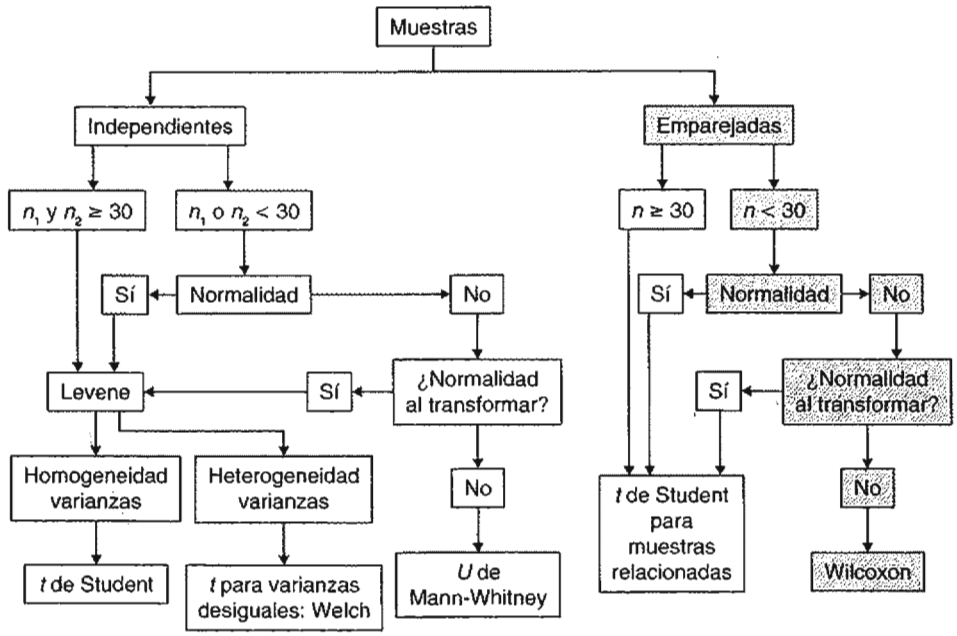



**Figura 6.11** Algoritmo completo para decidir cómo comparar dos medias. Se ha sombreado la trayectoria que conduce a decidir usar el test de Wilcoxon.

Resumiendo, la comparación puede ser de muestras independientes o de muestras emparejadas. A su vez, los métodos pueden ser paramétricos o no paramétricos. La combinación de estos dos criterios proporciona cuatro posibles análisis estadísticos principales.

## 6.10. TEST PARA MEDIAS CON STATA

Para la *t* de Student con varianzas homogéneas, y usando el primer ejemplo (dieta y cambio de peso), se procedería así:

```
. ttest cambpes, by(gr)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 1 | 19 | 18.57895 | 2.125198 | 9.263523 | 14.11407 | 23.04382 |
| 2 | 11 | 17.54545 | 2.541881 | 8.430464 | 11.88179 | 23.20912 |
| combined | 30 | 18.2 | 1.612737 | 8.833322 | 14.90158 | 21.49842 |
| diff | | 1.033493 | 3.400293 | | -5.931691 | 7.998677 |

```
    diff = mean(1) - mean(2)                                  t =   0.3039
Ho: diff = 0                     degrees of freedom =       28

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6183         Pr(|T| > |t|) = 0.7634          Pr(T > t) = 0.3817
```

Como puede observarse, STATA incluye los intervalos de confianza. La opción, **level(90)** proporcionaría intervalos de confianza al 90%.

Si se asume que las varianzas son heterogéneas, se pedirá el test de Welch del modo siguiente:

```
. ttest cambpes, by(gr) welch

Two-sample t test with unequal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 1 | 19 | 18.57895 | 2.125198 | 9.263523 | 14.11407 | 23.04382 |
| 2 | 11 | 17.54545 | 2.541881 | 8.430464 | 11.88179 | 23.20912 |
| combined | 30 | 18.2 | 1.612737 | 8.833322 | 14.90158 | 21.49842 |
| diff | | 1.033493 | 3.31325 | | -5.793251 | 7.860237 |

```
    diff = mean(1) - mean(2)                                  t =   0.3119
Ho: diff = 0                     Welch's degrees of freedom =  24.7867

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6211         Pr(|T| > |t|) = 0.7577          Pr(T > t) = 0.3789
```

El lector puede identificar cada uno de los resultados antes vistos, con la salvedad de que STATA contrasta tres hipótesis alternativas (Ha): la bilateral (a dos colas), que está en la parte central del listado (es la que se deberá elegir), y las correspondientes a cada cola.

Si se elige la opción **unequal** en vez de **welch**, STATA calculará los grados de libertad por un procedimiento alternativo (método de Satterthwaite), con pequeñas diferencias con respecto al test de Welch.

Para la *t* de una sola variable comparando su media respecto a un valor externo, se actuará del modo siguiente:

```
. ttest colester==200

One-sample t test
```

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| colester | 25 | 208 | 3.993745 | 19.96873 | 199.7573 | 216.2427 |

```
    mean = mean(colester)                                     t =   2.0031
Ho: mean = 200                                 degrees of freedom =       24

   Ha: mean < 200                Ha: mean != 200                 Ha: mean > 200
Pr(T < t) = 0.9717         Pr(|T| > |t|) = 0.0566          Pr(T > t) = 0.0283
```

Si se desea aplicar el test no paramétrico de la *U* de Mann-Whitney al ejemplo de las dos dietas para perder peso, se procederá así:

```
. ranksum cambpes, by(gr) porder

Two-sample Wilcoxon rank-sum (Mann-Whitney) test
```

| gr | obs | rank sum | expected |
|---|---|---|---|
| 1 | 19 | 292 | 294.5 |
| 2 | 11 | 173 | 170.5 |
| combined | 30 | 465 | 465 |

```
unadjusted variance      539.92
adjustment for ties       -1.08
                         ------
adjusted variance        538.84

Ho: cambpes(gr==1) = cambpes(gr==2)
                 z =  -0.108
    Prob > |z| =   0.9142

P{cambpes(gr==1) > cambpes(gr==2)} = 0.488
```

Desafortunadamente, STATA llama «Two-sample Wilcoxon rank-sum» al test de Mann-Whitney, aunque lo arregla de alguna manera al escribir luego el nombre (Mann-Whitney) entre paréntesis. Al incluir en la línea de instrucciones la opción **porder,** como se ha hecho arriba, STATA devuelve la interpretación de la *U* como la proporción (0,488, en el ejemplo) de comparaciones en que un grupo superó al otro.

Para la *t* emparejada con el ejemplo de la diferencia en los tiempos de reacción entre antes y después de consumir alcohol, se pedirá a STATA lo siguiente:

```
. ttest antes=desp

Paired t test
```

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| antes | 15 | 15.06667 | .7397983 | 2.865227 | 13.47996 | 16.65338 |
| desp | 15 | 11.86667 | .6609277 | 2.559762 | 10.44912 | 13.28422 |
| diff | 15 | 3.2 | .9421354 | 3.648875 | 1.179321 | 5.220679 |

```
    mean(diff) = mean(antes - desp)                           t =   3.3965
Ho: mean(diff) = 0                             degrees of freedom =       14

Ha: mean(diff) < 0            Ha: mean(diff) != 0             Ha: mean(diff) > 0
Pr(T < t) = 0.9978         Pr(|T| > |t|) = 0.0043          Pr(T > t) = 0.0022
```

Pueden reconocerse en la salida los cálculos que antes se han explicado.
Para el test de Wilcoxon en este mismo ejemplo:

```
. signrank antes = desp

Wilcoxon signed-rank test

      sign |      obs   sum ranks    expected
-----------+---------------------------------
  positive |       11       104.5        58.5
  negative |        2        12.5        58.5
      zero |        2           3           3
-----------+---------------------------------
       all |       15         120         120

unadjusted variance        310.00
adjustment for ties         -0.88
adjustment for zeros        -1.25
                           ------
adjusted variance          307.88

Ho: antes = desp
            z =   2.622
    Prob > |z| =   0.0088
```

Puede observarse que el valor *p* suele resultar inferior (más significativo) con el test *t* emparejado (paramétrico) que con el test no paramétrico de Wilcoxon.

## 6.11. TEST PARA MEDIAS CON OTROS PROGRAMAS

### 6.11.1. Test para medias con SPSS

Se pedirá así el test de la *t* para varianzas homogéneas:

```
T-TEST GROUPS=gr(1 2)
/VAR=cambpes
```

Una ventaja de SPSS es que la salida es muy completa y proporciona ya directamente el test de homogeneidad de varianzas, aunque no se le pida. Además, siempre proporciona dos opciones de resultados, una con varianzas homogéneas en la fila superior y otra con varianzas heterogéneas (test de Welch) en la fila inferior. Cuando se vea que el test de Levene es significativo (varianzas heterogéneas), se elegirá la fila inferior.

La *t* emparejada se solicitará así:

```
T-TEST PAIRS=antes WITH desp (PAIRED)
```

Los test no paramétricos se solicitan del modo siguiente:
Para la *U* de Mann-Whitney se escribirá la sintaxis:

```
NPAR TESTS
/M-W= cambpes BY gr(1 2)
```

Una ventaja de la *U* de Mann-Whitney en SPSS es que proporciona también la significación por un método exacto usando la distribución binomial.

Se pedirá el test Wilcoxon como:

```
NPAR TESTS
/WILCOXON=antes WITH desp (PAIRED)
```

### 6.11.2. Test para medias con R/Splus

En primer lugar se introducen los valores de cada grupo. En el ejemplo, se llamará «a» = dieta 1 y «b» = dieta 2. Después debe usarse la expresión **t.test** (*a,b,* **var.equal = TRUE**) para varianzas homogéneas, y se escribirá simplemente **t.test** (*a,b*) para varianzas heterogéneas.

```
> a<-c(2,9,10,10,12,13,15,15,16,17,18,19,19,21,27,28,29,35,38)
> b<-c(4,5,12,14,18,19,20,22,23,23,33)
> t.test (a,b, var.equal = TRUE) #proporcionat para muestras homogéneas
> t.test (a,b) #proporciona test de Welch
```

Para obtener una *t* emparejada, se introducirá como vectores separados cada una de los dos variables o mediciones. A continuación se usa la expresión **t.test,** seguida de un paréntesis en el que se indican los nombres de las dos variables, separadas por una coma. Tras otra coma se indicará que es emparejado mediante la opción **(paired = TRUE).**

```
>antes<-c(10,10,12,13,14,15,16,16,16,16,16,17,16,19,20)
>desp<-c(14,12,12,12,8,12,9,10,16,12,11,15,8,11,16)
> t.test(antes, desp, paired=T)
```

R/Splus no proporciona automáticamente los intervalos de confianza al hacer la *t*.
Para pedir el test Mann-Whitney con R/Splus deberá usarse la siguiente expresión:

```
>wilcox.test(a, b, paired = FALSE)
```

De nuevo, este programa no dice nada de la *U* de Mann-Whitney, sino de test de Wilcoxon no emparejado, de ahí la opción paired = FALSE. R/Splus ofrece una salida muy escueta. Además, el valor *p* que proporciona (*p* = 0,9313) es a dos colas y se le aplica una corrección por continuidad similar a la corrección de Yates en el test de la ji cuadrado. La cantidad que aparece, *W* = 102, corresponde a la menor de las dos cantidades *U* que se calculan en el test de Mann-Whitney.

Para el test de Wilcoxon con R/Splus, como ya se habrá deducido, deberá usarse la siguiente expresión:

```
> wilcox.test(antes, desp, paired = T)
```

Sin embargo, no funcionará si no se retiran antes las parejas con diferencias iguales a 0. Además, el listado de salida que devuelve R/Splus es menos completo y penaliza menos por empates y por ceros que STATA y SPSS. El valor de V para R/Splus es la suma de rangos que tiene mayor valor y que se puede comprobar en la tabla que antes hemos presentado cuando la muestra es pequeña.

## REFERENCIAS

1. Altman DG. Practical statistics for medical research. Londres: Chapman and Hall, 1991.
2. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23(1):151-69.
3. Bland JM, Altman DG. The use of transformation when comparing two means. BMJ 1996;312:1153.
4. Guallar E, Jiménez FJ, van 't Veer P, Bode P, Riemersma RA, Gómez-Aracena J, et al. Low toenail chromium concentration and increased risk of nonfatal myocardial infarction. Am J Epidemiol 2005;162(2):157-64.
5. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. En: Bailar JC III, Hoaglin DC, editors. Medical uses of statistics. 3rd ed. New Jersey: John Wiley and Sons, 2009. p. 259-79.
6. Sprent P. Applied nonparametric statistical methods. Londres: Chapman & Hall, 1996.

# 7 ESTIMACIÓN DEL TAMAÑO MUESTRAL

*M. Á. Martínez-González, M. Ruiz-Canela, F. Guillén-Grima*

## 7.1. INTRODUCCIÓN

Un aspecto clave antes de emprender una investigación es saber *qué tamaño* debe tener el estudio para poder valorar correctamente la hipótesis que se desea estudiar. La envergadura del estudio condicionará las necesidades de personal y de recursos, y la duración del trabajo. Por eso es imprescindible saberlo de antemano. Generalmente, el investigador ha decidido mucho antes cuál será el tamaño aproximado de su estudio, basándose en la factibilidad, el presupuesto disponible y, muy probablemente, también en el tamaño de los estudios previos sobre el tema (**1-3**). Pero esto no basta. Se requiere formalizar de la manera más exacta posible los diversos supuestos y adelantarse así de algún modo a los resultados que se obtendrán. Difícilmente se podrá lograr financiación para un proyecto de investigación si no se aporta una justificación bien documentada y calculada del tamaño del estudio que esté basada en supuestos realistas.

## 7.2. MARGEN DE ERROR

Se llamará aquí *margen de error (M)* a la amplitud que se espera o desea para *cada una* de las dos mitades simétricas de un intervalo de confianza. Esta definición se aplica tanto a la estimación de una media como a la de una proporción.

Por ejemplo, si al estimar la media de la presión arterial sistólica se espera que sea de 120 mmHg y su intervalo de confianza del 95% esté comprendido entre 118 y 122 mmHg, entonces M = 2, ya que:

$$M = \frac{LSC - LIC}{2} = \frac{122 - 118}{2} = 2$$

donde LSC y LIC son los límites superior e inferior de confianza.

El margen de error (M) se interpreta como una medida de la separación del intervalo de confianza con respecto a la media encontrada (M = ± 2 mmHg en el ejemplo).

## 7.3. ESTIMACIÓN DE UNA PROPORCIÓN

Podría pensarse que el tamaño muestral sería mayor al estimar la proporción de europeos con diabetes que si quisiera conocerse la misma proporción entre los navarros. Sin embargo, no sucede así. Aunque a primera vista parezca chocante, podrían buscarse ejemplos intuitivos para entenderlo. El cocinero que está preparando una sopa y para probarla toma siempre una cucharada. Hará lo mismo cuando la sopa sea para cinco personas o para un centenar. No se le ocurrirá tomar 20 cucharadas de prueba porque esté preparando más sopa (**4**). Como se verá después en las fórmulas del cálculo del tamaño muestral, *no* se considera el tamaño de la población de donde se obtiene la muestra.

Por ejemplo, supóngase que se desea conocer la proporción de españoles con obesidad. En esta situación, para el estudio se espera una prevalencia del 25% y se desea un *margen de error* (M) para el intervalo de confianza del 95% de ±1% (M = 0,01); es decir, se espera que el resultado sea una proporción igual a 0,25 (IC 95%: 0,24 a 0,26). ¿Cuántos sujetos deben incluirse en la muestra?

Para dar respuesta a esta pregunta se parte del intervalo de confianza de una proporción:

$$IC(\pi) = p \pm \left( z_{\alpha/2} \sqrt{\frac{pq}{n}} \right) = p \pm M$$

Lo que figura dentro del paréntesis es el *margen de error* (M). Por tanto:

$$M = z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

Se despeja n:

$$n = \frac{z^2_{\alpha/2} pq}{M^2}$$

Los supuestos eran:

$z$ = 1,96 (se desea un IC del 95%).

$p$ = 0,25 (y, por tanto, $q$ = 0,75).

M = 0,01.

De esta forma, de la fórmula anterior se obtendrá el resultado siguiente:

$$n = \frac{z^2_{\alpha/2} pq}{M^2} = \frac{1{,}96^2 \times 0{,}25 \times 0{,}75}{0{,}01^2} = 7.203$$

Se necesitarán, por tanto, 7.203 sujetos en la muestra. Puede comprobarse la precisión del resultado, pues el 25% de 7.203 es 1.801 y, si se calcula el intervalo de confianza del 95%, se obtendrá exactamente 0,24-0,26.

La lógica perplejidad en este momento surge de la pregunta sobre cómo obtener $p$ ($p$ = 0,25 en el ejemplo) si el estudio se inicia precisamente porque se desea estimar tal proporción, que es desconocida. Hay tres modos de hacerlo:

1. Buscar bibliografía y consultar estudios similares.
2. Realizar un pequeño estudio piloto con pocos sujetos para tener una idea sobre $p$.
3. Elegir el caso que necesita mayor tamaño muestral $n$, que es $p = q = 0{,}5$, ya que maximiza el producto $pq$. Si se supone que $p$ = 0,50, a no ser que se esté en el peor de los casos, seguro que *sobra* tamaño muestral.

## 7.4. ESTIMACIÓN DE UNA MEDIA

Siguiendo la metodología anterior, tendremos el resultado siguiente:

$$IC(\mu) = \bar{x} \pm \left( z_{\alpha/2} \frac{s}{\sqrt{n}} \right) = \bar{x} \pm M$$

$$M = z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Así, si se despeja $n$ de la segunda expresión, se obtendrá:

$$n = \frac{z^2_{\alpha/2} s^2}{M^2}$$

Imagínese que se desea estimar la media de la presión arterial sistólica y los supuestos eran:

$z$ = 1,96 (se desea un IC del 95%).

Media esperada = 120 mmHg.

Desviación estándar esperada = 8 mmHg.

M = 2 mmHg.

Se necesitarán 62 sujetos en la muestra.

$$n = \frac{z^2_{\alpha/2} s^2}{M^2} = \frac{1{,}96^2 \times 8^2}{2^2} = 62$$

Se habrá advertido que se usa $z$ en vez de $t$. Se debe a que, para introducir $t$ en la última fórmula, se necesitaría saber antes los grados de libertad que, a su vez, dependen de $n$ y por, tanto, se desconocen. Una vez hallado $n$, se podría recalcular todo usando $t$ con los grados de libertad $n - 1$. No osbtante hay que considerar que la estimación del tamaño muestral es siempre aproximada, pues depende de que posteriormente se cumplan los supuestos, que no dejan de ser una mera apuesta.

## 7.5. COMPARACIÓN DE DOS PROPORCIONES

Como ya se explicó en el apartado 4.13 (v. fig. 4.13), el planteamiento *a priori* de un contraste de hipótesis a dos colas puede representarse por una figura con dos campanas de Gauss: una mantiene que la hipótesis nula ($H_0$) es cierta y la otra sostiene que la hipótesis alternativa ($H_1$) es cierta. Así se ha representado de nuevo en la figura 7.1, en referencia a una diferencia de proporciones, aunque ahora la hipótesis alternativa mantiene que la diferencia $d$ de proporciones es *menor* que 0. Por esta razón, la campana formada por las posibles muestras que podrían extraerse de una población donde $H_1$ fuese cierta queda a la izquierda (es decir, por debajo) de la campana de las posibles muestras que podrían extraerse de una población en la que $H_0$ fuese cierta.

$H_0$ mantiene que la diferencia de proporciones poblacional vale 0. La campana de la derecha (bajo $H_0$) representa cómo se distribuirán las diferencias de proporciones en todas las posibles

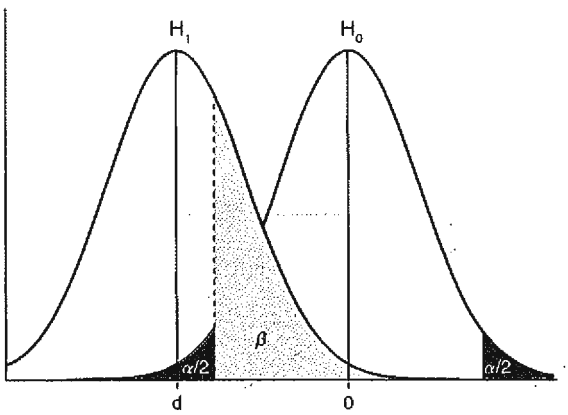



**Figura 7.1** Planteamiento del tamaño muestral en un contraste de hipótesis. La distancia entre ambas hipótesis tiene un valor $d$, que se puede descomponer en dos segmentos en función del error estándar (EE): $d = z_{\alpha/2}EE + z_{\beta}EE$. La línea vertical discontinua marca el criterio para la decisión que se tomará a favor de una u otra hipótesis. Solo si la muestra cae a la izquierda de esa línea discontinua se rechazará la hipótesis nula.

muestras tomadas de una población en la que $H_0$ fuese verdad. Las diferencias de proporciones calculadas en esas muestras seguirán una distribución normal centrada en 0. Las zonas sombreadas en esa campana corresponden a $\alpha/2$ a cada lado. Cuando el estimador muestral esté más lejos de $z$ errores estándar a un lado o a otro de la campana, se rechazará $H_0$, pero se puede haber cometido un error tipo 1. Las dos zonas pequeñas sombreadas en color más oscuro a cada lado de la campana bajo $H_0$ representan el porcentaje de muestras que proceden de una población donde $H_0$ es verdad y, sin embargo, conducen equivocadamente a rechazar $H_0$. La mitad de esas posibles muestras está en cada una de las dos colas.

$H_1$ (campana de la izquierda en la figura 7.1) mantiene que la verdadera diferencia poblacional es una cantidad $d$ que está por debajo de $H_0$. La campana bajo $H_1$ asume que $H_1$ es verdad en la población y representa cómo se distribuirán las posibles muestras tomadas de esa población. Habrá algunas de esas muestras que obtendrán diferencias de proporciones muy cercanas a 0 y llevarán a no poder rechazar $H_0$, aunque se hayan obtenido de una población en la que $H_1$ fuese cierta (pertenecen a esa campana). Serán errores tipo 2. Su probabilidad (riesgo $\beta$) es el área bajo la campana de la izquierda que queda sombreada en gris. El riesgo $\beta$ es siempre a una cola.

La métrica para moverse en este gráfico se basa en que las distancias están medidas en unidades de *error estándar* de una diferencia de proporciones (EEDP).

La distancia hacia la izquierda desde 0 hasta $d$ se descompone en dos fragmentos.

1. Desde 0 hasta el comienzo del área $\alpha/2$ de la izquierda: $z_{\alpha/2}$EEDP.
2. Desde el comienzo del área $\alpha/2$ de la izquierda hasta $d$: $z_{\beta}$EEDP.

Por tanto, la distancia total ($d - 0 = d$) es la suma de las dos:

$$d = (z_{\alpha/2}\text{EEDP}) + (z_{\beta}\text{EEDP}) = (z_{\alpha/2} + z_{\beta})\text{EEDP}$$

Es conocido ya el valor del EEDP:

$$\text{EEDP} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

Si se supone que los dos grupos tendrán igual tamaño (que suele ser lo ideal), entonces $n_1 = n_2$, y se usará $n$ para el tamaño de *cada* grupo:

$$\text{EEDP} = \sqrt{\frac{2pq}{n}}$$

Por lo que:

$$d = (z_{\alpha/2} + z_{\beta})\sqrt{\frac{2pq}{n}}$$

Despejando $n$, surge la fórmula del tamaño muestral (**5,6**):

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 \times 2pq}{d^2}$$

*Ejemplo:* en un estudio de seguimiento a 5 años, se desea comparar la proporción de ocurrencia de depresión en dos grupos de jóvenes según estuviesen inicialmente expuestos o no a *binge-drinking* los fines de semana.

Se espera que los resultados sean:

$P_{\text{expuestos}} = 10\%$

$P_{\text{no expuestos}} = 4\%$

Se está dispuesto a asumir los riesgos:

$\alpha$ bilateral del 5%.

$\beta$ (siempre es unilateral) del 10% (potencia, 90%).

La proporción $p$ que se escribirá en la fórmula es la media [(0,1 + 0,04)/2 = 0,07] de las 2 proporciones presentadas; la diferencia entre ellas *(d)* es 0,06 (0,10 – 0,04 = 0,06) y aparece en el denominador.

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2pq}{d^2} = \frac{(1{,}96 + 1{,}28)^2 \times 2 \times 0{,}07 \times 0{,}93}{0{,}06^2} = 380$$

Se necesitarán en total 760 sujetos, 380 en cada grupo.

Parece superfluo a estas alturas explicar por qué 1,96 es el valor de $z_{\alpha/2}$, pero quizá surjan dudas sobre la razón por la cual $z_\beta$ vale 1,28. Vale 1,28 porque $\beta$ es siempre unilateral, y para $z$ = 1,28 el área que queda en la cola es 0,10. Téngase en cuenta que si se pide a STATA:

```
display normal(-1.28)   #devolverá 0,10,
```

y si se le indica:

```
display invnormal(0.9) #devolverá 1,2816.
```

Cuando se desee una potencia del 80%, $z_\beta$ valdrá 0,84.

## 7.6. COMPARACIÓN DE DOS MEDIAS

Es conocido que la varianza de una proporción es el producto $pq$. Por eso, la fórmula anterior puede adaptarse para el tamaño muestral de la diferencia entre dos medias simplemente sin más que usar $s^2$ para reemplazar a $pq$.

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2s^2}{d^2}$$

*Ejemplo:* en un estudio que desea comparar la pérdida de peso con dieta mediterránea (MeDiet) o en grupo control se espera que los resultados sean:

Media cambio peso $_{\text{MeDiet}}$ = –5 kg ($s$ = 12).

Media cambio peso $_{\text{control}}$ = –1 kg ($s$ = 12).

Se está dispuesto a asumir los riesgos:

$\alpha$ bilateral del 5%.

$\beta$ (siempre es unilateral) del 20% (potencia, 80%).

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2s^2}{d^2} = \frac{(1{,}96 + 0{,}84)^2 \times 2 \times 12^2}{4^2} = 142$$

Es una *sana* costumbre redondear hacia arriba el resultado del cálculo del tamaño muestral. Se asegura así no quedarse corto.

Aquí se necesitarían 284 en total, 142 en cada grupo.

Un atajo aproximado para esta fórmula (7), siempre que se desee asumir $\alpha$ a dos colas = 5% y $\beta$ = 20%, consiste en usar el cociente $d/s$, en este caso 1/3 (4/12), y aplicar:

$$n = \frac{16}{(d/s)^2} = \frac{16}{\left(\frac{1}{3}\right)^2} = 144$$

## 7.7. CÁLCULO DE LA POTENCIA ESTADÍSTICA

Se despeja así $z_\beta$ de la fórmula usada en el apartado anterior.

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2s^2}{d^2} \Rightarrow (z_{\alpha/2} + z_\beta)^2 = \frac{nd^2}{2s^2}$$

Por tanto:

$$z_\beta = \sqrt{\frac{nd^2}{2s^2}} - z_{\alpha/2}$$

Un ejemplo inspirado en la referencia bibliográfica (**8**) consistiría en calcular la potencia para comparar las puntuaciones del *minimental test* entre un grupo ($n$ = 180) asignado a una dieta mediterránea y un grupo control ($n$ = 180). Los supuestos son:

Riesgo $\alpha$: 0,05 bilateral.

Media (DE) en MeDiet: 28 (3).

Media (DE) en grupo control: 27 (3).

$$z_\beta = \sqrt{\frac{nd^2}{2s^2}} - z_{\alpha/2} = \sqrt{\frac{180 \times 1^2}{2 \times 3^2}} - 1{,}96 = 1{,}2$$

Este valor se consulta en la distribución normal y se obtendrá la potencia; por ejemplo, en STATA:

```
display normal(1.2)        #devolverá 0,8849
```

Conclusión: el estudio planteado tiene una potencia >88% (riesgo $\beta < 0{,}12$).

Análogamente se puede proceder del mismo modo para una comparación de proporciones sin más que sustituir $s^2$ por $pq$.

El siguiente ejemplo se refiere a proporciones y está inspirado en la referencia (**9**). En un ensayo preventivo, se prevé que se terminará por diagnosticar cáncer de próstata en 1.000 de 9.500 asignados a finasterida y en 1.400 de los 9.500 asignados a placebo. Las proporciones son: 1.000/9.500 = 0,1053 y 1.400/9.500 = 0,1474. La proporción media $p$ es 2.400/19.000 = 0,1263. Por tanto, $q$ = 1 – 0,1263 = 0,8737. Se asume $\alpha$ = 5% a dos colas.

$$z_\beta = \sqrt{\frac{nd^2}{2pq}} - z_{\alpha/2} = \sqrt{\frac{9.500 \times (0{,}1474 - 0{,}1053)^2}{2 \times 0{,}1263 \times 0{,}8737}} - 1{,}96 = 6{,}8$$

Este valor se mira en la normal y se obtendrá la potencia. Por ejemplo, en Excel:

=DISTR.NORM.ESTAND(6,8)  #devolverá 1.

*Conclusión:* el estudio planteado tiene una potencia del 100%.

Tutoriales específicos (**10,11**) y la edición previa de este manual (**12**) contienen explicaciones más detalladas y otros aspectos del cálculo de tamaño muestral.

## 7.8. CURVAS DE POTENCIA

Un modo elegante de adelantar la potencia que tendrá un estudio consiste en valorar diversas posibilidades o escenarios para comprobar cómo se comportará la potencia del estudio ante diversas variaciones sobre los supuestos que se han asumido. Se asemejaría a un *análisis de*

*sensibilidad*, consistente en averiguar cómo variarán los resultados si cambiasen los supuestos que se han asumido.

Por ejemplo, para una comparación de dos proporciones, la figura 7.2 presenta la variación de la potencia estadística en función de diversos escenarios esperables para el estudio. Se asumen como constantes el riesgo $\alpha$ del 5% a dos colas ($\alpha/2 = 0{,}025$) y una proporción media *(p)* de $p = 0{,}20$. Es decir, se espera siempre que el 20% de los sujetos en total tendrán un episodio o evento.

Después se plantean dos posibles tamaños de muestra: la primera opción con 200 sujetos en total, $n = 100$ en cada grupo, y la segunda opción con 100 en total, $n = 50$ en cada grupo. Se calcula la potencia en estas dos opciones para todas las posibles diferencias $d$ entre las dos proporciones en el rango que va desde $d = 0{,}08$ hasta $d = 0{,}33$.

En el estudio de la figura 7.2 puede apreciarse que, con un tamaño muestral de 50 en cada grupo, no se alcanzará una potencia estadística del 80% a no ser que la diferencia $d$ entre ambos grupos sea al menos $d = 0{,}225$. Con el doble tamaño ($n = 100$) se tendrá >80% de potencia solo para diferencias entre grupos del 17% o superiores ($d \geq 0{,}17$).

Las órdenes del *Do-file* de STATA para componer la figura 7.2 fueron:

```
clear
set obs 100
g dif=0.0775+(_n/400)
scalar zalfa= invnormal(1-(0.05/ 2 ))
g p = 0.2
g zbeta1 = ((100*((dif)^2) / (2*p*(1-p)) )^0.5) - zalfa
g zbeta2 = ((50*((dif)^2) / (2*p*(1-p)) )^0.5) - zalfa
g potencia1 = normal(zbeta1)
lab var potencia1 "n cada grupo=100"
g potencia2 = normal(zbeta2)
lab var potencia2 "n cada grupo=50"
twoway (line potencia1 dif,lcolor(red) lwidth(medthick)) ///
(line potencia2 dif,lc(black) lw(medthick) lpattern(dash)), ///
title("Potencia según diferencia esperada") ///
subtitle("(alfa/2=0.025 y % de eventos=20%)") ///
ytitle(Potencia) xtitle("Diferencia (pA-pB) entre grupos") ///
xlabel(0.1(0.05)0.3, grid) ylabel(0(0.1)1, grid)
```

La variable ***dif*** va creciendo en intervalos de 0,0025, desde 0,08 hasta 0,3275.

La orden **scalar**, que es como **generate** (abreviado aquí como **g**), sirve para crear una *constante* que queda disponible para usarla a continuación.

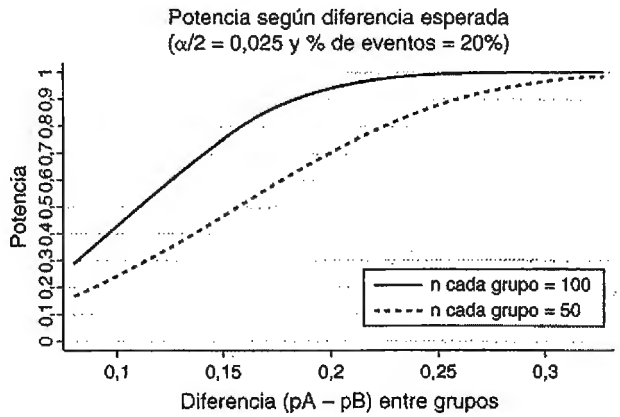



**Figura 7.2** Curvas de potencia para una comparación de dos proporciones (pA y pB). Se han asumido como constantes $\alpha$ y p. Se valoran los escenarios con diferencias ($d$ = pA – pB) entre 0,08 y 0,33, y con dos posibles tamaños de muestra.

## 7.9. USO DE STATA PARA ESTIMAR EL TAMAÑO MUESTRAL Y LA POTENCIA

El programa STATA facilita responder a las preguntas formuladas en este capítulo sobre comparaciones de proporciones o medias y sobre potencia estadística. STATA es más flexible en cuanto a los supuestos. Hasta ahora se ha asumido siempre que los grupos comparados tendrán el mismo tamaño (razón 1:1). Tiene su lógica, ya que se obtiene más partido de los datos cuando los grupos de comparación poseen el mismo tamaño. No obstante, a veces puede ser interesante que un grupo tenga mayor tamaño que otro, por ejemplo, que el tamaño de un grupo sea 1,5 veces mayor que el de otro por razones de costes y factibilidades. Se dirá entonces que los tamaños están en razón 1,5:1. Al comparar medias se ha supuesto también que las desviaciones estándar de los dos grupos serán siempre iguales, pero puede que esto no sea asumible. STATA calcula el tamaño también cuando se asume que los grupos son de distinto tamaño y las varianzas son diferentes.

Con la orden **sampsi** en STATA se obtendrán $n_1$ y $n_2$. Es preciso escribir después las dos proporciones. Por omisión, STATA incrementa ligeramente *n* debido a una corrección por continuidad. Se recomienda suprimir tal corrección con la opción:

```
, nocontinuity
```

Para el ejemplo anterior de *binge-drinking* y depresión:

```
. sampsi .1 .04, nocontinuity

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:

         alpha =   0.0500   (two-sided)
         power =   0.9000
            p1 =   0.1000
            p2 =   0.0400
         n2/n1 =   1.00

Estimated required sample sizes:

            n1 =      378
            n2 =      378
```

Si se deseasen distintos tamaños, por ejemplo, con un grupo doble que el otro, se usará **ratio:**

```
. sampsi .1 .04, nocont ratio(2)

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:

         alpha =   0.0500   (two-sided)
         power =   0.9000
            p1 =   0.1000
            p2 =   0.0400
         n2/n1 =   2.00

Estimated required sample sizes:

            n1 =      275
            n2 =      550
```

Para comparar medias, se obtendrá *n* con la misma orden **sampsi** en STATA seguida de los valores de las medias de cada grupo; como opción, se añadirán sus desviaciones estándar (entre paréntesis y precedidas de las expresiones **sd1** y **sd2**). Por omisión, el programa usa un error $\beta$ de 0,1 (potencia = 90%). Para una potencia del 80%, se añade la opción **power*(.80)***.

El ejemplo del cambio de peso daría el resultado siguiente:

```
. sampsi -5 -1, sd1(12) sd2(12) power(.8)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =    0.0500  (two-sided)
         power =    0.8000
            m1 =        -5
            m2 =        -1
           sd1 =        12
           sd2 =        12
         n2/n1 =      1.00

Estimated required sample sizes:

            n1 =       142
            n2 =       142
```

Se obtendría lo mismo con:

```
sampsi 0 4, sd(12) p(.8)
```

Para obtener la potencia se usa la misma orden, pero indicando el valor de los tamaños muestrales con las opciones **n1** y **n2**, o bien **n**, si ambas muestras son de igual tamaño.

En el ejemplo del *minimental test* según asignación a dieta mediterránea o a grupo control, la potencia se pediría del modo siguiente:

```
. sampsi 27 28, sd(3) n(180)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                  and m2 is the mean in population 2
Assumptions:

         alpha =    0.0500  (two-sided)
            m1 =        27
            m2 =        28
           sd1 =         3
           sd2 =         3
sample size n1 =       180
            n2 =       180
         n2/n1 =      1.00

Estimated power:

         power =    0.8854
```

Como se ha indicado ya **n**, STATA entiende que no tiene que calcular este valor e intepreta que ahora lo que le interesa al investigador es la potencia. Se recomienda consultar el vídeo llamado «SAMPLE SIZE» en http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

## 7.10. PROGRAMACIÓN DE EXCEL PARA TAMAÑO MUESTRAL Y POTENCIA

Se pueden encontrar y descargar diversas calculadoras programadas en una hoja Excel en http://www.unav.es/departamento/preventiva/recursos_bioestadistica, en el archivo denominado: «Programa sencillo en Excel de cálculo tamaño muestral».

En esa hoja de Excel se han introducido pequeñas modificaciones para afinar más en los supuestos, en concreto:

- Se recalcula $n$ para la estimación de la media usando $t$ en vez de $z$.
- En el cálculo de $n$ para diferencia de proporciones, se asume que $z_\beta$ está bajo $H_1$ y, por eso, las proporciones será diferentes ($p_1$ y $p_2$), en vez de usar la media de ambas.

## 7.11. OTROS PROGRAMAS DISPONIBLES PARA EL CÁLCULO DEL TAMAÑO MUESTRAL

Existen diversas opciones de *software* específicas para estimar el tamaño muestral en diversos procedimientos y con distintas especificaciones.

Algunos de los programas son:

- *StatCalc:Epiinfo.* Es un módulo del paquete *estadistico epiinfo,* gratuito y muy utilizado para el diseño de encuestas, cohortes y estudios de casos y controles, y para ensayos clínicos. Es descargable desde http://wwwn.cdc.gov/epiinfo/ y desde http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize.
- *PS: Power and Sample size calculation.* Gratuito, muy utilizado y descargable desde http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize.
- POWER *V3.0.* Gratuito (Windows) y descargable desde http://dceg.cancer.gov/tools/design/power.
- *G*Power 3.* Gratuito, presenta siempre una gráfica del tipo de la 7.1. www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/.
- *PASS12.* Este programa es mucho más completo, pero la licencia cuesta más de 1.000 dólares. Más información en http://www.ncss.com/online-store/.
- *Siz.* Constituye la opción más coste-efectiva de entre los programas comerciales. El coste anual de la licencia en 2013 es de unos 150 €. Tiene más prestaciones que los anteriores (http://www.cytel.com/software/siz).
- *Power And Precision.* Programa comercial bastante completo, cuyo coste es de 600 dólares (versión académica), que asciende hasta 1.300 dólares para la versión general (http://www.power-analysis.com/).

### 7.11.1. Otros programas específicos de ensayos clínicos

- *nQuery + nTerim t.* Es un programa general, con opciones especificas para ensayos clínicos. Es utilizado por las principales agencias, como la *Food and Drug Administration* (FDA) y la Agencia Europea del Medicamento (EMA), y por la industria. Tiene distintas opciones, desde 75 € en una versión para estudiantes a 1.300 € en la versión general (http://www.statistical-solutions-software.com/nquery-advisor-nterim/).
- *East.* Se considera el programa estándar. Es específico únicamente para ensayos clínicos, utilizado por la FDA, la EMA y por la industria farmacéutica para los ensayos clínicos. También es el más caro. El coste anual de la licencia es de más de 1.300 € para una institución académica y en torno a 3.000 € para la licencia general. Permite el diseño de ensayos clínicos secuenciales, análisis de futilidad o ensayos clínicos optimizados (http://www.cytel.com/software/east).

## 7.12. RESUMEN DE LAS FÓRMULAS DEL TAMAÑO MUESTRAL

| | Proporciones | Medias |
|---|---|---|
| **Estimación (un grupo)** | $n = \frac{z^2_{\alpha/2} pq}{M^2}$ | $n = \frac{z^2_{\alpha/2} s^2}{M^2}$ |
| **Comparar (dos grupos)** | $n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2pq}{d^2}$ | $n = \frac{(z_{\alpha/2} + z_\beta)^2 \times 2s^2}{d^2}$ |
| **Potencia** | $z_\beta = \sqrt{\frac{nd^2}{2pq}} - z_{\alpha/2}$ | $z_\beta = \sqrt{\frac{nd^2}{2s^2}} - z_{\alpha/2}$ |

## 7.13. RESUMEN DE LAS INSTRUCCIONES EN STATA

| Objetivo | Instrucción a STATA |
|---|---|
| Comparar dos proporciones ($p_A$ = 0,10 y $p_B$ = 0,04) | **sampsi .1 .04, nocontinuity** |
| Comparar dos proporciones (0,10 y 0,04) con el doble de sujetos en un grupo y potencia = 80% | **sampsi .1 .04, nocont /// power(0.8) ratio(2)** |
| Comparar dos medias | **sampsi -5 -1, sd1(12) sd2(12)** |
| Comparar dos medias con igual desviación estándar | **sampsi 0 4, sd(12)** |
| Obtener potencia (comparar dos medias) | **sampsi 0 1, sd(3) n1(90) n2(99)** |
| Obtener potencia (comparar dos proporciones) | **sampsi .05 .03, nocont n(2450)** |

## REFERENCIAS


1. Clayton D, Hill M. Statistical models in Epidemiology. Oxford: Oxford University Press; 1993.
2. Lemeshow S, Hosmer DW, Klar J, Lwanga SK. Adequacy of sample size in health studies. Chichester: John Wiley & Sons Ltd; 1990.
3. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272(2):122-4.
4. Nguyen P. Public Opinion Polls, Chicken Soup and Sample Size. Teaching Statistics 2005;27:89-92.
5. Julious SA. Sample Sizes for Clinical Trials. London: Chapman and Hall; 2009.
6. Julious SA. Tutorial in Biostatistics: sample sizes for clinical trials with normal data. Stat Med 2004;23(12):1921-86.
7. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample sizes estimates. Stat Med 1992;41:185-96.
8. Martínez-Lapiscina EH, Clavero P, Toledo E, Estruch R, Salas-Salvadó J, San Julián B, et al. Mediterranean diet improves cognition: the PREDIMED-NAVARRA randomised trial. J Neurol Neurosurg Psychiatry 2013;84(12):1318-25.
9. Thompson IM Jr, Goodman PJ, Tangen CM, Parnes HL, Minasian LM, Godley PA, et al. Long-term survival of participants in the prostate cancer prevention trial. N Engl J Med 2013;369(7):603-10.
10. Day SJ, Graham DF. Sample size estimation for comparing two or more groups. Stat Med 1991;10(1):33-43.
11. Julious SA, Campbell MJ. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. Stat Med 2012;31(24):2904-36.
12. Martínez-González MA, Alonso A, Bes-Rastrollo M. Estimación del tamaño muestral. En: Martínez-González MA, Sánchez-Villegas A, Faulin Fajardo J, editores. Bioestadística amigable. 2.ª ed. Madrid: Díaz de Santos; 2006. p. 373-96.

# COMPARACIONES DE K MEDIAS (TRES O MÁS GRUPOS)

*M. Á. Martínez-González, N. Martín-Calvo, J. B. Toledo*

## 8.1. INTRODUCCIÓN AL ANOVA DE UNA VÍA

Se aplicará el análisis de la varianza (ANOVA) de una vía *(oneway)* para comparar las medias de tres o más grupos. Por ejemplo, se desea comparar el volumen cerebral total (en una escala donde el máximo es 100) en 3 grupos de consumo de alcohol (abstemios, exbebedores y bebedores activos). Este ejemplo se basa en un estudio real, que se puede consultar en su fuente original (**1**). En estas páginas, para facilitar los cálculos, el ejemplo se ha simplificado mucho. Se realizará primero con un tamaño muestral muy pequeño, y después, en un segundo ejemplo, se ampliará un poco más.

Imagínese en primer lugar que hay solo 15 sujetos en el estudio, cinco en cada uno de los tres grupos de consumo de alcohol. La tabla 8.1 presenta los volúmenes cerebrales totales de cada uno de los 15 sujetos. La parte inferior de la tabla recoge las medias, las desviaciones estándar y el tamaño muestral de cada uno de los tres grupos.

La media total ($media_{TOTAL}$ = 91) corresponde, *en este caso,* al promedio de las medias de los tres grupos, ya que todos ellos tienen el mismo tamaño ($n_1 = n_2 = n_3 = 5$). Si los grupos fuesen de tamaño desigual, se obtendría la media total mediante un promedio *ponderado* de las medias de los grupos y los pesos vendrían dados por $n_i$.

La pregunta es: ¿difiere el volumen cerebral según el consumo de alcohol?

Se deberán comparar tres medias. La hipótesis nula ($H_0$) es:

$$H_0 \equiv \mu_1 = \mu_2 = \mu_3$$

Para resolver este contraste de hipótesis, la variabilidad total se repartirá en dos componentes, uno explicable por las diferencias entre grupos (varianza entre grupos, *between*), que viene a expresar el *efecto,* y el otro, residual, el queda dentro de cada grupo (intragrupo, *within*) y expresa el *error.*

Recuérdese que, en estadística, se denomina suma de cuadrados (SC) a la suma de los cuadrados de la diferencia entre cada valor y la media:

$$SC = \sum (x_i - \bar{x})^2$$

Una varianza ($s^2$) es una SC dividida entre unos grados de libertad (gl).

$$s^2 = \frac{SC}{gl}$$

Por tanto:

$$SC = s^2 \times gl$$

Al hacer un ANOVA, se calcularán dos SC. La suma de cuadrados intragrupos o residual ($SC_{within}$) corresponde al *error* y será:

$$SC_{within(RESID)} = \sum s_i^2 (n_i - 1)$$

$$SC_{within(RESID)} = 4^2 (5-1) + 4^2 (5-1) + 4^2 (5-1) = 64 + 64 + 64 = 192$$

*Tabla 8.1 Ejemplo simple de ANOVA de una vía: volumen cerebral en bebedores*

| | ABSTEMIOS | EXBEBEDORES | BEBEDORES ACTIVOS |
|---|---|---|---|
| | 100 | 96 | 89 |
| | 98 | 94 | 87 |
| | 98 | 94 | 87 |
| | 94 | 90 | 83 |
| | 90 | 86 | 79 |
| $\bar{x}_i$ | **96** | **92** | **85** |
| $s_i$ | 4 | 4 | 4 |
| $n_i$ | 5 | 5 | 5 |

La suma de cuadrados entre grupos *(between)* o *efecto* ($SC_{between}$) será:

$$SC_{between} = \sum n_i \left(\bar{x}_i - \bar{x}_{TOTAL}\right)^2$$

$$SC_{between} = 5(96-91)^2 + 5(92-91)^2 + 5(85-91)^2 = 125 + 5 + 180 = 310$$

Después se calculan los gl entre grupos y dentro de grupos. Si N es el total de los tres grupos (N = 15) y $k$ el número de grupos ($k = 3$), los gl son:

- $gl_{between} = k - 1 = 3 - 1 = 2$.
- $gl_{within} = N - k = 15 - 3 = 12$.

Con estos resultados se prepara la tabla 8.2.

Las dos primeras columnas ya se han explicado. La tercera, que debería denominarse *Varianzas* (pero los programas informáticos la llaman *Mean Squares* o MS), surge de dividir cada SC entre sus respectivos gl. Las varianzas en el ejemplo resultan ser 155 *(between)* y 16 *(within)*. Parece lógico que la varianza dentro de los grupos sea 16, ya que las desviaciones estándar de los tres grupos valen 4 cada una. El cociente entre ambas se llama F. La última columna de la tabla es este cociente F, que procede de dividir la varianza correspondiente al efecto *(between)* entre la relativa al error *(within)*.

$$F_{k-1,N-k} = \frac{s^2_{between}}{s^2_{within}}$$

Como el numerador de F se debe al *efecto* (diferencias entre grupos, *between*) y el denominador es atribuible al *error (within)*, es fácil deducir que, cuanto mayor sea F, más pruebas habrá para rechazar la hipótesis nula. Ya se había explicado la distribución F como un cociente entre dos varianzas (v. apartado 6.2). Entonces F se empleaba para tomar la decisión entre usar la $t$ de Student para varianzas homogéneas o el test de Welch.

En el presente ejemplo, el valor de F será:

$$F_{3-1;15-3} = \frac{155}{16} = 9{,}7$$

*Tabla 8.2 Tabla de ANOVA de una vía correspondiente a los datos de la tabla 8.1*

| FUENTE | SUMAS DE CUADRADOS | GL | VARIANZAS | F |
|---|---|---|---|---|
| Entre grupos *(between)* | 310 | 2 | 310/2 = 155 | 155/16 = 9,7 |
| Residual *(within)* | 192 | 12 | 192/12 = 16 | |
| Total | 502 | 14 | | |

F es muy superior a la unidad, lo que sugiere que se acabará por rechazar la hipótesis nula. Sin embargo, antes hay que consultar en las tablas (o, en un programa) su significación estadística. F es un cociente entre dos varianzas y tiene dos tipos de grados de libertad, los del numerador y los del denominador. Son los de las respectivas varianzas que están arriba y abajo. Una vez identificados sus grados de libertad, el valor *p* de significación estadística para F se mira en Excel o STATA.

En Excel:

=DISTRF(9,7;2;12)

En STATA:

```
display Ftail(2,12,9.7)
```

En R:

```
pf(9.7,2,12,lower=FALSE)
```

El ordenador devolverá $p = 0{,}003$, lo cual lleva al rechazo de $H_0$ si se acepta la convención de un riesgo alfa *(α)* del 5% (v. apartado 4.14). Con F sucede que los efectos (diferencias entre media del grupo y media total) están elevados al cuadrado y, por tanto, siempre estarán incluidas las dos colas en el valor de *p*, como sucedía con la ji cuadrado ($\chi^2$). No se puede calcular un valor *p* para F a una cola; siempre incluirá las dos colas.

Con $p = 0{,}003$, se concluirá que el volumen cerebral total es significativamente *distinto* según el consumo de alcohol. Las medias de los grupos muestran que el volumen cerebral es menor en consumidores de alcohol que en no consumidores (**1**). Una vez excluidos posibles sesgos, habría que valorar el diseño epidemiológico utilizado. Esto permitirá decidir si la diferencia se debe a que el alcohol reduce el volumen cerebral o a que quienes tienen menor volumen cerebral son más propensos a consumir alcohol (o a ambas cosas a la vez).

## 8.2. RELACIÓN ENTRE EL ANOVA Y LA *T* DE STUDENT

Los cálculos para el ANOVA se han realizado usando solo medias, desviaciones estándar y tamaños muestrales. Se podría haber hecho conociendo esos índices estadísticos de cada grupo, sin necesidad de conocer uno a uno los datos de los individuos que forman la muestra. Es posible, porque el ANOVA, como la *t* de Student, es un método *paramétrico*. El ANOVA es muy similar a la *t* de Student (más de lo que parece a primera vista). El ANOVA compara medias, como la *t*, pero lo puede hacer para más de dos grupos. La *t* solo sirve para un caso particular, cuando existen solo dos grupos. No es lícito aplicar la *t* de Student para comparaciones por parejas cuando hay más de una pareja. Siempre que haya más de dos grupos, se debe aplicar el ANOVA. La *t* de Student comparaba dos grupos y respondía a la pregunta sobre cuál de los dos tenía una media superior. El ANOVA solo contesta a la pregunta de si *todas* las medias de los diversos grupos comparados son iguales o no; bastaría con que una fuese distinta del resto para que el ANOVA resultase significativo. La hipótesis alternativa no es que un grupo en concreto sea superior a otro, sino, simplemente, que los grupos son distintos entre sí (*heterogeneidad* de medias).

Imagínese por un momento un universo en el que no existiesen exbebedores. Bórrese entonces la columna de exbebedores de la tabla 8.1 y vuelva a repetirse todo el proceso del cálculo anterior de la F con solo dos grupos: abstemios y bebedores activos.

Entonces N = 10, la media total sería 90,5, y los demás resultados serían los de la tabla 8.3, con una F = 18,91 y una *p* todavía más lejana al límite de la significación ($p = 0{,}0025$). Esta *p no sería válida* en el mundo real, sino solo en un universo ficticio sin exbebedores. No es válida en el mundo real porque en él existen tres grupos, y cuando hay más de dos grupos es imperativo usar el análisis ANOVA.

***Tabla 8.3 Tabla de ANOVA de una vía correspondiente a la primera y última columna de la tabla 8.1, prescindiendo de los exbebedores***

| FUENTE | SUMAS DE CUADRADOS | GL | VARIANZAS | F |
|---|---|---|---|---|
| Entre grupos *(between)* | 302,5 | 1 | 302,5 | 18,9 |
| Residual *(within)* | 128 | 8 | 16 | |
| Total | 430,5 | 9 | | |

Únicamente en el imaginario mundo sin exbebedores se podría haber aplicado una *t* de Student para varianzas homogéneas con vistas a la comparación del volumen cerebral entre dos grupos: abstemios y bebedores activos. Tendría esta forma:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{s_{común}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{96-85}{4\sqrt{\frac{1}{5}+\frac{1}{5}}} = 4{,}348$$

Sin perder cierto sentido del humor, podría decirse que el parecido de este resultado con la F antes calculada por ANOVA (F = 18,91, v. tabla 8.3) es *sobrecogedor.* Basta con elevar la *t* al cuadrado ($4{,}348^2 = 18{,}91$) para comprobarlo. Cuando F tiene un solo grado de libertad en el numerador, se cumple siempre que $t^2 = F$. En esta situación, los grados de libertad del denominador de la F sería los mismos que los de la *t*. Además, la significación estadística de *t* a dos colas coincidirá con la de F. Todo esto sucede porque la *t* de Student pertenece a la familia del ANOVA, de la que es simplemente un caso particular aplicable cuando solo hay dos grupos. Por tanto, para comparar dos grupos se podría usar tanto la *t* como el ANOVA.

## 8.3. ANOVA DE UNA VÍA CON STATA

Los datos anteriores se han introducido en STATA en dos columnas: una denominada ***vdep,*** que contiene los valores del volumen cerebral, y otra llamada ***gr,*** que incluye «1» para los abstemios, «2» para los exbebedores y «3» para los bebedores activos. Para el análisis ANOVA se procedería así con la orden **oneway** del modo indicado en la figura 8.1.

Se reconocerán la mayoría de estos resultados comparándolos con la tabla 8.2. STATA añade la varianza total (35,857). Se apreciará que corresponde a la suma de cuadrados total (310 + 192 = 502)

```
. oneway vdep gr, tab

                Summary of vdep
        gr |      Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |        96          4           5
         2 |        92          4           5
         3 |        85          4           5
-----------+------------------------------------
     Total |        91  5.9880834          15

                Analysis of Variance
    Source              SS         df      MS            F     Prob > F
-----------------------------------------------------------------------
Between groups          310         2       155         9.69     0.0031
 Within groups          192        12        16
-----------------------------------------------------------------------
     Total              502        14  35.8571429

Bartlett's test for equal variances:  chi2(2) =   0.0000  Prob>chi2 = 1.000
```

Valor *p* para igualdad de medias

Valor *p* para igualdad de varianzas

**Figura 8.1** ANOVA de una vía con STATA *(oneway).*

dividida entre los grados de libertad totales (15 – 1 = 14). También añade un test de homogeneidad de varianzas de los grupos (test de Bartlett). En este ejemplo, la *homogeneidad* de las varianzas de los tres grupos es *total* (la varianza es exactamente la misma en los tres), por lo que la $\chi^2$ del test de Bartlett vale 0 y su valor $p$ = 1,00. Cuando el test de Bartlett resulte significativo (Prob > chi2 < 0,05), se deberá a que las varianzas de los grupos son significativamente distintas entre sí. Esta situación puede dar problemas si el tamaño de los grupos es pequeño (menor que 30), sobre todo si, además, son de desigual tamaño. En tal caso suele usarse el test de Kruskal-Wallis (v. apartado 8.7).

La opción **tab** con la que acaban las instrucciones que se indicaron en STATA es imprescindible para obtener una tabla descriptiva con la media, $n$ y $s$ en cada grupo. Este tipo de tabla se ha de pedir siempre; de lo contrario, se puede perder el contacto con los datos.

## 8.4. REQUISITOS DEL ANOVA

1. Tipos de variables:
   a. La variable *dependiente* debe ser cuantitativa (numérica). Es la variable cuyas medias se quieren comparar (volumen cerebral en el ejemplo).
   b. La variable *independiente* es simplemente el factor o variable de agrupación y debe ser cualitativa (categórica). Es el factor que clasifica las observaciones en diferentes grupos. En el ejemplo serían los grupos de consumo de alcohol.
2. Normalidad:

   El ANOVA es un método paramétrico; por eso, se ha dicho algunas veces que su requisito es que los datos de la variable dependiente sigan una distribución normal. No es cierto. Lo realmente importante es que la distribución de los *residuales* se aproxime bien a una normal. Los *residuales* son la diferencia entre cada valor individual y la *media de su grupo*. Los *residuales* así calculados, si se elevan al cuadrado, sumarán exactamente la SC residual.

   En STATA, si *vdep* es la variable dependiente y *gr* la que define los grupos, y hubiese tres grupos, se deberán dar los siguientes pasos para comprobar la normalidad de residuales:

```
quietly summarize vdep if gr==1  #describe ocultamente vdep
(en el grupo 1).

g resid=vdep - r(mean) if  gr ==1  #r(mean) es la media del
grupo 1.
```

   La media fue obtenida (ocultamente) en el paso anterior.

```
qui su vdep if  gr==2

replace resid=vdep-r(mean) if  gr==2   #ahora r(mean) =
media del grupo 2.

qui su vdep if  gr==3

replace resid =vdep-r(mean) if  gr==3   #ahora r(mean) =
media del grupo 3.

ladder resid   #pide test de normalidad de residuales y
sus transformaciones.

pnorm resid   #gráfico de normalidad de residuales.
```

El prefijo *quietly* indica a STATA que ejecute una orden de manera oculta y no la muestre en los resultados. El nombre **r(mean)** solo puede usarse si antes se ha pedido una descriptiva (**summarize**). Por tanto, **r(mean)** devolverá la media de la *última* estadística descriptiva que se haya solicitado.

Cuando los residuales no superan el test de normalidad, puede haber problemas para aplicar un ANOVA. El ANOVA es un procedimiento *robusto,* lo que significa que no suelen distorsionarse sus resultados aunque se hagan transgresiones en sus condiciones de aplicación. Los problemas de falta de normalidad de los residuales solo son importantes cuando el tamaño de muestra es reducido (menor de 30 por grupo), y se agravan todavía más si los grupos son de desigual tamaño y tienen varianzas diferentes. En tales situaciones se debería aplicar el test no paramétrico de Kruskal-Wallis.

3. Homogeneidad de varianzas *(homoscedasticidad):* se mira en STATA con el test de Bartlett y en SPSS con el test de Levene. Lo ideal es que el test no sea significativo cuando los grupos son de pequeño tamaño. Si todos tienen un tamaño superior a 30, la hipótesis aquí exigida no debería preocupar en absoluto.

## 8.5. ANOVA DE UNA VÍA CON OTROS PROGRAMAS

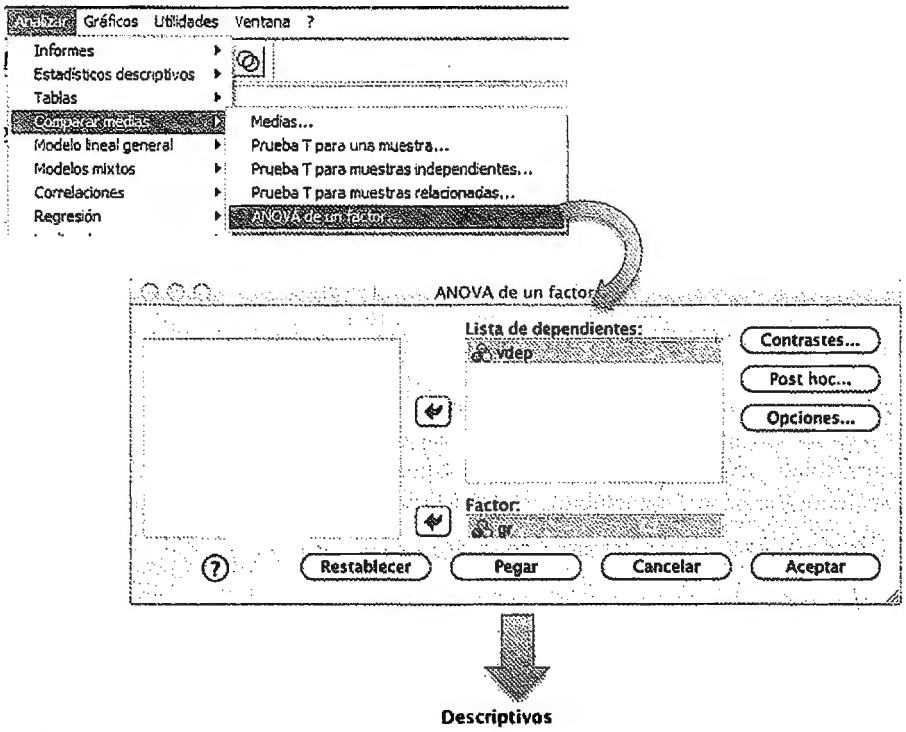
### 8.5.1. ANOVA de una vía con SPSS

En SPSS están programadas muchas opciones dentro del ANOVA de uso amigable y rápido. Es un buen programa para este procedimiento. Se deben seguir los pasos siguientes: «Analizar» → «Comparar medias» → «ANOVA de un factor...». Entonces aparece el cuadro de diálogo del ANOVA de un factor. En el recuadro de la izquierda se muestran las variables que tiene la base de datos; se seleccionarán la dependiente y el factor. Situándose sobre cada variable y pulsando el botón derecho del ratón, puede obtenerse información acerca de dicha variable. Si se pulsa sobre la variable cuantitativa que se utiliza como dependiente ***(vdep),*** se marcará como seleccionada; después se oprime el botón en forma de cabeza de flecha y esa variable ***(vdep)*** se traslada al recuadro central superior titulado «Lista de dependientes». A continuación se presiona sobre la variable independiente (Factor) y sobre la otra cabeza de flecha, con lo que el Factor (***gr*** en el ejemplo) se situará en la ventana inferior. Antes de oprimir el botón «Aceptar», siempre deben pedirse al menos dos características, que están en el menú «Opciones» → «Estadísticos» → «Descriptivos» y «Prueba de homogeneidad de las varianzas» (fig. 8.2).

En modo de sintaxis, para todo esto bastará con escribir:

```
ONEWAY vdep BY gr
/STAT DESCR HOMOG.
```

La salida incluirá una descripción de cada uno de los grupos, con la estimación de los intervalos de confianza para las medias, seguida de los mismos cálculos para el total de la muestra. A continuación, el programa ofrece la prueba de Levene para la igualdad u homogeneidad de varianzas entre los diversos grupos *(homoscedasticidad).* Cuando el test de Levene *no* alcance significación estadística ($p > 0{,}05$), puede asumirse que las varianzas son homogéneas, ya que la hipótesis nula del test de Levene es que las varianzas son iguales. En el ejemplo, donde $p = 1{,}00$, queda clara la *total* homogeneidad de varianzas. Al final aparece la tabla del ANOVA con los elementos ya vistos.




**Descriptivos**

vdep

| | N | Media | Desviación típica | Error típico | Intervalo de confianza para la media al 95% | | Mínimo | Máximo |
|---|---|---|---|---|---|---|---|---|
| | | | | | Límite inferior | Límite superior | | |
| 1 | 5 | 96,00 | 4,000 | 1,789 | 91,03 | 100,97 | 90 | 100 |
| 2 | 5 | 92,00 | 4,000 | 1,789 | 87,03 | 96,97 | 86 | 96 |
| 3 | 5 | 85,00 | 4,000 | 1,789 | 80,03 | 89,97 | 79 | 89 |
| Total | 15 | 91,00 | 5,988 | 1,546 | 87,68 | 94,32 | 79 | 100 |

**Prueba de homogeneidad de varianzas**

vdep

| Estadístico de Levene | gl1 | gl2 | Sig. |
|---|---|---|---|
| ,000 | 2 | 12 | 1,000 |

**ANOVA de un factor**

vdep

| | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|---|---|---|---|---|---|
| Inter-grupos | 310,000 | 2 | 155,000 | 9,687 | ,003 |
| Intra-grupos | 192,000 | 12 | 16,000 | | |
| Total | 502,000 | 14 | | | |

**Figura 8.2** ANOVA de una vía con SPSS *(comparar medias).*

### 8.5.2. ANOVA de una vía con R/Splus

Para resolver el ejercicio anterior se debe proceder de acuerdo con los pasos siguientes:

1. Se introducen los datos, lo cual se puede hacer como se explicó en el apartado 2.2.3 (primero se introducen en Excel, se guardan como .txt y se leen con la orden **read.table** desde R). Una vez leídos, se dará la orden **attach**.

   Otra posibilidad consiste en introducir directamente los datos como:

   ```
   >brain <-c(100,98,98,94,90,96,94,94,90,86,
   89,87,87,83,79)
   >oh<-factor(c(rep(0,5), rep(1,5), rep(2,5)))
   ```

   *Factor* convierte la variable ***gr*** en un factor (variable categórica) con tres niveles. Entonces se deben preparar los datos para que el programa entienda que forman una base de datos unida por columnas (**cbind** = *column bind*, es decir, unir las columnas) y una forma estructurada relacionada **(as.data.frame)**; se pondrá un nombre a la base de datos (***OH_brain***, en este ejemplo).

   ```
   >OH_brain<-as.data.frame(cbind(brain, oh))
   ```

   Si ahora se escribe:

   ```
   >OH_brain,
   ```

   el programa devolverá las dos columnas de las dos variables con todos los datos.
   Finalmente se indica a R/Splus que se va a usar ***OH_brain:***

   ```
   > attach(OH_brain)
   ```

2. Se pedirá el análisis de la varianza de una vía con la orden **aov** (de *analysis of variance*). Se dirigirán los resultados a un archivo temporal (***my_result*** en el ejemplo). Además, hay que señalar que el grupo no es una variable cuantitativa, sino categórica, es un factor **(as.factor)**.

   ```
   >my_result<-aov(brain ~ as.factor(oh))
   ```

   El símbolo ~ se puede obtener en Word desde «Insertar» (símbolo) y luego copiarlo y pegarlo en R/Splus. Otra posibilidad más sencilla es pulsar simultáneamente dos teclas (AltGr + 4) y luego la barra espaciadora.

3. Se obtendrá el resultado pidiendo **summary(my_result)**:

   ```
   > summary(myresult)
                Df Sum Sq Mean Sq F value  Pr(>F)
   as.factor(gr)  2    310     155  9.6875 0.00313 **
   Residuals     12    192      16
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ```

### 8.5.3. Programación de un ANOVA sencillo a partir de datos agregados en Excel

La figura 8.3 muestra una hoja de Excel en la que bastará con cambiar los datos en la zona sombreada para que se recalcule todo el ANOVA. Se puede descargar este archivo desde: http://www.unav.es/departamento/preventiva/recursos_bioestadistica. Se recomienda descargar esta hoja de cálculo y examinar detalladamente las ecuaciones que se han introducido en cada una de las casillas que no están sombreadas.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | | abstemios | exbebedores | activos | TOTAL |
| 2 | | | 100 | 96 | 89 | |
| 3 | | | 98 | 94 | 87 | |
| 4 | | | 98 | 94 | 87 | |
| 5 | | | 94 | 90 | 83 | |
| 6 | | | 90 | 86 | 79 | |
| 7 | | Promedio | 96 | 92 | 85 | 91,000 |
| 8 | | desvest | 4 | 4 | 4 | 5,988 |
| 9 | | n(contar) | 5 | 5 | 5 | 15 |
| 10 | | SC residual= | 64 | 64 | 64 | 192 |
| 11 | | SC between= | 125 | 5 | 180 | 310 |
| 12 | | SC TOTAL | | | | 502 |
| 13 | | | | | Var tot*N-1= | 502 |
| 14 | | | | | | |
| 15 | Fuente | SC | gl | Var | F | p |
| 16 | Entre (efecto) | 310 | 2 | 155 | 9,6875 | 0,00313 |
| 17 | Dentro (error) | 192 | 12 | 16 | | |
| 18 | TOTAL | 502 | 14 | | | |

**Figura 8.3** ANOVA programado en Excel (descargable desde http://www.unav.es/departamento/preventiva/recursos_bioestadistica).

## 8.6. EL ANOVA EN MANOS DE UN LECTOR SAGAZ DE ARTÍCULOS CIENTÍFICOS

### 8.6.1. Primer ejemplo

Si llega a manos de un lector sagaz (y bien formado sobre ANOVA) un artículo que compara la presión arterial sistólica en cinco grupos, cada uno de 40 sujetos, con medias (DE): 116 (11,0), 118 (11,0), 120 (11,6), 121 (13,1) y 125 (13,1) mmHg, y los autores mantienen que las diferencias entre estos cinco grupos son significativas con $p < 0{,}001$, ese lector podrá comprobar si este valor $p$ es verdad o no con los siguientes pasos:

$$\bar{x}_{TOTAL} = \frac{116+118+\ldots+125}{5} = 120$$

(Si los grupos fuesen de diferente tamaño, habría que hacer media ponderada; aquí no es necesario, porque los cinco grupos tienen cada uno 40 sujetos.)

$$SC_{between} = 40\left[(116-120)^2+\ldots+(125-120)^2\right] = 1.840$$

$$SC_{within} = (40-1)\left[(11)^2+\ldots+(13{,}1)^2\right] = 28.071{,}42$$

$$F_{k-1;\,N-k=195} = \frac{s^2_{between}}{s^2_{within}} = \frac{1.840/4}{28.071{,}42/(200-5)} = \frac{460}{144} = 3{,}2$$

A esta F le corresponde una $p = 0{,}014$.

```
=DISTRF(3,2;4;195)
```

En Excel devolverá $p = 0{,}014$.

El lector sabrá con seguridad que los autores del artículo han debido equivocarse cuando dicen que $p < 0{,}001$. Se puede descargar una hoja de Excel con estos datos y las ecuaciones ya preparadas desde: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

Se recomienda inventar supuestos sobre esta hoja de cálculo, variando las medias, *n* y *s*, para ver cómo se modifican los resultados del ANOVA en función de cada cambio. Este ejercicio servirá de ayuda para entender cómo funciona un ANOVA.

### 8.6.2. Segundo ejemplo

Otra posibilidad consiste en que los autores proporcionen las medias de los grupos (pero no sus desviaciones estándar particulares), además de indicar la media total y la desviación estándar total.

Por ejemplo, se podría leer en un artículo que el peso medio de tres grupos de 68, 99 y 70 participantes fue de 80,1 kg, 82,2 kg y 76,8 kg, respectivamente. La media total fue 80 kg (DE: 12,68). ¿Hay diferencias significativas entre los grupos?

Con la desviación estándar total se puede obtener la SC *total:*

$$SC_{TOTAL} = s^2_{TOTAL}(N-1) = 12{,}68^2(237-1) = 37.945$$

Con las medias de cada grupo y la media total se puede calcular la SC *between:*

$$SC_{between} = 68(80{,}1-80)^2 + 99(82{,}2-80)^2 + 70(76{,}8-80)^2 = 1.197$$

Como la SC total equivale a la suma $SC_{between} + SC_{within}$, puede despejarse la segunda:

$$SC_{within} = SC_{TOTAL} - SC_{between} = 37.945 - 1.197 = 36.748$$

$$F_{k-1;\,N-k=234} = \frac{s^2_{between}}{s^2_{within}} = \frac{1.197/2}{36.748/(237-3)} = \frac{598{,}5}{157} = 3{,}8$$

A esta F le corresponde una $p = 0{,}024$, si se mira en:

```
=DISTRF(3,8;2;234)
```

Así, el lector sabrá que las medias de los pesos de los tres grupos son significativamente distintas. También se puede descargar una hoja de Excel (ANOVA para un lector sagaz 2) con estos datos y las ecuaciones ya preparadas desde: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

### 8.6.3. Tercer ejemplo

Más difícil todavía. En algunos artículos, la única información que se presenta corresponde a los intervalos de confianza al 95% para las medias de los grupos. Por ejemplo, se compara el colesterol HDL entre tres grupos, de 19 (grupo A), 16 (grupo B) y 15 (grupo C) personas. Los intervalos de confianza al 95% para las medias de los grupos son:

Grupo A: 50,0 a 64,7.

Grupo B: 46,2 a 54,1.

Grupo C: 48,4 a 51,9.

¿Es el colesterol HDL diferente entre los tres grupos?

Habrá que dar los siguientes pasos:

1. Obtener las medias de los grupos. La media está en el centro de los dos límites del intervalo de confianza, ya que los intervalos son simétricos:
   a. Media A = (50 + 64,7)/2 = 57,35.
   b. Media B = (46,2 + 54,1)/2 = 50,15.
   c. Media C = (48,4 + 51,9)/2 = 50,15.
2. Obtener la media total por ponderación:

$$\bar{x}_{\text{TOTAL}} = \frac{\sum n_i \bar{x}_i}{n_i} = \frac{(19\times 57{,}35)+(16\times 50{,}15)+(15\times 50{,}15)}{19+16+15} = 52{,}9$$

3. Calcular la suma de cuadrados entre grupos:

$$SC_{\text{between}} = 19(57{,}35-52{,}9)^2 + 16(50{,}15-52{,}9)^2 + 15(50{,}15-52{,}9)^2 = 611$$

4. Obtener los errores estándar (EE) de la media de cada grupo. Es lo más complicado. Si el intervalo de confianza = media ± $t$ EE:

$$EE = (LSC - LIC)/(2\times t)$$

Los valores de $t$ de Student para 18, 15 y 14 gl son 2,101; 2,131 y 2,145, respectivamente. Por tanto:

$$EE_{\text{grupoA}} = \frac{64{,}7-50}{2\times 2{,}101} = 3{,}5 \quad EE_{\text{grupoB}} = \frac{54{,}1-46{,}2}{2\times 2{,}131} = 1{,}85$$

$$EE_{\text{grupoC}} = \frac{51{,}9-48{,}4}{2\times 2{,}145} = 0{,}82$$

5. Obtener las desviaciones estándar de cada grupo a partir de los errores estándar.
   Si EE = $s/n$^0,5, por tanto, $s$ = EE × $n$^0,5.
   Así:
   a. $s_{\text{grupo A}} = 3{,}5 \times 19^{0{,}5} = 15{,}26$.
   b. $s_{\text{grupo B}} = 1{,}85 \times 16^{0{,}5} = 7{,}4$.
   c. $s_{\text{grupo C}} = 0{,}82 \times 15^{0{,}5} = 3{,}18$.
6. Con esta última información se puede obtener ya la SC que faltaba:

$$SC_{\text{within}} = \left[(19-1)(15{,}26)^2 + \ldots + (15-1)(3{,}18)^2\right] = 5.155$$

7. Calcular F:

$$F_{k-1;\,N-k=47} = \frac{s^2_{\text{between}}}{s^2_{\text{within}}} = \frac{611/2}{5.155/(50-3)} = \frac{305{,}5}{109{,}7} = 2{,}78$$

A esta F le corresponde una $p$ = 0,072, si se mira en:

=DISTR.F(2,78;2;47)

o en STATA:

```
di Ftail(2,47,2.78)
```

o en R:

```
pf(2.78,2,47,lower=FALSE)
```

De este modo se sabrá que el test no ha resultado estadísticamente significativo según el umbral convencional de riesgo $\alpha$. Se puede descargar un Excel con este ejemplo desde: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

Durante este proceso es posible que, más allá del automatismo de completar estos cálculos, al lector se le haya pasado una interesante idea por la cabeza. ¿No ha existido algún resultado intermedio que haya llamado la atención? Un lector sagaz seguro que habrá advertido que hay algo que *falla*. Antes se ha dicho que un requisito de aplicación del ANOVA es la *homoscedasticidad*. ¿Se podría decir afirmar que las varianzas son iguales? No, de ningún modo. Las varianzas son:

Varianza de A = $15{,}25^2 = 232{,}6$.

Varianza de B = $7{,}41^2 = 54{,}9$.

Varianza de C = $3{,}16^2 = 9{,}99$.

La primera varianza es 23 veces mayor (!) que la última. En estas situaciones de transgresión tan desmesurada de un supuesto, no hace falta aplicar ningún test estadístico para darse cuenta de que no se cumplen las condiciones de aplicación. ¿Es grave esto? Se dijo que el ANOVA era *robusto*, es decir, soportaba bien ciertas transgresiones. Sin embargo, a pesar de ser robusto, en este ejemplo no soportará las violaciones de los supuestos, primero porque la separación de la homoscedasticidad es de gran magnitud, segundo porque todos los grupos son pequeños (menores que 30) y tercero porque los grupos son de distinto tamaño. En estos casos, el ANOVA no sirve y hay que aplicar el test de Kruskal-Wallis.

## 8.7. TEST NO PARAMÉTRICO ALTERNATIVO AL ANOVA: KRUSKAL-WALLIS

La alternativa no paramétrica al análisis de la varianza que más se ha usado en la investigación biomédica es el test de Kruskal-Wallis. Compara *de una sola vez* tres o más muestras independientes. Más que las medias, podría decirse que compara las *medianas* de los grupos, pues usa solo la posición de cada dato, y no su valor. Debe reemplazar al ANOVA si los datos son ordinales, o bien cuando el tamaño de muestra es pequeño, si no hay normalidad de los residuales y las varianzas de los grupos son claramente heterogéneas.

Permite prescindir de los supuestos del ANOVA, pero tiene tres inconvenientes:

- Requiere usar *todos* los datos y no es calculable a partir de índices ($\bar{x}$, s, n).
- Tiene menor potencia que el ANOVA (si se cumplen los requisitos del ANOVA).
- No permite construir de manera directa intervalos de confianza.

Se asemeja al test de la *U* de Mann-Whitney. Si se usase Kruskal-Wallis para comparar dos grupos, el valor *p* (a 2 colas) sería idéntico al de la *U* de Mann-Whitney. Se explicará con un ejemplo (tabla 8.4). Supóngase que se valoraron los cambios a 6 meses en la presión arterial sistólica (mmHg) con tres regímenes terapéuticos (A, B y C). Si se extrajesen los residuales, se vería que no siguen la normal.

*Tabla 8.4 Ejemplo para el test de Kruskal-Wallis: cambios en la presión arterial sistólica (mmHg) a 6 meses de seguimiento con tres tratamientos*

| TRATAMIENTO A | TRATAMIENTO B | TRATAMIENTO C |
|---|---|---|
| +3,5 | –4 | 0 |
| +3 | –4,5 | –0,5 |
| +2,5 | –5 | –1 |
| 0 | –5,5 | –31 |
| –2 | –7 | |
| **Media = +1,4** | **Media = –5,2** | **Media = –8,125** |

Un breve vídeo titulado *Normalidad resid o ANOVA o KWALLIS* explica cómo comprobar la normalidad de residuales con STATA y cómo realizar el test de Kruskal-Wallis. Otro vídeo titulado *Comparar 3+ grupos (k medias): ANOVA y Kruskal-Wallis* explica otros aspectos de estos procedimientos con STATA. Ambos se pueden consultar en: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

Para calcular a mano el test de Kruskal-Wallis, primero se ordenarán de menor a mayor todos los valores observados en las *k* muestras. A continuación se les asignan rangos según su *posición*, teniendo siempre en cuenta los empates (en el empate se asigna la media de los puestos empatados). Todo esto se recoge en la tabla 8.5, que proporciona la información de que las medias de los rangos en cada grupo ($\bar{R}_i$) son:

A: 11,3.

B: 4.

C: 7,125.

El rango medio esperado (o total, $\bar{R}_{total}$) será la suma total de rangos dividida entre el total de sujetos:

$$\bar{R}_{total} = \frac{\sum R_i}{N} = \frac{56,5+20+28,5}{14} = \frac{105}{14} = 7,5$$

*Tabla 8.5 Cálculo del test de Kruskal-Wallis*

| | | | RANGOS | | |
|---|---|---|---|---|---|
| VALOR (TAS) | GRUPO (TTO.) | RANGO | A | B | C |
| **–31** | C | 1 | | | 1 |
| **–7** | B | 2 | | 2 | |
| **–5,5** | B | 3 | | 3 | |
| **–5** | B | 4 | | 4 | |
| **–4,5** | B | 5 | | 5 | |
| **–4** | B | 6 | | 6 | |
| **–2** | A | 7 | 7 | | |
| **–1** | C | 8 | | | 8 |
| **–0,5** | C | 9 | | | 9 |
| **0** | A | 10,5 | 10,5 | | |
| **0** | C | 10,5 | | | 10,5 |
| **2,5** | A | 12 | 12 | | |
| **3** | A | 13 | 13 | | |
| **3,5** | A | 14 | 14 | | |
| Sumas | | | 56,5 | 20 | 28,5 |
| $n_i$ | | | 5 | 5 | 4 |
| **Media de rangos** | | | **11,3** | **4** | **7,125** |
| Esperados: (N + 1)/2 | | | 7,5 | 7,5 | 7,5 |
| Numerador: sumandos ($ji^2$) | | | $(11,3-7,5)^2$ | $(4-7,5)^2$ | $(7,125-7,5)^2$ |
| Denominador: N(N + 1)/12 | | | 14(15)/12 | | |

donde $R_i$ es el rango de cada *sujeto* y N el tamaño muestral total. Como la suma total de rangos debe ser N(N + 1)/2, también se cumple que:

$$\bar{R}_{total} = \frac{N(N+1)/2}{N} = \frac{N-1}{2} = \frac{14+1}{2} = 7{,}5$$

Finalmente, se calcula una $\chi^2$, con $k-1$ grados de libertad, siendo $k$ el número de grupos:

$$\chi^2 = \frac{\sum n_i \left(\bar{R}_i - \bar{R}_{TOTAL}\right)^2}{N(N+1)/12}$$

donde $n_i$ es el tamaño de cada grupo y la primera media que hay dentro del paréntesis ($\bar{R}_i$) es el rango medio de cada grupo. El número 12 del denominador es una constante, como ocurría en el test de la *U* de Mann-Whitney.

$$\chi^2_{2\,gl} = \frac{5(11{,}3-7{,}5)^2 + 5(4-7{,}5)^2 + 4(7{,}125-7{,}5)^2}{14(15)/12} = 7{,}7$$

Si se consulta en STATA un valor de $\chi^2 = 7{,}7$ con dos grados de libertad:

```
. di chi2tail(2,7.7)
```

devolverá $p = 0{,}021$.

En consecuencia, si se asume un riesgo $\alpha$ del 5%, se puede afirmar que el cambio de la presión arterial fue significativamente distinto en los tres grupos.

### 8.7.1. Kruskal-Wallis en STATA

Se procederá del modo siguiente:

```
. kwallis cambio, by(gr)

Kruskal-Wallis equality-of-populations rank test
```

| gr | Obs | Rank Sum |
|---|---|---|
| 1 | 5 | 56.50 |
| 2 | 5 | 20.00 |
| 3 | 4 | 28.50 |

```
chi-squared =      7.658 with 2 d.f.
probability =      0.0217

chi-squared with ties =      7.675 with 2 d.f.
probability =      0.0216
```

Si solo se posee esta salida, los rangos *medios* de cada grupo podrían obtenerse dividiendo 56,5 entre 5, etc.

Cuando existen empates en los rangos, hay una pequeña diferencia entre los cálculos a mano y los realizados con ordenador, ya que el programa tiene incorporada una rutina que efectúa una mínima corrección para tener en cuenta los empates. STATA ofrece las dos posibilidades de cálculo: arriba, el valor de $\chi^2 = 7{,}658$ corresponde al cálculo sin corrección por empates (el que se

**Rangos**

| | gr | N | Rango promedio |
|---|---|---|---|
| cambio | 1 | 5 | 11,30 |
| | 2 | 5 | 4,00 |
| | 3 | 4 | 7,13 |
| | Total | 14 | |

**Estadísticos de contraste[a,b]**

| | cambio |
|---|---|
| Chi-cuadrado | 7,675 |
| gl | 2 |
| Sig. asintót. | ,022 |

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: gr

**Figura 8.4** Salida del test de Kruskal-Wallis en SPSS.

ha mostrado antes) y, abajo, el otro valor de $\chi^2 = 7{,}675$ está corregido por empates. Si se realizan los cálculos con ordenador, debe preferirse esta segunda opción.

### 8.7.2. Kruskal-Wallis en otros programas

En SPSS se usarán las siguientes instrucciones:

```
NPAR TESTS

/K-W=cambio BY gr(1 3)
```

La numeración dentro del paréntesis es el rango de códigos o etiquetas de los grupos que se van a comparar. El listado de salida es el que aparece en la figura 8.4.

Con R/Splus, si los datos no están disponibles y hay que introducirlos, se procederá del modo siguiente:

```
> A<-c(3.5,3,2.5,0,-2)
> B<-c(-4,-4.5,-5,-5.5,-7)
> C<-c(0,-.5,-1,-31)
> kruskal.test(list(A,B,C))

        Kruskal-Wallis rank sum test

data:  list(A, B, C)
Kruskal-Wallis chi-squared = 7.6747, df = 2,
p-value = 0.02155
```

Si los datos ya están introducidos y se ha hecho el paso de **attach,** se actuará así:

```
> kruskal.test(camb,gr)

        Kruskal-Wallis rank sum test

data:  camb and gr
Kruskal-Wallis chi-squared = 7.6747, df = 2,
p-value = 0.02155
```

## 8.8. COMPARACIONES MÚLTIPLES: CONTRASTES *A PRIORI*

Cuando el ANOVA resulte estadísticamente significativo, se sabrá que las medias de los grupos son diferentes entre sí, pero hay que profundizar más y preguntarse *dónde* están específicamente esas diferencias. Para ello se usan los *contrastes* del ANOVA, que buscan diferencias particulares entre grupos y pueden seguir dos aproximaciones:

1. *A priori.* Se realizan pocas comparaciones y se planifican antes de iniciar el análisis de los datos. Se seleccionan solo las comparaciones más interesantes, juiciosas y razonables para la investigación. Es la mejor opción, pues hace explícitas las hipótesis de interés y evita el mal recurso automático de comparar *todo con todo* (llamado a veces *excursión de pesca,* en la jerga estadística). Las comparaciones *a priori* pueden ser *ortogonales* o no ortogonales. Las ortogonales extraen el máximo partido a los datos.
2. *A posteriori* (o *post hoc*). No están planificadas y en ellas se compara *todo con todo*, es decir, se establecen todas las posibles comparaciones de medias (dos a dos). Se podrían hacer $k(k-1)/2$ contrastes dos a dos. Es decir, si hay, por ejemplo, cinco grupos ($k = 5$), se podrán realizar 10 comparaciones por parejas.

El siguiente ejemplo (tabla 8.6) amplía a cuatro grupos la anterior comparación del volumen cerebral según niveles de consumo de alcohol. El ANOVA podría realizarse a partir de los datos de la tabla 8.6.

Una vez que se sabe que resultan significativas las diferencias entre los 4 grupos ($F_{3,116} = 57{,}7$), se pueden usar los *contrastes a priori* para hacer tres comparaciones, como muestra esa figura:

1. Contraste 1: abstemios frente a bebedores.
2. Contraste 2: exbebedores frente a bebedores.
3. Contraste 3: bebedores moderados *(light)* frente a bebedores intensos *(heavy).*

Cada contraste *a priori* supone agrupar las categorías en solo *dos* sectores y asignar unos pesos o coeficientes ($w_i$) con signo positivo o negativo para cada grupo según a cuál de esos dos sectores se asigne cada grupo. Por ejemplo, en el primer contraste, los abstemios se incluirán en un sector (signo negativo) y el conjunto de bebedores *light* + *heavy* en otro (signo positivo), mientras que los exbebedores no intervienen en este primer contraste. El signo sirve para distinguir a qué *equipo* se pertenece.

Además de signo, los coeficientes tienen magnitud. El grupo o grupos que no entran en la comparación llevarán un coeficiente igual a 0 (exbebedores en el primer contraste). Cuando un grupo es el único en su sector, la magnitud de su coeficiente será el *doble* del coeficiente de los grupos presentes en el otro sector que contenga *dos* grupos. Si hubiese un grupo en un sector y *tres* grupos en el otro, los coeficientes para el grupo que está solo deberían valer el *triple.* Considerando lo anterior, el contraste 1 (abstemios frente a bebedores) tendrá estos coeficientes:

$$-2\,(\text{abstemios}) + 0\,(\text{exbebedores}) + 1\,(\textit{light}) + 1\,(\textit{heavy})$$

Los coeficientes multiplicados por las medias de cada grupo serán el numerador de una *t* de Student. Su denominador es un error estándar:

$$t_{gl=\text{residuales}} = \frac{\sum w_i \bar{x}_i}{\sqrt{s^2_{\text{residual}} \left( \sum \frac{w_i^2}{n} \right)}}$$

*Tabla 8.6 Contrastes* a priori

| | VOLUMEN CEREBRAL | | |
|---|---|---|---|
| ALCOHOL | MEDIAS | N | S |
| Abstemio | 96 | 30 | 3,68 |
| Exbebedor | 92 | 30 | 3,57 |
| *Light* | 86 | 30 | 4,11 |
| *Heavy* | 84 | 30 | 4,21 |
| Total | | 120 | |

| ANOVA | SUMA DE CUADRADOS | GL | VARIANZAS |
|---|---|---|---|
| Entre grupos | 2.730,0 | 3 | 910,00 |
| Residual | 1.766,2 | 116 | 15,23 |
| Totales | 4.496,2 | 119 | |
| | | F = | 59,75 |
| | | *p* = | <0,001 |

| Contraste 1 | | | | |
|---|---|---|---|---|
| | COEF. ($W_i$) | MEDIA | $W_i \times$ MEDIA | $W_i^2/N_i$ |
| Abstemio | −2 | 96 | −192 | 4/30 |
| Exbebedor | 0 | 92 | 0 | 0 |
| *Light* | 1 | 86 | 86 | 1/30 |
| *Heavy* | 1 | 84 | 84 | 1/30 |
| Numerador/denominador | | | −22 | (15,23 × (6/30))^0,5 |
| | | | $t_{116}$ = | −12,6 |

| Contraste 2 | | | | |
|---|---|---|---|---|
| Abstemio | 0 | 96 | 0 | |
| Exbebedor | −2 | 92 | −184 | 4/30 |
| *Light* | 1 | 86 | 86 | 1/30 |
| *Heavy* | 1 | 84 | 84 | 1/30 |
| Numerador/denominador | | | −14 | (15,23 × (6/30))^0,5 |
| | | | $t_{116}$ = | −8,02 |

| Contraste 3 | | | | |
|---|---|---|---|---|
| Abstemio | 0 | 96 | 0 | 0 |
| Exbebedor | 0 | 92 | 0 | 0 |
| *Light* | −1 | 86 | −86 | 1/30 |
| *Heavy* | 1 | 84 | 84 | 1/30 |
| Numerador/denominador | | | −2 | (15,23 × (2/30))^0,5 |
| | | | $t_{116}$ = | −1,98 |

Los grados de libertad son los de la varianza residual, ya que es la varianza usada en el error estándar del contraste. Así, en el ejemplo, para el primer contraste, *t* valdrá:

$$t_{gl=116} = \frac{(-2\times96)+(+1\times86)+(+1\times84)}{\sqrt{57,23\left(\frac{-2^2}{30}+\frac{+1^2}{30}+\frac{+1^2}{30}\right)}} = -12,6$$

La tabla 8.6 muestra los resultados de *t* para los tres contrastes y los cálculos intermedios. El error estándar de los contrastes se parece mucho al error estándar de la diferencia de dos medias (EEDM, en la *t* de Student), pero en vez de la varianza ponderada que se usaba en la *t* de Student, en este caso se utiliza la varianza residual del ANOVA. Además, en vez de la suma del inverso del tamaño de cada grupo ($1/n_i$), se usa la suma de los coeficientes al cuadrado divididos por el tamaño de su respectivo grupo. Téngase en cuenta que si $w_i$ fuesen +1 y −1, la expresión sería muy similar a la de la *t* de Student para comparar dos medias.

En el ejemplo, con el nivel convencional de $\alpha = 0{,}05$, los dos primeros contrastes serán significativos. El tercero, en cambio, estará en el límite de la significación. Se interpretará que el volumen cerebral medio es significativamente inferior en bebedores que en abstemios; también es significativamente inferior en bebedores activos que en exbebedores, y tiende a ser menor (en el límite de la significación) en bebedores intensos que en moderados. Todo esto apoyaría un efecto dosis-respuesta.

En definitiva, con los tres contrastes mostrados en la tabla 8.6 se han efectuado tres comparaciones que han *diseccionado* en detalle la heterogeneidad global entre los cuatro grupos que había detectado al principio el ANOVA. Estos tres contrastes *no necesitan corregirse por comparaciones múltiples*, ya que son *ortogonales*. Los contrastes ortogonales suponen el máximo aprovechamiento de los datos, porque evitan las redundancias.

Para que los contrastes *a priori* sean *ortogonales*, se deben cumplir las siguientes tres condiciones, que aseguran que estos contrastes no sean redundantes:

1. Pueden obtenerse tantos contrastes como grupos haya menos 1.
2. La suma de los pesos de cada contraste debe ser 0.
3. La suma de todos los posibles productos de los pesos asignados a cada grupo en un contraste por los pesos asignados a ese mismo grupo en otro contraste debe valer 0.

Así puede comprobarse en la parte superior de la tabla 8.7. La primera condición se cumple, ya que hay cuatro grupos y solo se han hecho tres contrastes (contrastes = $k - 1$). La segunda condición se comprueba al ver que $-2 + 0 + 1 + 1$ sumados dan 0, que $0 - 2 + 1 + 1$ también suman 0 y que $0 + 0 - 1 + 1$ también son igual a 0. La tercera condición requiere ir multiplicando el peso de los contrastes de cada columna $(-2)(0)(0) = 0$; $(0)(-2)(0) = 0$; $(+1)(+1)(-1) = -1$; $(+1)(+1)(+1) = +1$ y luego sumarlo todo $(0 + 0 - 1 + 1 = 0)$. Se comprueba que los productos por columnas también suman 0.

*Tabla 8.7* ***Contrastes* a priori**

| ORTOGONALES | | PESOS ($w_i$) | | | | SUMA $w_i$ |
|---|---|---|---|---|---|---|
| CONTRASTES | HIPÓTESIS NULAS | $W_1$ | $W_2$ | $W_3$ | $W_4$ | $\sum w_i$ |
| **Esquema de contrastes usado en el ejemplo** | | | | | | |
| Contraste 1 | 1 = 3 + 4 | −2 | 0 | +1 | +1 | 0 |
| Contraste 2 | 2 = 3 + 4 | 0 | −2 | +1 | +1 | 0 |
| Contraste 3 | 3 = 4 | 0 | 0 | −1 | +1 | 0 |
| | PRODUCTOS (por columna) | 0 | 0 | −1 | +1 | 0 |
| **Otro posible esquema ortogonal** | | | | | | |
| Contraste 1 | 1 = 2 + 3 + 4 | −3 | +1 | +1 | +1 | 0 |
| Contraste 2 | 2 = 3 + 4 | 0 | −2 | +1 | +1 | 0 |
| Contraste 3 | 3 = 4 | 0 | 0 | −1 | +1 | 0 |
| | PRODUCTOS (por columna) | 0 | 0 | −1 | +1 | 0 |
| **Otro posible esquema ortogonal** | | | | | | |
| Contraste 1 | 1 + 2 = 3 + 4 | −1 | −1 | +1 | +1 | 0 |
| Contraste 2 | 1 + 2 + 3 = 4 | −1 | −1 | −1 | +3 | 0 |
| Contraste 3 | 1 = 2 | +1 | −1 | 0 | 0 | 0 |
| | PRODUCTOS (por columna) | +1 | −1 | 0 | 0 | 0 |
| **Esquema *no* ortogonal** | | | | | | |
| Contraste 1 | 1 = 4 | −1 | 0 | 0 | +1 | 0 |
| Contraste 2 | 1 + 2 + 3 = 4 | −1 | −1 | −1 | +3 | 0 |
| Contraste 3 | 3 = 4 | 0 | 0 | −1 | +1 | 0 |
| | PRODUCTOS (por columna) | 0 | 0 | 0 | +3 | +3 |

En la tabla 8.7 se presentan otros ejemplos con diversos esquemas de contrastes. El primero corresponde al del ejemplo. El segundo esquema establecería una primera comparación de abstemios frente al resto de grupos, una segunda comparación de exbebedores frente a bebedores, y una tercera entre bebedores moderados e intensos. El tercer esquema compararía los que ahora no beben (abstemios + exbebedores) con los que sí lo hacen, después los bebedores intensos frente al resto de grupos, y, por último, abstemios frente a exbebedores. Todos estos esquemas corresponden a contrastes *ortogonales*. En cambio, el último esquema no es ortogonal, porque la suma de los productos de los coeficientes por columnas resulta ser +3 y debería haber sido 0. En esta última situación, el contraste necesitaría *penalización,* ya que se han obtenido tres valores *p*, cada uno de ellos con un riesgo $\alpha$ del 5%; así, al haber más intentos, es más fácil cometer un error tipo 1. Por el contrario, cuando son ortogonales, se logra un reparto equilibrado de las sumas de cuadrados que permite subdividir y jerarquizar las hipótesis y ahorrarse las penalizaciones. En consecuencia, los contrastes *a priori* de tipo ortogonal son muy recomendables. Su inconveniente reside en que necesitan hacer agrupaciones de categorías, y, a veces, tales agrupaciones pueden no estar justificadas.

## 8.9. CONTRASTES *A POSTERIORI (POST HOC)*: CORRECCIONES POR COMPARACIONES MÚLTIPLES

Son todas las posibles comparaciones de medias entre parejas de grupos que se pueden efectuar tras un ANOVA significativo y aplicarse cuando no haya ninguna forma lógica de agrupar o combinar varias categorías conjuntamente. Estarían justificados solo si realmente se desease hacer todas las comparaciones por parejas e interesase comparar *todo con todo.*

Se debe tener en cuenta que, al haberse hecho muchas comparaciones, tal vez alguna resulte significativa simplemente por azar (**2-4**). La idea general es que se requiere un valor *p menor* del habitual para poder concluir que las diferencias fueron significativas, pues se realizaron numerosas comparaciones con los mismos datos. No obstante, este planteamiento de penalizar las comparaciones múltiples ha sido criticado (**5**). Actualmente existe consenso, para todo tipo de comparaciones múltiples, más allá del caso particular del ANOVA, acerca de que *se deben penalizar* las comparaciones múltiples, especialmente en el análisis de subgrupos en los ensayos clínicos (**6,7**) y en las múltiples comparaciones que se realizan cuando se estudian muchos posibles polimorfismos genéticos como potencialmente asociados a una enfermedad (**8,9**).

Cada contraste de hipótesis tiene una probabilidad individual de cometer un error tipo 1. El error tipo 1 consiste en equivocarse y rechazar la hipótesis nula ($H_0$) cuando realmente es verdadera. Tal probabilidad de errar es el riesgo $\alpha$. Por lo tanto, la probabilidad de *acertar* y *no* rechazar una $H_0$ que es verdadera será $1 - \alpha$. Si el riesgo $\alpha$ es del 5%, la probabilidad de acertar cuando $H_0$ es correcta es 0,95. Pero si se hacen tres comparaciones y son independientes entre sí, la probabilidad de acertar en las tres comparaciones será $0,95 \times 0,95 \times 0,95 = 0,86$. Si la probabilidad de acertar es del 86%, la de equivocarse al menos una vez será del 14%. Esta será la probabilidad global de cometer al menos un error tipo 1 cuando se hacen tres comparaciones: $1 - (1 - 0,05)^3 = 0,14$. En general, para *c* comparaciones se dirá que la probabilidad global de un error tipo 1 es el error alfa global *(family wise alpha error),* y se calcula como:

$$\alpha_{global} = 1 - (1 - \alpha)^c$$

Dicen que el número 13 trae mala suerte. Si se hiciesen 13 comparaciones y las 13 hipótesis nulas fueran ciertas, ninguna debería haber resultado significativa. Ahora bien, por los repetidos intentos, y según la fórmula anterior, la probabilidad de que al menos una sea significativa ya sería casi del 50%. En la mitad de las ocasiones en que se realicen 13 comparaciones cuya $H_0$ sea verdad, se rechazará al menos una hipótesis nula, que no se debería haber rechazado.

$$1 - (1 - 0,05)^{13} = 0,49$$

Cometer un error tipo 1 el 50% de las veces es equivocarse *mucho*. Al realizar 13 comparaciones la probabilidad de este error es la misma que al echarlo a cara o cruz. Por eso hay que penalizar las comparaciones múltiples. Existen varios métodos para corregir el valor *p* por el número de comparaciones múltiples realizadas, de modo que globalmente consiguen que el $\alpha$ *global* quede siempre en el 5%. El error alfa global *(family wise alpha error)* se entiende como la probabilidad de obtener uno o más valores *p* significativos si todas las hipótesis nulas de las múltiples comparaciones fuesen ciertas.

### 8.9.1. Método *post hoc* de Bonferroni

La corrección de Bonferroni consiste en calcular un valor $p'$ penalizado:

$$p' = 1 - (1 - p)^c$$

Así, para c = 10 comparaciones, un valor *p* no penalizado de 0,005 se transforma con el procedimiento de Bonferroni en el siguiente valor $p'$ penalizado

$$p' = 1 - (1 - 0{,}005)^{10} = 0{,}04889$$

La anterior expresión se corresponde casi completamente con el simple producto:

$$p' = p \times c$$

$$p' = 0{,}005 \times 10 = 0{,}05$$

que es una buena aproximación y resulta más fácil y cómodo.

De este modo, cuando se aplica el procedimiento de Bonferroni a un estudio en el que se han hecho 10 comparaciones, los valores *p* individuales que se obtengan se deberían multiplicar por 10 para obtener $p'$ y ese valor $p'$ (penalizado) es el que se presentará en la publicación ($p' = p \times 10$). Así, solo valores iniciales $p < 0{,}005$ (es decir, 10 veces inferiores a lo convencional) podrán considerarse significativos.

El procedimiento de Bonferroni es estricto y casi desesperado, porque penaliza mucho la *p*, especialmente en presencia de numerosas comparaciones. No debe perderse de vista que, cuando unos resultados siguen siendo significativos incluso a pesar de aplicarles un procedimiento tan conservador como el de Bonferroni, nadie dudará de la credibilidad estadística de esa conclusión.

### 8.9.2. Métodos *post hoc* de Scheffé, Dunnett, Tukey y Sidak

Son otros métodos, entre muchos, que sirven para corregir comparaciones múltiples. Habitualmente se obtendrán con el ordenador. No es práctico explicar su cálculo manual.

El método de Scheffé puede ser más conservador todavía que el de Bonferroni, especialmente cuando las comparaciones sean pocas. El método de Dunnett compara un conjunto de grupos, uno a uno, todos frente a una única media, la de un solo grupo de referencia (grupo *control*). Es el procedimiento indicado cuando esta sea la situación experimental, y ahorra el número de comparaciones. Por ejemplo, si hay cuatro grupos, tres de ellos con tratamiento activo y un cuarto con placebo, solo se harían tres comparaciones: cada grupo activo siempre frente a placebo, en vez de las seis comparaciones de todas las posibles parejas. El método de Tamhane corrige las comparaciones múltiples cuando hay *heteroscedasticidad* (varianzas desiguales).

El método de Tukey (*honest significant differences,* HSD) es de los que menos penaliza los valores *p*. Pueden suscitarse dudas sobre la validez de este método, sobre todo cuando los grupos no tengan todos el mismo tamaño. Su uso puede levantar sospechas y algún lector podría pensar que el investigador lo eligió tendenciosamente para obtener resultados significativos. El método de Sidak se considera intermedio.

Cuando los grupos sean ordenables (p. ej., dosis crecientes de un mismo fármaco), es más interesante comprobar si existe un gradiente dosis-respuesta. Se debe evaluar si hay una tendencia

progresiva de cambio de la media conforme aumenta la dosis. Es preferible hacerlo con regresión lineal, como se verá en el apartado 10.6. Además, así se evita tener que recurrir al poco elegante sistema de comparar *todo con todo.*

### 8.9.3. Contrastes *a priori* con STATA

En STATA, los contrastes *a priori* requieren usar la orden **anova** en vez de usar **oneway** y además se debe generar un vector con los coeficientes, por ejemplo (–2, 0, +1, +1). Se hace con la orden **matrix.** Se da un nombre a ese vector tras escribir **matrix** y se introducen entre paréntesis, separados por comas, cada uno de los coeficientes del contraste. Al final hay que escribir, también separado por coma, un cero como si fuese un coeficiente más. Por último, se usa la orden **test.**

```
anova brain OH, tab
matrix C1=(-2, 0, +1, +1, 0)
test, test(C1)
```

Con esta secuencia de órdenes, STATA devolverá el siguiente resultado:

```
(1) -2*1b.OH + 3.OH + 4.OH = 0

F(1, 116) = 152.74
Prob > F =  0.0000
```

En la parte superior *–2*1b.OH* indica que se dio un peso negativo y de doble magnitud ($w_i = -2$) para el primer grupo de consumo de alcohol.

El valor $F_{gl=1,116} = 152{,}7$ que aparece en este resultado debe ser exactamente el cuadrado de la $t_{gl=116}$ (contraste 1) que se ha presentado en el ejemplo (v. tabla 8.7). Las pequeñas diferencias se deben al mayor número de decimales con que trabaja STATA.

Se procederá del mismo modo para los otros dos contrastes.

### 8.9.4. Contrastes *post hoc* con STATA

Se pueden añadir como opciones, separadas por una coma, tras la orden **oneway.**

```
oneway brain OH, sidak
oneway brain OH, scheffe
oneway brain OH, bonferroni
```

El resultado se presenta siempre con el mismo formato:

Comparison of brain by RECODE of id
(Bonferroni)

| Row Mean-Col Mean | abstemio | ex-beb. | light |
|---|---|---|---|
| ex-beb. | -3.99997<br>0.001 | | |
| light | -10<br>0.000 | -6.00004<br>0.000 | |
| heavy | -12<br>0.000 | -8.00003<br>0.000 | -1.99999<br>0.324 |

**Prueba de homogeneidad de varianzas**

brain

| Estadístico de Levene | gl1 | gl2 | Sig. |
|---|---|---|---|
| 1,664 | 3 | 116 | ,179 |

**Coeficientes de los contrastes**

| Contraste | OH | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | -2 | 0 | 1 | 1 |
| 2 | 0 | -2 | 1 | 1 |
| 3 | 0 | 0 | -1 | 1 |

**Pruebas para los contrastes**

| | | Contraste | Valor del contraste | Error típico | t | gl | Sig. (bilateral) |
|---|---|---|---|---|---|---|---|
| brain | Asumiendo igualdad de varianzas | 1 | -22,00 | 1,780 | -12,359 | 116 | ,000 |
| | | 2 | -14,00 | 1,780 | -7,865 | 116 | ,000 |
| | | 3 | -2,00 | 1,028 | -1,946 | 116 | ,054 |
| | No asumiendo igualdad de varianzas | 1 | -22,00 | 1,545 | -14,239 | 72,603 | ,000 |
| | | 2 | -14,00 | 1,961 | -7,140 | 52,754 | ,000 |
| | | 3 | -2,00 | 1,053 | -1,899 | 55,504 | ,063 |

**Figura 8.5** Contrastes *a priori* en SPSS.

En el cruce de cada fila y cada columna aparece la diferencia de medias entre esos dos grupos, e inmediatamente debajo, el valor *p* corregido. En el ejemplo, con el procedimiento de Bonferroni, todas las comparaciones por parejas, una vez penalizadas, resultarían estadísticamente significativas, salvo la diferencia entre bebedores ligeros e intensos ($p = 0{,}32$). Téngase en cuenta que, cuando se planificaron bien los contrastes *a priori*, esta misma comparación alcanzó un valor $p = 0{,}05$, lo cual apoya el interés de planificar bien contrastes ortogonales, en vez de optar por *excursiones de pesca* y comparar todo con todo.

### 8.9.5. Contrastes en el ANOVA en otros paquetes de estadística

SPSS ofrece una salida doble para los contrastes *a priori*, con una opción para varianzas homogéneas y otra para varianzas heterogéneas (fig. 8.5).

Se debe seleccionar una u otra en función del resultado del test de Levene. Si este test fuese significativo, indicaría que las varianzas son desiguales y se elegirán los tests para varianzas heterogéneas que SPSS presenta en la parte inferior. Estos resultados se pueden obtener con la siguiente sintaxis:

```
ONEWAY brain BY oh
  /CON= -2 0 1 1
  /CON=0 -2 1 1
  /CON=0 0 -1 1
  /STAT HOMOG.
```

SPSS también presenta muchas opciones de contrastes *post hoc* y, además, las acompaña de intervalos de confianza para las diferencias. Las opciones son las que muestra la figura 8.6.

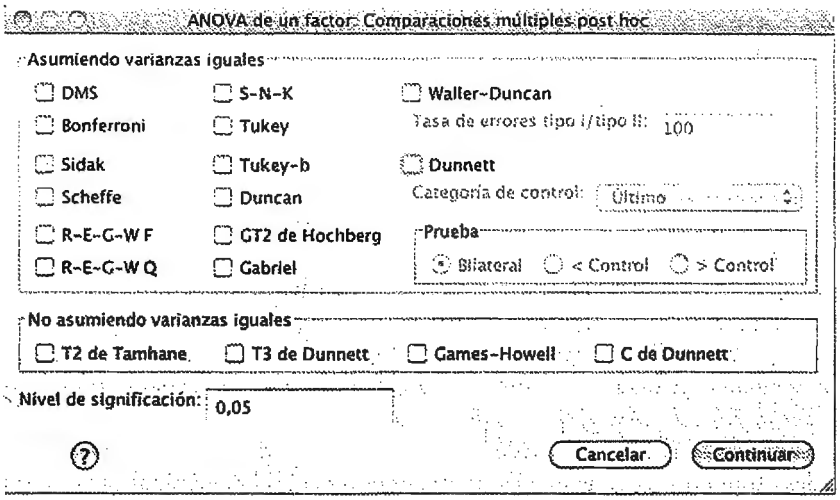



**Figura 8.6** Opciones para los contrastes *post hoc* en SPSS. Se debe seguir la siguiente ruta en el menú: Analizar → Comparar medias → ANOVA de un factor... → post hoc...

Se obtendría un contraste *post hoc* penalizado por el método de Bonferroni, de Tukey y de Dunnett, con la siguiente sintaxis:

```
ONEWAY brain BY oh
/POST=BONFERRONI
/POST=TUKEY
/POST=DUNNETT(1).
```

Al solicitar el método de Dunnett, se ha fijado como categoría de referencia (frente a la que se compararán todas las demás) el primer grupo (abstemios, OH = 1); por eso se indica el (1) entre paréntesis. Si no se procede así, por omisión, SPSS elegirá el último grupo como referencia (v. fig. 8.6).

En R se puede ejecutar directamente las comparaciones ajustadas mediante el método de Bonferroni.

```
pairwise.t.test(brain, oh, p.adj = "bonf")
```

Para otros métodos se debe crear primero el objeto que contiene el ANOVA:

```
Resultado <-aov(brain~oh)  # se debe haber usado attach
(OH_brain) previamente.
```

Posteriormente se puede realizar la comparación deseada sobre el objeto:

```
TukeyHSD(Resultado)
```

Para el test de Dunnett, es necesario cargar la librería **multcomp** y posteriormente ejecutar el comando:

```
library(multicop)
summary(glht(Resultado, linfct = mcp(gr = "Dunnett")))
```

## 8.10. MÉTODO DE BENJAMINI-HOCHBERG BASADO EN ORDENAR LOS VALORES *P*

Como se ha visto antes, se deben distinguir dos conceptos:

1. Riesgo $\alpha$ *particular* (el convencional): probabilidad de equivocarse al rechazar una $H_0$ que es verdadera, cuando se valora solo esa hipótesis nula.
2. Riesgo $\alpha$ *global*: probabilidad de rechazar equivocadamente al menos una $H_0$ en comparaciones múltiples con muchas hipótesis nulas, todas ellas verdaderas.

La figura 8.7 representa el riesgo $\alpha$ global a medida que aumenta el número de comparaciones, y en todas ellas $H_0$ es verdad. La contemplación atenta de esta figura ayudará mucho a entender por qué las comparaciones múltiples pueden crear un problema.

Además del riesgo $\alpha$ global, cuyo numerador es el número de errores tipo 1 y cuyo denominador es el total de hipótesis valoradas, en el supuesto de que $H_0$ siempre sea verdad existe otro concepto, que es la *tasa de falsos descubrimientos* (FDR o *false discovery rate*).

Se entiende por FDR el porcentaje de las decisiones a favor de la hipótesis alternativa que son equivocadas. Aquí no se supone que $H_0$ sea siempre verdad. El denominador no es el total de hipótesis valoradas, sino el total de *decisiones* a favor de $H_1$ (es decir, el total de *descubrimientos*). La tabla 8.8 explica la diferencia entre el error alfa y el FDR.

El procedimiento de Benjamini-Hochberg **(10-12)** surge para controlar el FDR. Como los demás métodos vistos, corrige la significación estadística cuando se han hecho muchas comparaciones (*c*), pero es más creativo, porque va *gastando* poco a poco de una manera inteligente y progresiva el riesgo $\alpha$. En vez de que acptar que siempre que $p < 0{,}05$ el valor será estadísticamente significativo, se usará un umbral o *valor crítico* distinto de 0,05 en cada comparación. *En cada comparación* va cambiando este valor crítico (es decir, el riesgo $\alpha$), y así consigue que el riesgo $\alpha$ global se mantenga en el 5%. Se basa en ordenar ascendentemente los valores *p* obtenidos y asignar rangos (i):

c = número total de comparaciones hechas.

i = puesto (n.° de orden) de cada valor *p* obtenido.

(i = 1 para el menor, i = c para el mayor).

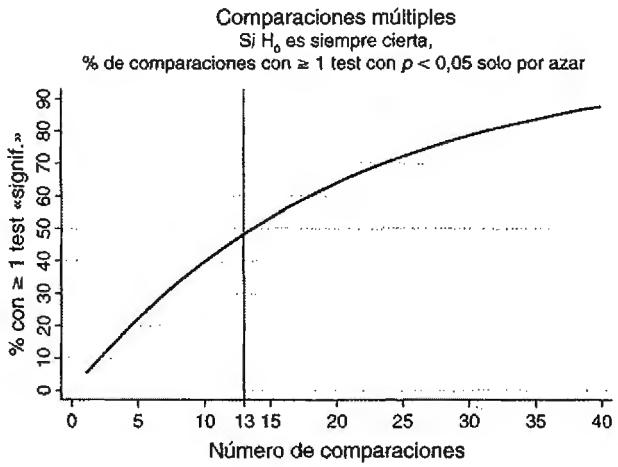



**Figura 8.7** Probabilidad de que al menos un test resulte significativo cuando se hacen muchos test en un universo donde todas las hipótesis nulas son verdaderas (ningún test debería resultar, por tanto, significativo). Cuando se hacen más de 13 test, la probabilidad de que alguno de ellos resulte equivocadamente significativo solo por azar es mayor del 50%.

*Tabla 8.8 Riesgo α y tasa de descubrimientos falsos (false discovery rate, FDR) cuando se han hecho 100 contrastes de hipótesis*

| | VERDAD | | |
|---|---|---|---|
| DECISIÓN | $H_0$ | $H_1$ | TOTAL |
| $H_0$ | 40 | 25 | 65 |
| $H_1$ | 10 | 25 | 35 |
| Total | 50 | 50 | |
| Riesgo $\alpha$ = 10/50 | | | |
| FDR = 10/35 | | | |

El valor crítico que sustituirá al riesgo $\alpha$ para cada comparación es:

$$\text{Valor crítico}(\approx \alpha) = 0{,}025\frac{i}{c}$$

Ejemplo: Hosking y Danthiir (**13**) valoraron la asociación de patrones dietéticos a lo largo de la vida con predictores sociodemográficos. Hicieron 12 comparaciones de la relación entre 12 de estos patrones y la edad. Los 12 valores *p* obtenidos, ordenados de menor (i = 1) a mayor (i = 10), iban desde $p < 0{,}001$ a $p = 0{,}987$.

La tabla 8.9 recoge estos valores *p* (sin corregir por comparaciones múltiples) en la primera columna, ordenados de menor a mayor. La segunda columna es su número de orden (i), y la tercera contiene el umbral o valor crítico que los autores deberían haber utilizado para considerar si era significativa o no cada *p* encontrada, teniendo en cuenta las comparaciones múltiples según el procedimiento de Benjamini-Hochberg. *Solo cuando la p encontrada sea inferior al valor crítico se podrá afirmar que hay significación estadística.* Este es el precio que se debe pagar por haber realizado muchas comparaciones. Así, aunque en el artículo original parecía que las seis primeras comparaciones eran significativas, realmente solo la primera de ellas superaba el umbral de la significación al corregir mediante múltiples tests con este procedimiento. Como se observa en este ejemplo, el procedimiento de Benjamini-Hochberg va más allá del ANOVA y se puede aplicar en cualquier situación en la que se hayan efectuado muchos test.

*Tabla 8.9 Método de Benjamini-Hochberg para c = 12 comparaciones*

| *P* ENCONTRADA | I | VALOR CRÍTICO | *P'* < 0,05 |
|---|---|---|---|
| *p* ordenadas | n.° | $0{,}025\frac{i}{c}$ | |
| <0,001 | 1 | 0,0025 | sí |
| 0,011 | 2 | 0,0042 | no |
| 0,012 | 3 | 0,0063 | no |
| 0,013 | 4 | 0,0083 | no |
| 0,024 | 5 | 0,0104 | no |
| 0,047 | 6 | 0,0125 | no |
| 0,053 | 7 | 0,0146 | no |
| 0,060 | 8 | 0,0167 | no |
| 0,083 | 9 | 0,0188 | no |
| 0,652 | 10 | 0,0208 | no |
| 0,795 | 11 | 0,0229 | no |
| 0,987 | 12 | 0,0250 | no |

*Solo cuando el valor* p *encontrado sea inferior al valor crítico se podrá considerar significativo.*

## 8.11. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Test | STATA | SPSS |
|---|---|---|
| Comprobación de normalidad de residuales | **quietly summarize *vdep* ///**<br>**if *gr*==1**<br>**g *resid* = *vdep*-r(mean) ///**<br>**if *gr*==1** (etc.)<br>**ladder *resid***<br>**pnorm *resid*** | **IF gr = 1 resid = vdep-##**(media gr1)<br>(etc.)<br>**EXE.**<br>**EXAM VAR = resid**<br>**/PLOT NPPLOT.** |
| ANOVA | **oneway *vdep gr*, tab** | **ONEWAY vdep BY gr**<br>**/STAT DESCR HOMOG.** |
| Kruskal-Wallis | **kwallis *vdep*, by(*gr*)** | **NPAR TESTS**<br>**/K-W = vdep BY gr(1 3).** |
| Contrastes *a priori* | **anova *vdep* gr**<br>**matrix *C1* = (-2,0,1,1,0)**<br>**test, test(*C1*)** | **ONEWAY vdep BY gr**<br>**/CON = -2 0 1 1**<br>**/CON = 0 -2 1 1**<br>**/CON = 0 0 -1 1**<br>**/STAT HOMOG.** |
| Contrastes *post hoc* | **oneway *vdep gr*, sidak**<br>**oneway *vdep gr*, scheffe**<br>**oneway *vdep gr*, bonferroni** | **ONEWAY vdep BY gr**<br>**/POST = BONFERRONI**<br>**/POST = TUKEY**<br>**/POST = DUNNETT(1).** |

## REFERENCIAS


1. Paul CA, Au R, Fredman L, Massaro JM, Seshadri S, Decarli C, et al. Wolf PA. Association of alcohol consumption with brain volume in the Framingham study. Arch Neurol 2008;65(10):1363-7.
2. Altman DG, Bland JM. Comparing several groups using analysis of variance. BMJ 1996;312(7044):1472-3.
3. Martín Andrés A, Luna del Castillo JD. Bioestadística para las Ciencias de la Salud. 5.ª ed. Madrid: Norma-Capitel; 2004.
4. Bender R, Lange S. Adjusting for multiple testing- when and how? J Clin Epidemiol 2001;54(4):343-9.
5. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1(1):43-6.
6. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ 2001;322(7292):989-91.
7. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357(21):2189-94.
8. Hunter DJ, Kraft P. Drinking from the fire hose -- statistical issues in genomewide association studies. N Engl J Med 2007;357(5):436-9.
9. Van den Oord EJ. Controlling false discoveries in genetic studies. Am J Med Genet B Neuropsychiatr Genet 2008;147(5):637-44.

10. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.

11. Benjamini Y, Hochberg. Y Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995;57:289-300.

12. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. J Educ Behav Stat 2002;27:77-83.

13. Hosking D, Danthiir V. Retrospective lifetime dietary patterns are associated with demographic and cardiovascular health variables in an older community-dwelling Australian population. Br J Nutr 2013 Dec;110(11):2069-83.

# ANOVA FACTORIAL, MODELOS LINEALES GENERALIZADOS Y ANCOVA

9

*C. Sayón-Orea, E. Toledo, J. M. Núñez-Córdoba, M. Á. Martínez-González*

## 9.1. INTRODUCCIÓN

En el capítulo anterior se explicó el ANOVA de una vía (o de un factor), que es el procedimiento más adecuado para comparar más de dos medias entre sí, y su equivalente no paramétrico (test de Kruskal-Wallis). Se trataba de una sola variable independiente (o *factor* de agrupación) con más de dos categorías (policotómica). Sin embargo, había un solo factor independiente.

Cuando los grupos se pueden clasificar atendiendo a más de un criterio (factor) aparecen los diseños *factoriales*, en los que hay más de una variable de agrupación. Es decir, valoran combinaciones de *dos o más factores*. Los diseños factoriales pertenecen a un grupo más amplio de procedimientos estadísticos conocidos como *modelos lineales generalizados*. El ANOVA de dos vías es el modelo más simple, ya que solo hay dos factores o variables independientes (de agrupación) y una variable dependiente (la que se compara). Aunque en el ejemplo que se presentará los factores tengan únicamente dos categorías, hay que tener en cuenta que en un ANOVA de dos factores cualquiera de los dos factores puede ser policotómico ($k \geq 3$).

En todo lo dicho hasta ahora se ha hablado de grupos independientes, pero a veces hay diseños relacionados o emparejados (medidas repetidas en el mismo sujeto). El ANOVA de medidas repetidas sirve para analizar tales diseños de $k$ medias relacionadas y tiene un equivalente no paramétrico (test de Friedman). En este capítulo se analizarán los ejemplos más elementales de ANOVA factorial y ANOVA de medidas repetidas. Se verán los cuatro procedimientos sombreados en el esquema mostrado en la figura 9.1.

## 9.2. ANOVA FACTORIAL (DE DOS VÍAS O DOS CRITERIOS)

### 9.2.1. Planteamiento e hipótesis en el ANOVA factorial

Se desea estimar la eficacia de dos métodos para perder peso. En el siguiente ejemplo ficticio, se han formado cuatro grupos, cada uno de 5 individuos que deseaban perder peso. Se han combinado 2 factores: ejercicio físico (sí/no) y dieta (control/hipocalórica). En la tabla 9.1 se observan los datos, en los que para cada individuo se indica en cuántos kilogramos varió su peso (peso final – peso inicial) tras finalizar el tratamiento:

La situación sería la siguiente:

- En la base de datos existirían, al menos, 3 variables:
  - Cambio de peso (variable *«dependiente»*): cuantitativa.
  - Ejercicio *(factor 1):* categórica (sí/no).
  - Dieta *(factor 2):* categórica (control/hipocalórica).
- Las hipótesis que deben comprobarse serían tres:
  - Efecto del factor ejercicio sobre el cambio de peso.

$$H_0 \equiv \mu_{\text{con ejercicio}} = \mu_{\text{sin ejercicio}}$$

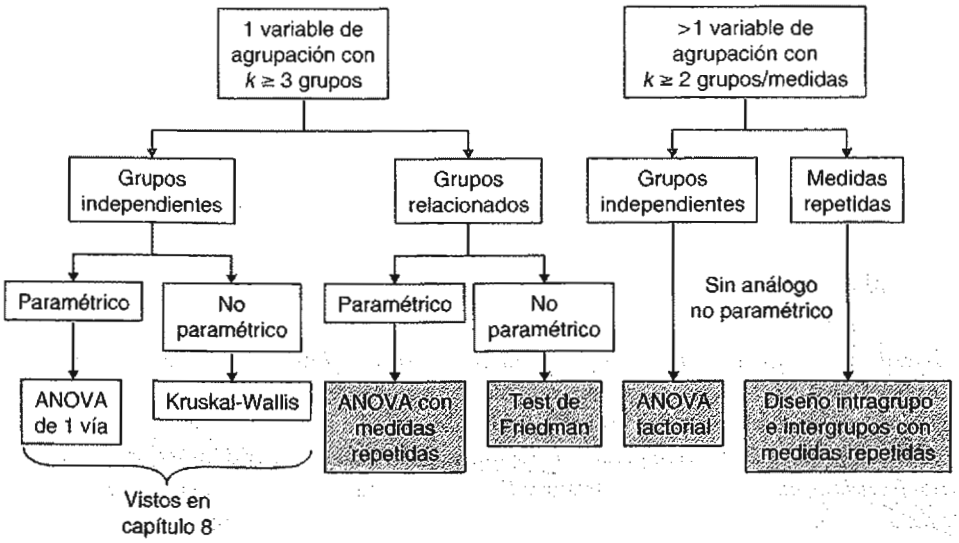



**Figura 9.1** Comparación de $k$ medias.

*Tabla 9.1 Cambio de peso (kg) al finalizar el tratamiento*

| | CONTROL (NO DIETA) | DIETA HIPOCALÓRICA |
|---|---|---|
| **Ejercicio no** | −1 | −6 |
| | −3 | −10 |
| | +2 | −3 |
| | +2 | −2 |
| | +4 | −8 |
| **Ejercicio sí** | −3 | −10 |
| | −5 | −12 |
| | +3 | −15 |
| | −1 | −16 |
| | −3 | −9 |

- Efecto del factor dieta sobre el cambio de peso.

$$H_0 \equiv \mu_{\text{con dieta}} = \mu_{\text{sin dieta}}$$

- Diferencias en el efecto del ejercicio según la dieta (*interacción:* ejercicio × dieta).

$$H_0 \equiv (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{con ejercicio}} = (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{sin ejercicio}}$$

Gráficamente se representaría como en la figura 9.2.

En esta situación se debe recurrir al análisis de la varianza de dos vías o dos criterios. Cabe señalar que existen varias observaciones por casilla, tal como se muestra en la tabla 9.1, porque en cada casilla formada por la combinación de ejercicio y dieta hay más de un individuo.

En el ejemplo anterior se observa que, en cada grupo, se tienen en cuenta simultáneamente dos factores: dieta, con 2 niveles (control/hipocalórica), y ejercicio, también con 2 niveles (sí/no). Los efectos causados por estos factores se denominan *efectos principales.* En este tipo de diseños, la combinación de factores lleva a la aparición de otro efecto importante, que se considera el primero que debe explorarse: la *interacción* (**1-3**). Se dice que existe *interacción* entre los 2 factores cuando

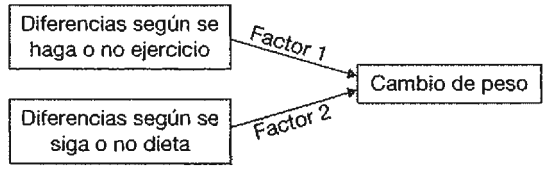



**Figura 9.2** ANOVA de dos vías o dos criterios.

las diferencias entre los niveles de un factor (p. ej., entre seguir una dieta control o una dieta hipocalórica) varían en función del nivel del segundo factor que esté presente (no ejercicio/sí ejercicio). Es decir, el efecto de un factor depende del efecto del otro. Es una relación siempre recíproca.

En la figura 9.3 se representa cómo serían los efectos de ambos factores *si no hubiese interacción entre ellos.* Los individuos sometidos a dieta perderían más peso que los que no hacen dieta. Además, el efecto de la dieta en el cambio de peso sería el mismo tanto para los que realizasen ejercicio como para los que no lo hicieran, es decir, el efecto de la dieta es independiente de que se haga o no ejercicio. La diferencia entre los grupos (sí ejercicio/no ejercicio) se mantendría constante fuese cual fuese la dieta seguida.

Sin embargo, si hubiese *interacción,* se producirían situaciones como la representada en la figura 9.4. En esa figura *sí* existe interacción. En este caso, la interacción consiste en que el efecto

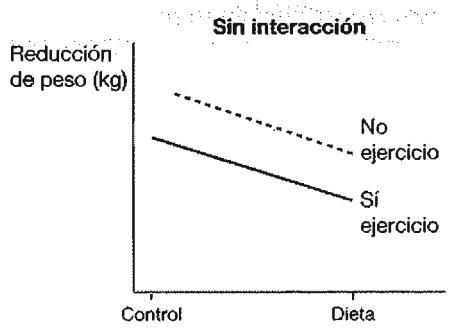



**Figura 9.3** Una de las posibles situaciones donde no existe interacción: el efecto de la dieta sobre el peso es el mismo sea cual sea el nivel de ejercicio.

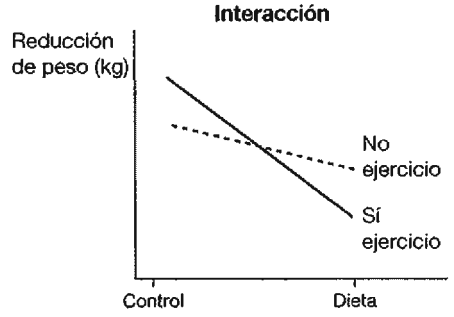



**Figura 9.4** Una de las posibles situaciones donde existe interacción: el efecto de la dieta sobre el peso es mayor si se hace ejercicio.

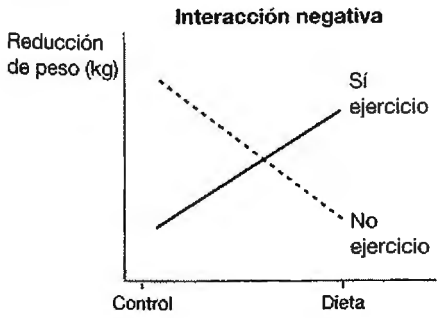



**Figura 9.5** Existe interacción negativa: el efecto de la dieta sobre el peso se invierte de sentido (interacción cualitativa) si se hace ejercicio.

de la dieta sería mayor cuando se realiza simultáneamente ejercicio físico. Es decir, habría más diferencias entre los del grupo de dieta y los del grupo control si los individuos realizasen ejercicio a la vez. La interacción refleja en este caso la existencia de una *sinergia* o potenciación entre dos factores. Si el resultado fuese el indicado, se concluiría que un programa de dieta resultaría más eficaz en individuos que hacen ejercicio.

También podría ocurrir exactamente lo contrario, como muestra la figura 9.5: el efecto de la dieta es distinto en quienes realizan ejercicio que en quienes no lo realizan. Entonces, la interacción sería negativa o antagónica. Si se obtuviera este resultado, se concluiría que la dieta ensayada es eficaz en individuos que no hacen ejercicio, pero perjudicial en los que hacen ejercicio.

Así, en este tipo de diseños hay 4 componentes de la variabilidad total:

1. Debido al primer factor: DIETA.
2. Debido al segundo factor: EJERCICIO.
3. Interacción entre ambos.
4. Residual (que coincide con el concepto ya estudiado en el capítulo anterior).

Las fuentes de variabilidad 1) y 2) son los *efectos principales*. La interacción (punto 3), puede explicarse como un tercer efecto añadido a la suma de DIETA + EJERCICIO, que solo actúa cuando los otros dos (efectos principales) están presentes. La suma de 1)+ 2) + 3) es *lo explicado* por los factores que se han tenido en consideración (variabilidad *intergrupos*). El *residual* es lo que queda sin explicar (variabilidad *intragrupo* o *error*) tras haber tenido en cuenta los otros tres componentes.

Más adelante se explicarán los pasos que deben seguirse para realizar un ANOVA de dos factores con los distintos programas estadísticos (v. apartados 9.8 y 9.9).

Por ahora, simplemente se muestra el resultado de esta comparación, donde se observa que la interacción ***(dieta#ejercicio)*** tiene un valor $p = 0{,}163$, cuya interpretación es que no existe interacción, es decir, no podemos rechazar la hipótesis nula. Por otro lado, el valor $p = 0{,}004$ permite rechazar la hipótesis nula de que el ejercicio no tiene efecto sobre el cambio de peso; a su vez, el valor $p < 0{,}001$ también lleva a rechazar la hipótesis nula de que la dieta no tiene efecto sobre el cambio de peso. En general, se concluiría que tanto el ejercicio como la dieta tienen un efecto significativo sobre el cambio de peso.

```
      Number of obs =       20     R-squared     =  0.7681
      Root MSE      =  3.05778     Adj R-squared =  0.7247
```

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 495.6 | 3 | 165.2 | 17.67 | 0.0000 |
| dieta | 369.8 | 1 | 369.8 | 39.55 | 0.0000 |
| ejercicio | 105.8 | 1 | 105.8 | 11.32 | 0.0040 |
| dieta#ejercicio | 20 | 1 | 20 | 2.14 | 0.1630 |
| Residual | 149.6 | 16 | 9.35 | | |
| Total | 645.2 | 19 | 33.9578947 | | |

### 9.2.2. Interpretación de los test de interacción y precauciones en el análisis de subgrupos

En la tabla 9.2 se aprecia con más detalle el concepto de *interacción*. Se trata de una tabla 2 × 2, donde se cruzan las dos variables y se forman cuatro grupos, según se esté expuesto o no cada uno de los dos factores. Se presenta la media del cambio de peso en cada grupo, redondeando las cifras para facilitar la comprensión de la situación. Quienes no hacen dieta y tampoco ejercicio aumentan en 1 kg (+1 kg) de peso; los que hacen dieta, pero no ejercicio, reducen su peso en 6 kg (–6 kg), y quienes hacen ejercicio, pero no dieta, reducen su peso en 2 kg (–2 kg). Puede apreciarse que hay 7 kg de diferencia en el cambio de peso entre quienes hacen dieta y los que no la hacen, es decir, el tránsito de no hacer dieta a hacerla supone una diferencia de –7 kg en el cambio de peso; el tránsito de no hacer ejercicio a hacerlo supone una diferencia de –3 kg. ¿Qué esperaríamos en los expuestos simultáneamente a ambos factores?

En un modelo de ANOVA, *en ausencia de interacción,* lo esperado para la media del grupo expuesto a ambos factores (ejercicio + dieta) sería que *se sumasen* los dos efectos. Si se contempla el problema desde el punto de vista de la dieta (primero en una fila de la tabla y luego en otra), pasar a hacer dieta supone reducir el peso en 7 kg entre los que no hacen ejercicio. Por tanto, si no existiese interacción, también se reduciría el peso en 7 kg al pasar de no hacer dieta a hacerla entre los que sí hacen ejercicio (fig. 9.6). En ausencia de interacción, el efecto esperado en los que hacen ejercicio sería también –7 kg, y sería de esperar una media de cambio de peso de –9 kg en los expuestos a ambos factores.

Lo mismo se esperaría si se contemplase el problema desde el punto de vista del ejercicio (primero en una columna y luego en otra), ya que pasar de no hacer ejercicio a hacerlo supone perder 3 kg entre los que no hacen dieta. Por lo tanto, en ausencia de interacción, entre los que sí hacen dieta también se esperarían 3 kg menos, y pasarían de –6 a –9 kg.

Se dice que no hay interacción si lo que sucede cuando se combinan ambos factores es lo esperado meramente por la suma de esos dos factores. En cambio, si lo que se obtiene para la combinación de ambos factores es significativamente distinto de lo esperado por la simple suma e sus efectos, se dirá que existe *interacción*. Nunca existirá estrictamente una correspondencia exacta con lo esperado por la suma de los efectos de los 2 factores. El problema de la interacción

***Tabla 9.2 Medias del cambio de peso (kg) según se siga una dieta y/o se haga ejercicio físico***

| | DIETA CONTROL | DIETA HIPOCALÓRICA |
|---|---|---|
| Ejercicio no | +1 | –6 |
| Ejercicio sí | –2 | ¿Esperado? |

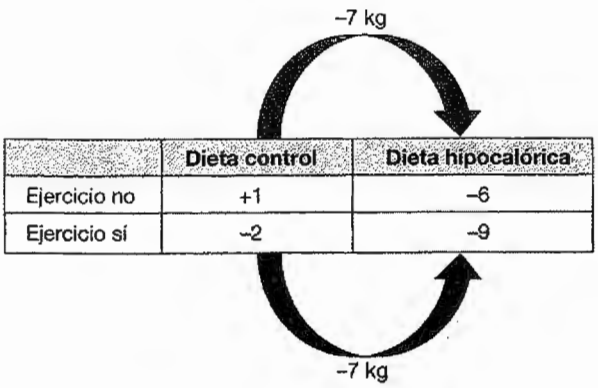



**Figura 9.6** Medias del cambio de peso (kg) en ausencia de interacción.

se resuelve considerando si el apartamiento de lo esperado por la suma puede ser explicado simplemente por la variabilidad aleatoria, como suele ser habitual en los contrastes de hipótesis. El ANOVA factorial incluye un test para averiguar si la interacción es estadísticamente significativa. Se considera que lo observado es significativamente distinto de lo esperado si el valor *p* es inferior a 0,05 (tabla 9.3).

La existencia de interacción complica la interpretación de los resultados, ya que el efecto de un factor depende de que el otro factor esté presente. Una interacción significativa quizá puede llevar a tener que dividir en dos la base de datos y realizar análisis separados para cada factor. Esta opción recibe el nombre de *análisis de subgrupos* (4,5). El análisis de subgrupos consiste en repetir la misma comparación dentro de cada subgrupo de participantes. Los subgrupos vendrán definidos en función de alguna característica (sexo, grupo de edad, gravedad, presencia de alguna enfermedad concomitante, etc.). Este análisis de subgrupos no se debe hacer indiscriminadamente, sino solo cuando se planificó (para muy pocos subgrupos) *a priori* o cuando una interacción resulte estadísticamente significativa.

Cuando se observan efectos opuestos en los subgrupos, por ejemplo, si en un subgrupo (digamos en mujeres) el efecto del factor bajo estudio es protector, mientras que en el otro subgrupo (hombres) resulta perjudicial, entonces se habla de interacción *cualitativa* (4). La interacción *cuantitativa* simplemente llevaría a que el efecto fuera significativamente más fuerte en un subgrupo que en otro. La interacción *cualitativa* en cambio significa que el efecto cambia de sentido.

Existe un peligro para los investigadores, que es valorar el efecto en demasiados subgrupos. Esto es más grave cuando en un ensayo globalmente no se encuentra ningún efecto, pero los investigadores intentan encontrarlo en algún subgrupo peculiar. Esta situación ocurre con más frecuencia en ensayos financiados por industrias que pueden estar interesadas en tales efectos especiales dentro de subgrupos, lo que puede crear sesgos (5). Entonces se produce una inflación del

***Tabla 9.3 Medias del cambio de peso (kg) cuando existe interacción***

| | DIETA CONTROL | DIETA HIPOCALÓRICA |
|---|---|---|
| Ejercicio no | +1 | –6 |
| Ejercicio sí | –2 | Cualquier resultado distinto de –9 ($p < 0,05$) |

error tipo 1. Para prevenirlo, se ha propuesto limitar mucho los análisis de subgrupos y penalizar los valores *p* que se encuentren para el efecto principal dentro de cada subgrupo con un esquema similar al de Bonferroni (4-7). En cualquier caso, hay que ser cauto al interpretar un valor *p* significativo dentro de un subgrupo. Del mismo modo, se debe ser conservador al interpretar los valores *p* de los test de interacción que se analizarán en este capítulo. No se debe aceptar como importante una interacción simplemente porque el valor *p* para dicha interacción sea <0,05. La significación tiene que estar clara y, además, ha de haber una magnitud en la diferencia del efecto entre los subgrupos que interaccionan que sea clínicamente relevante.

A pesar de lo dicho, cuando se tienen en cuenta varios factores a la vez, debe valorarse el test de interacción entre ellos, porque la interpretación variará mucho dependiendo de si hay interacción cualitativa o, al menos, una interacción cuantitativa que sea fuerte (8). En cambio, cuando la interacción no resulta significativa, un diseño factorial representa un ahorro de tiempo y esfuerzo, porque se obtiene simultáneamente información sobre 2 factores con menor número de participantes que si se hubiesen hecho dos estudios.

### 9.2.3. Supuestos del modelo de ANOVA factorial

Antes de aplicar el test de ANOVA factorial, se ha de comprobar previamente que se cumplen los supuestos básicos del ANOVA, como la normalidad e igualdad de varianzas (homoscedasticidad), del mismo modo que se explicó en el capítulo anterior (análisis de residuales). No obstante, cuando no se cumplen estas condiciones y la muestra es pequeña, en este caso existe un problema mayor, al no disponer de buenos test no paramétricos que sustituyan al ANOVA factorial.

## 9.3. ANOVA CON MEDIDAS REPETIDAS (COMPARACIÓN DE *K* MEDIAS RELACIONADAS)

En ocasiones, en una investigación se toma más de una medición de una misma característica en cada sujeto. Por ejemplo, se puede medir el peso corporal antes de empezar una dieta, al cabo de 1 mes y a los 6 meses. Nos interesa saber si la dieta hace que cambie el peso corporal. Hay que comparar, por tanto, la media de la misma variable (peso corporal) medida en tres ocasiones distintas en los mismos sujetos.

Cuando hay que comparar más de 2 medidas repetidas de un desenlace y se desea responder a la pregunta *«¿la media de la variable desenlace cambia a lo largo del tiempo?»*, se podría proceder mediante dos aproximaciones: a través de una aproximación *multivariante* mediante un MANCOVA (análisis multivariante de la varianza) de medidas repetidas, conocido también como modelo lineal generalizado para medidas repetidas, o bien mediante una aproximación *univariante* conocida como ANOVA de medidas repetidas. Esta aproximación es comparable a lo expuesto en el capítulo anterior sobre ANOVA de una vía, que se basa en la suma de cuadrados.

Las condiciones de aplicación de un ANOVA de medidas repetidas son:

1. Las observaciones tienen que ser independientes.
2. Las variables de medidas repetidas deben seguir una distribución normal.
3. *Esfericidad,* que implica, en primer lugar, que todas las correlaciones de la variable desenlace entre las medidas repetidas sean iguales, independientemente del intervalo de tiempo entre las medidas, y, en segundo lugar, que las varianzas de la variable desenlace sean las mismas en cada una de las medidas repetidas.

La esfericidad sería el análogo a la igualdad de varianzas (homogeneidad de varianzas) vista en el apartado 6.2.

La condición de *esfericidad* se expresa mediante el *coeficiente épsilon (ε).* La situación ideal sería que $\varepsilon = 1$; cuando esta condición no se cumple, dicho coeficiente valdrá menos de 1. De esta

forma, si se cuenta con más de dos mediciones ($k > 2$), será preciso realizar una corrección en el valor $p$ del ANOVA de medidas repetidas para tener en cuenta la *esfericidad*. En muchos programas estadísticos, el coeficiente de esfericidad se calcula automáticamente. STATA, por ejemplo, calcula $\varepsilon$ mediante tres métodos: 1) Huynh-Feldt; 2) Greenhouse-Geisser, y 3) *Box's conservative*. Los tres difieren un poco entre sí; se recomienda utilizar el ajuste de Greenhouse-Geisser, aunque sea un tanto conservador (**9**).

Ahora bien, la principal hipótesis nula en el ANOVA de medidas repetidas es que no hay cambio de una medida a otra, es decir, que en la población se mantiene constante la media de esa variable de desenlace a lo largo de todos los momentos de tiempo en que está siendo observada. Este ANOVA para medidas repetidas se puede conceptualizar como una extensión del test de la $t$ de Student para datos emparejados. En ambas situaciones, como en cualquier diseño emparejado, se tiene la ventaja de que las comparaciones que se efectuarán estarán basadas en diferencias dentro de cada sujeto (autoemparejamiento). Así, se reduce el «ruido» o variabilidad aleatoria. Como la variación entre sujetos suele ser mucho mayor que la variación intrasujeto (es decir, de un tiempo a otro en un mismo sujeto), al prescindir de la variación entre sujetos se afina más la capacidad de detectar diferencias, porque el objeto es lo que ocurre dentro de cada sujeto. Cada sujeto es, de alguna manera, su propio control.

Al final, como es habitual en los procedimientos de ANOVA, se calculará una cantidad F, que es el cociente entre una varianza explicada por las $k$ repeticiones de la medición (efecto) y otra varianza que se llama residual (error), no explicada por las distintas mediciones. Para calcular la varianza residual habrá que extraer, de la variabilidad total, además de la variabilidad entre repeticiones de la medición (dentro de cada sujeto), la variabilidad entre sujetos. En este caso, las «mediciones repetidas» que se realizan para cada sujeto se asemejan a los «grupos» que se vieron al tratar del ANOVA de un factor. Como existen dos fuentes de variación además de la residual, el problema es bastante similar al análisis de la varianza de dos vías.

En el cuadro 9.1 y en la figura 9.7 están representadas la variabilidad residual y la descomposición de las sumas de cuadrados en el ANOVA de una vía. Esta descomposición se compara con las que se efectúan en el ANOVA de dos vías y en el ANOVA de medidas repetidas.

A continuación se utilizará un ejemplo ilustrativo. Algunos estudios han demostrado que existe una relación inversa entre la adherencia a la dieta mediterránea y la enfermedad coronaria. Se sabe que los efectos beneficiosos de esta dieta se deben, en buena parte, a que su principal fuente de grasa es el aceite de oliva (virgen extra, en el ejemplo). Por tanto, es lógico pensar que el

### CUADRO 9.1 VARIABILIDAD RESIDUAL EN EL ANOVA DE MEDIDAS REPETIDAS COMPARADO CON EL ANOVA DE UNO Y DOS FACTORES

**En el ANOVA de un factor**

Variabilidad residual = Variabilidad total – variabilidad entre grupos

**En el ANOVA de dos factores**

Variabilidad residual = Variabilidad total – (variabilidad entre grupos del factor 1 + variabilidad entre grupos del factor 2 + variabilidad de la interacción)

**En el ANOVA para medidas repetidas**

Variabilidad residual = Variabilidad total – (variabilidad entre medidas + variabilidad entre sujetos)

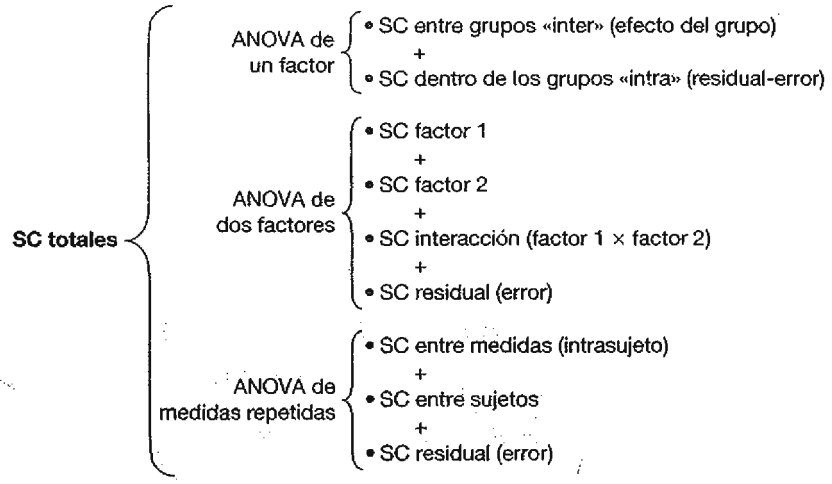



**Figura 9.7** Descomposición de la suma de cuadrados (SC) en el ANOVA de un factor, de dos factores y de medidas repetidas.

aumento del consumo de aceite de oliva es beneficioso para la salud. De esta forma, un grupo de investigadores estudió la efectividad de una intervención nutricional para aumentar el consumo de aceite de oliva en sujetos con alto riesgo cardiovascular (**10,11**). Los investigadores querían saber si una intervención nutricional podía modificar el consumo de aceite de oliva. Midieron el consumo de aceite de oliva (en gramos) en tres momentos: basal (previo a la intervención), a los 3 meses y al año de la intervención. Los datos de la tabla 9.4 son ficticios (se han simplificado y reducido notablemente para no complicar el ejemplo), pero están inspirados en el trabajo real de este grupo.

Se tienen así $k = 3$ medidas repetidas del consumo de aceite de oliva (***oliva1***, ***oliva2*** y ***oliva3***) en cada participante. La siguiente pregunta que se plantea es: ¿existen diferencias a lo largo del tiempo en el consumo de aceite de oliva?

En el ejemplo anterior, $H_0$ sería:

$$H_0 \equiv \mu_{\text{oliva1}} = \mu_{\text{oliva2}} = \mu_{\text{oliva3}}$$

Bajo $H_0$, las variaciones con respecto a la media total serían explicables solo por azar.

A continuación, se detallan los pasos que deben seguirse para realizar un ANOVA de medidas repetidas con STATA.

En primer lugar, se deberá hacer un **reshape** en la base de datos, para pasar de un formato *«wide»* a un formato *«long»* mediante la siguiente orden:

```
reshape long oliva, i(id) j(tiempo)
```

donde:

**i(*id*)** corresponde al nombre de la variable que identifica a los sujetos dentro de la base de datos; en el ejemplo es id.

**j(*tiempo*)** es una variable nueva que se creará y tendrá tantos valores como mediciones existan; en el ejemplo, tomará valores de 1 a 3, porque hay tres mediciones.

*Tabla 9.4 Consumo de aceite de oliva (g), medidos en tres momentos (basal, a los 3 meses y al año de intervención)*

| SUJETO (id) | BASAL (oliva1) | A LOS 3 MESES (oliva2) | A 1 AÑO (oliva3) |
|---|---|---|---|
| 1 | 10 | 2 | 11 |
| 2 | 12 | 3 | 10 |
| 3 | 4 | 4 | 11 |
| 4 | 12 | 14 | 13 |
| 5 | 20 | 14 | 10 |
| 6 | 11 | 24 | 18 |
| 7 | 20 | 13 | 27 |
| 8 | 29 | 10 | 15 |
| 9 | 17 | 15 | 10 |
| 10 | 10 | 9 | 12 |
| 11 | 4 | 14 | 8 |
| 12 | 15 | 20 | 18 |
| 13 | 5 | 8 | 14 |
| 14 | 2 | 4 | 6 |
| 15 | 18 | 19 | 22 |
| 16 | 8 | 21 | 20 |
| 17 | 6 | 10 | 22 |
| 18 | 6 | 12 | 24 |
| 19 | 6 | 22 | 24 |
| 20 | 6 | 30 | 28 |
| 21 | 8 | 13 | 28 |
| 22 | 10 | 10 | 22 |
| 23 | 8 | 12 | 24 |
| 24 | 10 | 15 | 18 |
| 25 | 12 | 16 | 30 |
| **Media de cada medida (columna)** | **10,76** | **13,36** | **17,80** |

En este nuevo formato *long*, la base de datos pasará a tener 75 observaciones en lugar de 25, porque ahora cada sujeto se ha multiplicado por 3. La nueva variable ***tiempo*** tendrá tres posibles valores (1, 2 y 3), que corresponden a los diferentes momentos (basal, a los 3 meses y al año) y, en cambio, habrá una única variable de medida ***(oliva)***.

```
. reshape long oliva, i(id) j(tiempo)
(note: j = 1 2 3)

Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        25   ->     75
Number of variables                    4   ->      3
j variable (3 values)                      ->   tiempo
xij variables:
                          oliva1 oliva2 oliva3   ->   oliva
```

Una vez que la base de datos se encuentra en formato **long**, se podrá realizar el test de medidas repetidas mediante la siguiente orden:

```
anova olive id tiempo, repeat(tiempo)
```

```
. anova oliva id tiempo, repeat(tiempo)
```

Number of obs = 75 R-squared = 0.5627
Root MSE = 5.94717 Adj R-squared = 0.3258

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 2184.24 | 26 | 84.0092308 | 2.38 | 0.0047 |
| id | 1550.61333 | 24 | 64.6088889 | 1.83 | 0.0378 |
| tiempo | 633.626667 | 2 | 316.813333 | 8.96 | 0.0005 |
| Residual | 1697.70667 | 48 | 35.3688889 | | |
| Total | 3881.94667 | 74 | 52.4587387 | | |

Between-subjects error term: id
Levels: 25 (24 df)
Lowest b.s.e. variable: id

Repeated variable: tiempo

Huynh-Feldt epsilon = 0.9486
Greenhouse-Geisser epsilon = 0.8836
Box's conservative epsilon = 0.5000

| | | | Prob > F | | | |
|---|---|---|---|---|---|---|
| Source | df | F | Regular | H-F | G-G | Box |
| tiempo | 2 | 8.96 | 0.0005 | 0.0006 | 0.0009 | 0.0063 |
| Residual | 48 | | | | | |

Se observa que se ha calculado un valor $p$ = 0,0005, que permitirá rechazar la hipótesis nula de que las mediciones durante los tres tiempos son iguales. Existen, por tanto, diferencias estadísticamente significativas entre las medias de las tres mediciones repetidas. La F de ***id,*** que vale 1,83 ($p$ = 0,0378), informa sobre la heterogeneidad entre sujetos. Este resultado es secundario y anodino (se interpretaría como el rechazo de la hipótesis nula de igualdad de medidas entre sujetos; es decir, no todos los sujetos de la muestra reaccionan igual, sino que existen diferencias significativas entre ellos).

Como se mencionó anteriormente, al tratarse de más de dos medidas reperidas, el valor $p$ calculado en la primera tabla tiene que ser corregido. De esta forma, en la tabla inferior de la salida de STATA se obtiene el cálculo del coeficiente $\varepsilon$ según tres métodos distintos. A continuación, se observa que los grados de libertad para tiempo son los mismos que en la tabla superior, el valor F también coincide y se muestran los valores $p$; en primer lugar, aparece el valor $p$ (0,0005) original, y los tres siguientes son los valores $p$ corregidos por los tres métodos descritos. Como se recomienda emplear el método de Greenhouse-Geisser, se escogerá el tercer valor de $p$ ($p$ = 0,0009).

Este tipo de análisis (ANOVA de medidas repetidas) era un método muy utilizado hace algunas décadas, pero ha ido cayendo en desuso, ya que las ecuaciones estimación generalizadas (GEE, por sus siglas en inglés, que se explicarán en el apartado 19.8) representan el método de elección casi siempre que se tienen diseños longitudinales de medidas repetidas. De hecho, no es necesario utilizar este método cuando se puede aplicar el GEE, dado que es más sencillo de ajustar y no tan limitado como el ANOVA de medidas repetidas (9). No obstante, es útil conocer la interpretación del ANOVA clásico de medidas repetidas, ya que algunos investigadores aún lo siguen utilizando en sus publicaciones.

## 9.4. EQUIVALENTE NO PARAMÉTRICO DEL ANOVA CON MEDIDAS REPETIDAS: TEST DE FRIEDMAN

Si el tamaño de muestra fuese pequeño y además no se pudiese asumir que los residuales se adaptan bien a la normalidad, no debería realizarse un ANOVA de medidas repetidas. En esta situación existe un test *no paramétrico* para el análisis de medidas repetidas: el test de Friedman. Debe reiterarse que el ANOVA es una técnica robusta y relativamente resistente a la falta de normalidad, especialmente para muestras grandes (**12,13**). Por otra parte, en medicina es frecuente encontrar asimetría positiva en variables cuantitativas. En esta situación suele mejorar la adaptación a la normal si se transforma logarítmicamente la variable dependiente. También debe tenerse en cuenta que muchas veces, ante la duda, será interesante repetir el análisis mediante los dos métodos para comprobar que los resultados coinciden. Cuando difieran mucho, la muestra sea pequeña y exista mala adaptación de los residuales a la normalidad, se preferirán los resultados del test de Friedman.

Imagínese que se desea comparar la puntuación obtenida en una escala de adhesión a la dieta mediterránea (mínimo 0 puntos, máximo 14 puntos) medida en 10 sujetos en tres tiempos diferentes tras una intervención nutricional dirigida, precisamente, a aumentar la adhesión a la dieta mediterránea entre los participantes (tabla 9.5).

Los pasos necesarios en un test de Friedman son los siguientes:

1. *Reemplazar los datos de cada sujeto por su rango* dentro de cada fila (su posición, ordenados de mayor a menor en el conjunto de cada sujeto). Estos rangos están indicados en la tabla 9.5 en cursiva y entre paréntesis. En caso de empate entre dos o más datos, el rango asignado es el promedio de los empatados. Por ejemplo, si los datos que ocupan las posiciones 2 y 3 tienen el mismo valor, el rango asignado a cada uno es la media de 2 y 3 = 2,5.
2. *Sumar los rangos por columnas.* Como comprobación de que las sumas se han realizado correctamente, se verificará que la suma total de rangos es (n.° de medidas(n.° de medidas + 1)/2) × (n.° de sujetos).

$$\sum R_i = 3 \times ((3+1)/2)) \times 10 = 60$$

3. *Calcular los rangos medios.* Se divide la suma de rangos *Ri* de cada columna entre el número de observaciones en cada una, es decir, se obtiene simplemente la media de los rangos.
   Rangos medios:
   Primera medida = 12/10 = 1,2.
   Segunda medida = 22/10 = 2,2.
   Tercera medida = 26/10 = 2,6.

***Tabla 9.5 Test de Friedman. Puntuación de adherencia a la dieta mediterránea (0-14) medida en 10 sujetos en tres tiempos diferentes tras una intervención nutricional***

| SUJETO | MED1 | MED2 | MED3 |
|---|---|---|---|
| 1 | 12 *(1)* | 13 *(2)* | 14 *(3)* |
| 2 | 9 *(1)* | 13 *(2,5)* | 13 *(2,5)* |
| 3 | 11 *(2)* | 10 *(1)* | 13 *(3)* |
| 4 | 10 *(2)* | 11 *(3)* | 9 *(1)* |
| 5 | 10 *(1)* | 13 *(2)* | 14 *(3)* |
| 6 | 10 *(1)* | 11 *(2,5)* | 11 *(2,5)* |
| 7 | 7 *(1)* | 10 *(2)* | 13 *(3)* |
| 8 | 8 *(1)* | 9 *(2)* | 13 *(3)* |
| 9 | 9 *(1)* | 12 *(3)* | 11 *(2)* |
| 10 | 9 *(1)* | 10 *(2)* | 11 *(3)* |
| **Suma de rangos** | **12** | **22** | **26** |

4. *Calcular una ji cuadrado* ($\chi^2$), según la siguiente expresión:

$$\chi^2 = \frac{\sum[R_i - \frac{n(k+1)}{2}]^2}{\frac{nk(k+1)}{12}} \rightarrow (\text{grados de libertad} = k-1)$$

donde:
$k$ es el número de observaciones repetidas (número de medidas).
$n$ es el número de sujetos.
$R_i$ es la suma de rangos para la medición i.

$$\chi^2 = \frac{(12-20)^2 + (22-20)^2 + (26-20)^2}{\frac{10\times 3\times 4}{12}} = \frac{104}{10} = 10,4$$

Puede entenderse que $20 = \left(\frac{n(k+1)}{2}\right)$ es el valor esperado de la suma de rangos si la hipótesis nula fuese cierta. La diferencia entre la suma de rangos $R_i$ realmente observada para cada medición y la $R_i$ esperada bajo la hipótesis nula es la estimación del efecto. Como se calcula una $\chi^2$, esta diferencia se eleva al cuadrado. El denominador $\frac{nk(k+1)}{12}$ correspondería al *seudoerror* estándar de esa diferencia (elevado también al cuadrado).

En caso de empates, el resultado de la $\chi^2$ debe ajustarse dividiendo entre el siguiente factor de corrección (FC):

$$FC = 1 - \frac{\sum_i T_i}{nk(k^2-1)}$$

donde: $T_i = \sum_h t_{ih}^3 - \sum_h t_{ih}$

$t_{ih}$ es el número de observaciones empatadas para un rango dado en el i-ésimo individuo. En este ejemplo había dos empates, correspondientes a los individuos 2 y 6, con dos empates ($t_{ih} = 2$) cada uno, por lo que $T_1 = 2^3 - 2 = 6$ y $T_2 = 2^3 - 2 = 6$.

$$FC = 1 - \frac{(6+6)}{10\times 3(3^2-1)} = 0,95$$

$$\chi_2^2 = \frac{10,4}{0,95} = 10,95$$

Si hubiese algún sujeto en el que se diese un triple empate en las tres medidas, entonces el valor de $T_i$ sería $T_i = 3^3 - 3 = 24$; esta situación no se ha producido en el ejemplo.

5. *Mirar en las tablas de $\chi^2$ para $k - 1$ grados de libertad.*

$\chi_2^2 = 10,95 \rightarrow$ significativo a $p < 0,005$ para dos grados de libertad.
Con STATA:

```
display chi2tail(2,10.95)
```

Con Excel:
=DISTR.CHI ($X^2$; $gl$)= DISTR.CHI(10,95;2) = 0,004

6. *Conclusión.* La adhesión a la dieta mediterránea varió significativamente a lo largo del tiempo tras una intervención nutricional. Observando las sumas de rangos obtenidas en cada uno de los tres momentos (12, 22 y 26), se puede afirmar que esa variación corresponde a una mayor adhesión a la dieta mediterránea entre la medición anterior al inicio de la intervención y las mediciones posteriores. Para saber si ese aumento ya es significativo desde los dos primeros momentos se deberían realizar comparaciones para datos apareados entre cada dos medidas (test de Wilcoxon). Habría que penalizar los valores *p* encontrados mediante el ajuste de Bonferroni, al tratarse de comparaciones múltiples.

## 9.5. AJUSTE DEL ANOVA POR VARIABLES CONTINUAS: ANCOVA

ANCOVA quiere decir *análisis de la covarianza* y corresponde a una técnica estadística que combina una regresión y un ANOVA. En este caso, el modelo incluye una variable dependiente (el desenlace) y una o más variables cuantitativas (llamadas covariables), que actúan como predictoras. Las covariables representan la fuente de la variabilidad que, según se piensa, influye en la variable dependiente, pero no se ha podido controlar durante el procedimiento experimental. De esta forma, el ANCOVA determina la correlación entre estas covariables y la variable dependiente, y posteriormente retira esa varianza asociada con las covariables a la variable dependiente antes de determinar si las diferencias en la variable dependiente son significativas. Así, se puede decir que el objetivo del ANCOVA es eliminar cualquier error sistemático fuera del control del investigador que pudiera llegar a sesgar los resultados, además de tener en cuenta las diferencias debidas a las características propias de los sujetos incluidos en el estudio.

Las condiciones de aplicación del ANCOVA incluyen las ya vistas en el capítulo anterior del ANOVA de una vía y las que se verán para la regresión:

1. Distribución normal de los residuales.
2. Homogeneidad de varianzas entre los grupos.
3. Independencia.
4. Linealidad: tiene que existir una relación lineal entre las covariables y la variable dependiente, así como entre cada par de covariables.
5. Ausencia de multicolinealidad, definida como la presencia de correlaciones fuertes ente las covariables.

A continuación, se ilustra el método ANCOVA con un ejemplo. Se desea comparar la media de índice de masa corporal (IMC) en 3 grupos de participantes (no fumadores, fumadores actuales o exfumadores); si se deseara únicamente realizar esta comparación, la forma más adecuada sería a través de un ANOVA de una vía, tal como se explicó en el capítulo anterior. Pero, si se pensara, además, que el IMC está fuertemente relacionado con la actividad física, se tendría que plantear un modelo con ajuste por actividad física. En el apartado 9.8.2, se explicará este ejemplo de ANCOVA y la manera de realizarlo con STATA.

## 9.6. COMPARACIONES INTRAGRUPO E INTERGRUPOS CON MEDIDAS REPETIDAS

En el apartado 9.3, se trató el ANOVA de medidas repetidas con un desenlace medido varias veces en el tiempo. En este apartado, la situación es muy similar y únicamente difiere en que el desenlace cambiante en el tiempo se compara ahora entre dos grupos. Este diseño se llama comparación *intragrupo (within)* e *intergrupos (between)*. Cabe mencionar que el grupo puede ser tanto dicotómico como con más categorías.

Las condiciones de aplicación de este test son las mismas que para el ANOVA de medidas repetidas, a las que se añade una más: las matrices de covarianza de los diferentes grupos de

***Tabla 9.6 Peso de seis pacientes medido en tres momentos (basal, a 1 mes y a los 3 meses) tras ser asignados a una dieta hipocalórica = 1 o normocalórica = 2***

| SUJETO | PESO1 | PESO2 | PESO3 | DIETA |
|---|---|---|---|---|
| 1 | 76 | 65 | 63 | 1 |
| 2 | 82 | 70 | 68 | 1 |
| 3 | 80 | 78 | 70 | 1 |
| 4 | 84 | 80 | 84 | 2 |
| 5 | 79 | 75 | 79 | 2 |
| 6 | 84 | 84 | 84 | 2 |

comparación deben ser homogéneas; esto sería análogo a la igualdad de varianzas en dos grupos. A pesar de ser una condición importante, la violación de esta condición no es un problema grave en una muestra grande.

Con este tipo de diseño se puede contestar a las siguientes preguntas:

1. ¿Existe un efecto durante el tiempo para toda la muestra?
2. ¿Se produce un efecto durante el tiempo en cada grupo?
3. ¿El cambio con el tiempo es diferente para cada grupo?

La última de estas preguntas es la más importante de responder.

La tabla 9.6 contiene datos ficticios para ejemplificar el planteamiento de este tipo de diseño. Cuenta con los datos de seis pacientes a los cuales se les asignó el seguimiento de una dieta hipocalórica = 1 o normocalórica = 2, y cuyo peso se registró antes de iniciar el estudio, al cabo de 1 mes y a los 3 meses (*peso1*, *peso2* y *peso3*).

Los pasos para realizar este análisis serían los siguientes:

Tal y como se explicó en el apartado 9.3, para realizar este test en STATA se tendrá que realizar un **reshape** a la base de datos, para pasar de un formato <<*wide*>> a un formato <<*long*>> mediante la siguiente orden:

```
reshape long peso, i(id) j(tiempo)
```

```
. reshape long peso, i(id) j(tiempo)
(note: j = 1 2 3)

Data                                    wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                             6   ->      18
Number of variables                        5   ->       4
j variable (3 values)                          ->   tiempo
xij variables:
                                pesol peso2 peso3  ->   peso
-----------------------------------------------------------------------------
```

Al ejecutar la instrucción **reshape,** la base de datos pasa a tener un formato largo. Ahora, en lugar de seis observaciones hay 18, porque cada sujeto tiene tres medidas. También se observa que se ha creado una nueva variable llamada tiempo.

Una vez que la base de datos está en formato largo, a través del comando ANOVA se podrá pedir a STATA lo siguiente:

```
anova peso dieta / id | dieta tiempo ///
  tiempo#dieta,repeat(tiempo)
```

Después de la orden **anova** debe ir la variable dependiente (el desenlace), que en el ejemplo es ***peso;*** después se debe escribir la variable de agrupación ***(dieta)*** y, en seguida, se tiene que indicar el término del error intergrupos, que sería el ***id,*** y el grupo ***(dieta);*** a continuación se incluiría la variable nueva que se creó durante el **reshape** ***(tiempo),*** seguida del término de interacción ***tiempo#dieta;*** finalmente, se indica la variable de repetición, que, en este caso, es el tiempo. Se obtendría así la salida que se recoge a continuación:

```
. anova peso dieta / id|dieta tiempo tiempo#dieta ,repeated(tiempo)

                   Number of obs =       18     R-squared     =  0.9534
                   Root MSE      =  2.18581     Adj R-squared =  0.9011

          Source |  Partial SS    df       MS          F     Prob > F
-----------------+------------------------------------------------------
           Model |  782.722222     9  86.9691358      18.20     0.0002
                 |
           dieta |       364.5     1       364.5       8.76     0.0416
         id|dieta |  166.444444     4  41.6111111
-----------------+------------------------------------------------------
          tiempo |  137.444444     2  68.7222222      14.38     0.0022
    tiempo#dieta |  114.333333     2  57.1666667      11.97     0.0039
                 |
        Residual |  38.2222222     8  4.77777778
-----------------+------------------------------------------------------
           Total |  820.944444    17  48.2908497

Between-subjects error term:  id|dieta
                     Levels:  6          (4 df)
         Lowest b.s.e. variable:  id
         Covariance pooled over:  dieta      (for repeated variable)

Repeated variable: tiempo
                                 Huynh-Feldt epsilon        =  0.7289
                                 Greenhouse-Geisser epsilon =  0.5250
                                 Box's conservative epsilon =  0.5000

                                            ------------ Prob > F ------------
          Source |  df       F      Regular     H-F     G-G     Box
-----------------+------------------------------------------------------
          tiempo |   2     14.38     0.0022    0.0071  0.0172  0.0192
    tiempo#dieta |   2     11.97     0.0039    0.0108  0.0235  0.0258
        Residual |   8
-----------------+------------------------------------------------------
```

Se puede apreciar que la suma de cuadrados totales corresponde a la suma de cuadrados de: factor 1 (dieta) + factor 2 (tiempo) + error intersujetos (id|dieta) + interacción (tiempo#dieta) + residual.

En la primera tabla se presenta la suma de cuadrados, los grados de libertad, el estadístico F y la significación estadística para cada una de estas fuentes de variabilidad. Como se mencionó al tratar del ANOVA de medidas repetidas, cuando se tienen más de dos medidas en el tiempo siempre deberá hacerse una corrección en el valor *p,* para tener en cuenta la *esfericidad.* Estos son los resultados que se presentan en la segunda tabla de la salida de STATA. Con estos datos se pueden responder a las preguntas que se planteaban al inicio de este apartado:

- *¿El cambio en el tiempo es diferente en cada grupo?* Se respondería mediante el valor *p* de la *interacción* ***tiempo#dieta.*** Como se ha dicho, la corrección más utilizada es la de Greenhouse-Geisser (G-G), por lo que la respuesta es afirmariva, con diferencias significativas basadas en un valor $p = 0{,}0235$.
- *¿Existe un efecto durante el tiempo?* Sí, se observa una *p* sin corrección de *0,0022* y una *p* corregida de *0,017.*

Estos resultados se pueden comprobar de otro modo, para lo cual podría obtenerse un modelo de regresión lineal. Esta operación es muy sencilla en STATA con la orden postestimación **regress**, que no necesita complementarse con ninguna otra información cuando se indica *justamente detrás* del ANOVA anterior:

```
regress
```

Esta sencilla orden obtendría el siguiente resultado, que es muy informativo:

```
. regress
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 782.722222 | 9 | 86.9691358 | Number of obs | = 18 |
| Residual | 38.2222222 | 8 | 4.77777778 | F( 9, 8) | = 18.20 |
| | | | | Prob > F | = 0.0002 |
| | | | | R-squared | = 0.9534 |
| | | | | Adj R-squared | = 0.9011 |
| Total | 820.944444 | 17 | 48.2908497 | Root MSE | = 2.1858 |

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 2.dieta | 10 | 2.304049 | 4.34 | 0.002 | 4.686853 | 15.31315 |
| id\|dieta | | | | | | |
| 1 2 | 0 | (empty) | | | | |
| 2 1 | 5.333333 | 1.784709 | 2.99 | 0.017 | 1.217788 | 9.448879 |
| 2 2 | 0 | (empty) | | | | |
| 3 1 | 8 | 1.784709 | 4.48 | 0.002 | 3.884454 | 12.11555 |
| 3 2 | 0 | (empty) | | | | |
| 4 1 | 0 | (empty) | | | | |
| 4 2 | -1.333333 | 1.784709 | -0.75 | 0.476 | -5.448879 | 2.782212 |
| 5 1 | 0 | (empty) | | | | |
| 5 2 | -6.333333 | 1.784709 | -3.55 | 0.008 | -10.44888 | -2.217788 |
| 6 1 | 0 | (empty) | | | | |
| 6 2 | 0 | (omitted) | | | | |
| tiempo | | | | | | |
| 2 | -8.333333 | 1.784709 | -4.67 | 0.002 | -12.44888 | -4.217788 |
| 3 | -12.33333 | 1.784709 | -6.91 | 0.000 | -16.44888 | -8.217788 |
| tiempo#dieta | | | | | | |
| 2 2 | 5.666667 | 2.523959 | 2.25 | 0.055 | -.1535938 | 11.48693 |
| 3 2 | 12.33333 | 2.523959 | 4.89 | 0.001 | 6.513073 | 18.15359 |
| _cons | 74.88889 | 1.629209 | 45.97 | 0.000 | 71.13193 | 78.64585 |

Aquí se comprueba que los resultados en valores *p* coinciden con el ANOVA. Quizá este nuevo listado de salida es más fácil de interpretar que el del ANOVA, ya que también sirve para estimar la magnitud de las diferencias. Se aprecia que la suma de cuadrados de la regresión (782,72) y la varianza de la regresión (86,97) corresponden a lo obtenido en el ANOVA para el modelo. Lo mismo ocurre para la fuente de variabilidad residual *(error)*, con una suma de cuadrados = 38,22 y una varianza = 4,78. En general, el modelo de regresión sirve para estimar la media de cada grupo y proporciona valores *p* para cada una de las comparaciones específicas. Otra ventaja derivada de resolver estos problemas por métodos de regresión en vez de por ANOVA es que permite calcular intervalos de confianza (se explicará en los apartados 10.5 y 12.10 al hablar de regresión lineal).

Si se quisiera realizar este mismo test, pero utilizando SPSS, se tendría que indicar la siguiente orden en la ventana de sintaxis:

```
GLM
peso1 peso2 peso3 BY dieta
/WSFACTOR = tiempo 3 Polynomial
/METHOD = SSTYPE(3)
/WSDESIGN = tiempo
      /DESIGN = dieta.
```

Se obtendrían, entonces, las siguientes tablas:

**Pruebas de efectos intrasujetos**

**Medida: MEASURE_1**

| Fuente | | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|---|
| Tiempo | Esfericidad asumida | 137,444 | 2,000 | 68,722 | 14,384 | 0,002 |
| | Greenhouse-Geisser | 137,444 | 1,050 | 130,894 | 14,384 | 0,017 |
| | Huynh-Feldt | 137,444 | 1,458 | 94,286 | 14,384 | 0,007 |
| | Límite-inferior | 137,444 | 1,000 | 137,444 | 14,384 | 0,019 |
| Tiempo * dieta | Esfericidad asumida | 114,333 | 2,000 | 57,167 | 11,965 | 0,004 |
| | Greenhouse-Geisser | 114,333 | 1,050 | 108,884 | 11,965 | 0,023 |
| | Huynh-Feldt | 114,333 | 1,458 | 78,432 | 11,965 | 0,011 |
| | Límite-inferior | 114,333 | 1,000 | 114,333 | 11,965 | 0,026 |
| Error (tiempo) | Esfericidad asumida | 38,222 | 8,000 | 4,778 | | |
| | Greenhouse-Geisser | 38,222 | 4,200 | 9,100 | | |
| | Huynh-Feldt | 38,222 | 5,831 | 6,555 | | |
| | Límite-inferior | 38,222 | 4,000 | 9,556 | | |

**Pruebas de los efectos intersujetos**

**Medida: MEASURE_1**

**Variable transformada: promedio**

| Fuente | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|
| Intersección | 106568,056 | 1 | 106568,056 | 2561,048 | 0,000 |
| Dieta | 364,500 | 1 | 364,500 | 8,760 | 0,042 |
| Error | 166,444 | 4 | 41,611 | | |

En ellas se observan los mismos resultados calculados antes con STATA. Con la primera parte de la tabla se contestaría a la pregunta de si el efecto es diferente en el tiempo; en los 4 valores *p* de significación estadística, el primero está sin corregir y los 3 siguientes están corregidos. Como se dijo,

es preferible elegir el método de Greenhouse-Geisser con $p = 0{,}017$. La siguiente $p$ correspondería a la de interacción ($p = 0{,}023$) que es la más importante, pues informa que el cambio en el tiempo fue significativamente diferente en los 2 grupos.

## 9.7. ANÁLISIS ESTADÍSTICO DE ENSAYOS *CROSS-OVER*

En este tipo de ensayos se desea comparar dos tratamientos diferentes (A y B) para una determinada patología. La particularidad del diseño *cross-over* radica en que el mismo sujeto recibe ambos tratamientos: en primer lugar, la mitad de la muestra (grupo 1) es sometida al tratamiento A y la otra mitad (grupo 2) recibe el tratamiento B; posteriormente, tras un período de lavado *(washout)*, el grupo 1 ahora recibe el tratamiento B y el grupo 2 es sometido al tratamiento A.

Para poder plantear un *cross-over* se requieren algunas condiciones básicas que dependen tanto del desenlace que se estudia como de la intervención:

1. El desenlace (problema clínico) ha de ser estable en el tiempo o muy similar en los períodos comparados.
2. La intervención (tratamiento) deberá ejercer un efecto rápido (suficiente para ser evaluado en el período establecido como tiempo de seguimiento en el protocolo del ensayo) y será de corta duración (su efecto no debe extenderse al siguiente período).

Las principales ventajas de este tipo de ensayos son dos: cada paciente es su propio control, de forma que todas las variables que pueden considerarse como confusoras se eliminan y se facilita el análisis comparativo, y el tamaño muestral que se requiere será menor.

El análisis estadístico, como se verá más adelante, es un tanto complejo. El objetivo principal sería la comparación de los efectos sobre los desenlaces finales. Sin embargo, al hacer esta comparación se debe descartar que el resultado obtenido esté condicionado por una variación en el tiempo del problema clínico en estudio o por factores relacionados con la secuencia en la que se administraron los dos tratamientos.

Los datos de la tabla 9.7 son ficticios y se utilizarán como ejemplo del análisis estadístico de un ensayo *cross-over*. El problema clínico que se intenta afrontar es la reducción de la tensión arterial sistólica (TAS) en pacientes hipertensos, con comparación de un fármaco A y un fármaco B.

En la tabla se observan varias columnas. Hay 6 pacientes (id); la columna f_1 se refiere al fármaco que recibieron en primer lugar; tas_basal es la tensión arterial sistólica basal al inicio del estudio y sin haber iniciado el fármaco; tas_fp1 es la tensión arterial sistólica al final del período 1; tas_lav es la tensión arterial sistólica en el período de lavado; f_2 es el fármaco que recibieron los pacientes en segundo lugar; tas_fp2 es la presión al final del período 2, y las columnas ta_a y ta_b son la presión al término del fármaco A y al final del fármaco B. Posteriormente, la columna dif_ab corresponde a la diferencia entre la presión al final de A-B, y la siguiente columna es la diferencia entre el período 1 y 2 (dif_p1_p2). La última columna es la de secuencia.

*Tabla 9.7 Ejemplo de estudio* **cross-over**

| id | f_1 | tas_basal | tas_fp1 | tas_lav | f_2 | tas_fp2 | ta_a | ta_b | dif_ab | dif_p1-p2 | sec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 162 | 147 | 160 | B | 152 | 147 | 152 | −5 | −5 | 1 |
| 2 | A | 170 | 147 | 169 | B | 158 | 147 | 158 | −11 | −11 | 1 |
| 3 | A | 177 | 136 | 178 | B | 150 | 136 | 150 | −14 | −14 | 1 |
| 4 | B | 160 | 126 | 162 | A | 127 | 127 | 126 | 1 | −1 | 2 |
| 5 | B | 189 | 166 | 185 | A | 158 | 158 | 166 | −8 | 8 | 2 |
| 6 | B | 177 | 151 | 175 | A | 122 | 122 | 151 | −29 | 29 | 2 |

Lo primero que se tiene que comprobar es si hubo un efecto diferente de los fármacos (A-B) al final del estudio sobre la tensión arterial sistólica (variable cuantitativa). Se debería asumir que la muestra era de mayor tamaño muestral y seguía una distribución normal. Así, el método estadístico indicado en este caso sería la *t* de Student para muestras relacionadas. Mediante este test se evaluará si la media de la diferencia es distinta de 0, y se obtiene el siguiente resultado:

```
. ttest ta_a=ta_b

Paired t test
```

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ta_a | 6 | 139.5 | 5.566267 | 13.63452 | 125.1915 | 153.8085 |
| ta_b | 6 | 150.5 | 5.475704 | 13.41268 | 136.4243 | 164.5757 |
| diff | 6 | -11 | 4.171331 | 10.21763 | -21.72275 | -.277253 |

```
    mean(diff) = mean(ta_a - ta_b)                                t =  -2.6370
Ho: mean(diff) = 0                             degrees of freedom =        5

Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
Pr(T < t) = 0.0231         Pr(|T| > |t|) = 0.0461          Pr(T > t) = 0.9769
```

Con este resultado se podrá rechazar la hipótesis nula y se concluirá que el tratamiento A reduce más la TAS que el tratamiento B ($p = 0{,}0461$).

Queda por evaluar si existieron efectos propios del diseño que pudieran haber influido en el resultado. Para esto sería preciso valorar si existió un efecto residual del primer fármaco administrado, simplemente comparando la media de la TAS basal y la media de la TAS en el período de lavado.

```
. ttest tas_basal=tas_lav

Paired t test
```

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tas_ba~l | 6 | 172.5 | 4.417767 | 10.82128 | 161.1438 | 183.8562 |
| tas_lav | 6 | 171.5 | 3.93912 | 9.648834 | 161.3742 | 181.6258 |
| diff | 6 | 1 | .8944272 | 2.19089 | -1.299198 | 3.299198 |

```
    mean(diff) = mean(tas_basal - tas_lav)                        t =   1.1180
Ho: mean(diff) = 0                             degrees of freedom =        5

Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
Pr(T < t) = 0.8428         Pr(|T| > |t|) = 0.3144          Pr(T > t) = 0.1572
```

Se observa que no existen diferencias significativas entre la TAS basal y la TAS en el período de lavado. Esto no proporciona evidencias de que la TAS cambiase respecto a sus niveles basales tras suspender el fármaco inicial. En cambio, si el resultado hubiese sido significativo, podría pensarse en un efecto residual *(carry over)*.

El siguiente paso consistiría en comprobar si existió un efecto del período. Para llevar a cabo este análisis basta con comparar la TAS al final del período 1 con la TAS al término del período 2.

```
. ttest tas_fp1=tas_fp2

Paired t test
```

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tas_fp1 | 6 | 145.5 | 5.554278 | 13.60515 | 131.2223 | 159.7777 |
| tas_fp2 | 6 | 144.5 | 6.489735 | 15.89654 | 127.8176 | 161.1824 |
| diff | 6 | 1 | 6.434283 | 15.76071 | -15.53985 | 17.53985 |

```
    mean(diff) = mean(tas_fp1 - tas_fp2)                      t =   0.1554
Ho: mean(diff) = 0                         degrees of freedom =        5

Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
Pr(T < t) = 0.5587         Pr(|T| > |t|) = 0.8826          Pr(T > t) = 0.4413
```

Se observa que la media de la TAS al final del período 1 fue de 145,5 (DE: 13,6) y la media al final del período 2, de 144,5 (DE: 15,9). No se observaron diferencias estadísticamente significativas. Si el resultado hubiese sido significativo, se habría concluido que las diferencias observadas eran atribuibles al período.

Para terminar el análisis, se debe comprobar si existió interacción entre el primer fármaco y la respuesta al segundo. Para ello, se analizará si la diferencia entre los dos fármacos A y B dependió de la secuencia con la que fueron administrados. Se aplicará un test de *t* de Student para muestras independientes comparando las diferencias A-B en los pacientes que recibieron la secuencia 1 (fármaco A-fármaco B) con aquellos a los que se administró la secuencia 2 (fármaco B-fármaco A).

```
. ttest dif_ab, by(sec)

Two-sample t test with equal variances
```

| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| 1 | 3 | -10 | 2.645751 | 4.582576 | -21.38375 | 1.383749 |
| 2 | 3 | -12 | 8.888194 | 15.3948 | -50.24281 | 26.24281 |
| combined | 6 | -11 | 4.171331 | 10.21763 | -21.72275 | -.277253 |
| diff | | 2 | 9.273618 | | -23.74769 | 27.74769 |

```
    diff = mean(1) - mean(2)                                  t =   0.2157
Ho: diff = 0                               degrees of freedom =        4

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.5801         Pr(|T| > |t|) = 0.8398          Pr(T > t) = 0.4199
```

Se obtiene una $p = 0{,}8398$ no significativa. Si este test hubiese sido significativo, indicaría que no es que el fármaco A siempre indujese mayor efecto que el B, sino que su efecto sería dependiente de la secuencia en que fue administrado.

Mediante estos tres últimos pasos no se obtuvo evidencia de efectos residuales ni interacción, por lo que se concluirá que el fármaco A es mejor que el fármaco B. Si existiera interacción, el tratamiento estadístico se haría más complejo, lo que desborda los objetivos de este capítulo.

## 9.8. ANOVA FACTORIAL Y ANCOVA: ESTIMACIÓN DE MEDIAS AJUSTADAS EN STATA

### 9.8.1. ANOVA factorial

La opción del ANOVA factorial de STATA está integrada en los modelos lineales y se ha diseñado pensando más en la regresión que en el ANOVA clásico. Por eso exige que se incluya explícitamente la interacción, ya que, por omisión, ajusta un ANOVA factorial sin término de interacción.

A partir del mismo ejemplo explicado en el apartado 9.2, se debe indicar la siguiente orden a STATA:

```
anova dif_peso dieta ejercicio dieta#ejercicio
```

La salida que proporciona el programa es:

```
. anova dif_peso dieta ejercicio dieta#ejercicio
```

| | | | | | |
|---|---|---|---|---|---|
| | Number of obs = | 20 | R-squared | = | 0.7681 |
| | Root MSE = | 3.05778 | Adj R-squared | = | 0.7247 |

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 495.6 | 3 | 165.2 | 17.67 | 0.0000 |
| dieta | 369.8 | 1 | 369.8 | 39.55 | 0.0000 |
| ejercicio | 105.8 | 1 | 105.8 | 11.32 | 0.0040 |
| dieta#ejercicio | 20 | 1 | 20 | 2.14 | 0.1630 |
| Residual | 149.6 | 16 | 9.35 | | |
| Total | 645.2 | 19 | 33.9578947 | | |

***Interpretación:*** *el programa presenta al principio el tamaño muestral total (N = 20), el coeficiente de determinación o $R^2$, que es el porcentaje de la variabilidad total en el cambio de peso explicada por la dieta, el ejercicio y su interacción (explican el 76,81% de la variabilidad total). También proporciona un valor de $R^2$ ajustado para cuando se introduce más de un factor. STATA denomina «Model» a la suma de los dos efectos principales más la interacción. Presenta la suma de cuadrados o «Sum of Squares» (Partial SS), los grados de libertad (degrees of freedom, df) y las varianzas o «Means squared» (MS) para cada fuente de variabilidad («Source»). Las fuentes de variabilidad son cada factor, su interacción y el residual. Finalmente, el programa presenta la suma de cuadrados total (645,2), sus grados de libertad totales (N – 1 = 20 – 1 = 19) y el cociente entre ambos, que es la varianza del cambio de peso (645,2/19 = 33,9578947).*

*Con el valor p = 0,163 encontrado, no se puede rechazar la hipótesis nula de que el efecto de un factor sea constante para los diversos niveles del otro. Como siempre que un valor p no resulta significativo, hay que proponer una interpretación correcta. No se habrá demostrado la ausencia de interacción; simplemente, con estos datos se carece de argumentos o pruebas para ponerla de manifiesto. De todos modos, la ausencia de interacción facilita la interpretación de estos resultados.*

*Los valores p = 0,004 y p < 0,001 llevan a rechazar la hipótesis nula de que el ejercicio no tiene efecto sobre el cambio de peso y la dieta carece de efecto sobre el cambio de peso, respectivamente. Es decir, se concluirá que tanto el ejercicio como la dieta tienen un efecto significativo sobre el cambio de peso.*

*De esta forma, en ausencia de interacción significativa, se puede asumir que, mientras no se demuestre lo contrario, el efecto de la dieta sería el mismo para los que hacen ejercicio y para los que no lo hacen. También se cumple lo contrario: el efecto del ejercicio sería el mismo para quienes siguen una dieta*

*y para quienes no lo hacen. Esta es la hipótesis nula de la interacción. En cambio, en presencia de interacción habría dos efectos diferentes de la dieta: uno para los que hacen ejercicio y otro para los que no lo hacen. Esto implica, necesariamente, que también habría dos efectos distintos del ejercicio, uno para los que siguen una dieta y otro para los que no la siguen.*

### 9.8.2. ANCOVA: estimación de medias ajustadas

Para realizar un ANCOVA con STATA se puede proceder de dos maneras: con **anova** o con **regress.** En STATA, las dos instrucciones asumen una respuesta continua (variable dependiente). Si se utiliza **regress,** los predictores son considerados variables continuas; en cambio, cuando se use **anova,** los predictores por defecto se consideran factores categóricos. En ambos casos, ya se use una orden u otra, se tendrá que especificar que la variable es continua en **anova,** o que es categórica en **regress.**

De esta forma, las órdenes que se deben indicar a STATA para llevar a cabo un ANCOVA son:

1. Utilizando **anova,** se desea obtener la media del IMC ajustado por tabaco (factor con tres niveles) y actividad física (variable continua):

La **c.** indica a STATA que es una variable continua

anova *imc tabaco c.mettotal*

```
. anova imc tabaco c.mettotal
```

```
                          Number of obs =       193     R-squared     =  0.1759
                          Root MSE      =   2.13798     Adj R-squared =  0.1629
```

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 184.450336 | 3 | 61.4834453 | 13.45 | 0.0000 |
| tabaco | 159.566924 | 2 | 79.7834618 | 17.45 | 0.0000 |
| mettotal | 22.2958736 | 1 | 22.2958736 | 4.88 | 0.0284 |
| Residual | 863.91518 | 189 | 4.57097979 | | |
| Total | 1048.36552 | 192 | 5.46023706 | | |

En este ejemplo, la variable dependiente es el índice de masa corporal ***(imc),*** la variable de agrupación es el ***tabaco,*** que tenía tres categorías, y la variable ***mettotal*** (actividad física) es la covariable cuantitativa. A esta variable se le tiene que anteponer **c.** para informar a STATA de que se trata de una variable continua. En este ejemplo se tendría que rechazar la hipótesis nula, ya que el valor *p* del modelo es *<0,001.* También se observa que la *p* del tabaco es significativa ($p < 0{,}001$), lo que se interpretaría diciendo que las medias de IMC son diferentes en las tres categorías del factor tabaco, con independencia de la actividad física. La *p* de ***mettotal*** (actividad física) también es significativa ($p = 0{,}028$), lo que se interpreta como que la media del IMC es diferente según la actividad física que se realice, independientemente del tabaco.

Si, además, se quisieran conocer las medias de IMC para cada una de las tres categorías del factor tabaco *ajustadas* por actividad física, lo que se tendría que pedir a STATA (inmediatamente después del ANOVA) sería lo siguiente:

**margins *tabaco***

```
. margins tabaco

Predictive margins                              Number of obs   =        193

Expression   : Linear prediction, predict()
```

| | Margin | Delta-method Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tabaco | | | | | | |
| 0 | 20.82304 | .2020635 | 103.05 | 0.000 | 20.427 | 21.21908 |
| 1 | 22.38754 | .2997337 | 74.69 | 0.000 | 21.80007 | 22.975 |
| 2 | 23.01993 | .3904511 | 58.96 | 0.000 | 22.25466 | 23.7852 |

Si *después* del ANOVA se pide **margins,** seguido del nombre de la variable de agrupación, STATA mostrará las medias ajustadas con los intervalos de confianza al 95% en las tres categorías de agrupación (factor tabaco). En este ejemplo, se obtendrán tres medias ajustadas de IMC para las tres categorías de la variable tabaco: 0 son los nunca fumadores (media de IMC = 20,82 kg/m$^2$), 1 son los fumadores actuales (22,39 kg/m$^2$) y 2 los exfumadores (23,02 kg/m$^2$). Se obtienen dos ventajas: estas medias están ajustadas por actividad física y, además, se obtienen sus intervalos de confianza. Este método se puede aprovechar para ajustar por múltiples factores. Lo más habitual es ajustar, al menos, por sexo y edad.

2. Utilizando **regress:**

Si se repitiese el mismo ejemplo, pero ahora con la orden **regress,** la programación en STATA sería:

La **i.** indica a STATA que es un factor (categorías)

regress *imc i.tabaco mettotal*

```
. regress imc i.tabaco mettotal
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 184.450336 | 3 | 61.4834453 |
| Residual | 863.91518 | 189 | 4.57097979 |
| Total | 1048.36552 | 192 | 5.46023706 |

| | |
|---|---|
| Number of obs | = 193 |
| F( 3, 189) | = 13.45 |
| Prob > F | = 0.0000 |
| R-squared | = 0.1759 |
| Adj R-squared | = 0.1629 |
| Root MSE | = 2.138 |

| imc | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| tabaco | | | | | | |
| 1 | 1.564497 | .3616511 | 4.33 | 0.000 | .8511059 | 2.277888 |
| 2 | 2.19689 | .4395503 | 5.00 | 0.000 | 1.329835 | 3.063945 |
| mettotal | .0030915 | .0013998 | 2.21 | 0.028 | .0003303 | .0058526 |
| _cons | 20.53755 | .2376187 | 86.43 | 0.000 | 20.06882 | 21.00627 |

Cabe destacar que, ahora, se escribió una **i.** antes de la variable categórica para que STATA la considere un factor con varias categorías. Se puede comprobar que los resultados del ANCOVA y de la regresión coinciden. Además, puede constatarse que los coeficientes de la regresión (1,564497 para el grupo 1 de tabaco y 2,19689 para el grupo 2) coinciden exactamente con la diferencia de las medias ajustadas que se obtuvieron anteriormente con la orden **margins,** ya que:

22,38754 – 20,82304 = **1,5645**
23,01993 – 20,82304 = **2,19689**

## 9.9. ANOVA FACTORIAL EN SPSS

Las opciones de SPSS en este tipo de modelos lineales generalizados son muy amplias. Muchos de los conceptos que se necesitan para comprender las opciones que aparecen programadas corresponden más bien a cuestiones sobre regresión, que se estudiarán en capítulos siguientes.

Siguiendo con el mismo ejemplo anterior, se debe proceder de acuerdo con el siguiente recorrido por los menús de SPSS:

**Analizar → Modelo lineal general → Univariante... → Dependiente = dif_peso → Factores fijos = ejercicio y dieta → Opciones... → Estadísticos descriptivos → Continuar → Aceptar**

O, si se hace mediante sintaxis:

```
UNIANOVA
mal> dif_pesoBY ejercicio dieta.
  /METHOD = SSTYPE(3)
   /INTERCEPT = INCLUDE
   /PRINT = DESCRIPTIVE
   /CRITERIA = ALPHA(.05)
   /DESIGN = ejercicio dieta ejercicio*dieta
```

El resultado es fácilmente interpretable con lo visto hasta ahora. Un primer listado de tipo descriptivo resulta muy útil para tener una especie de *mapa* de las diferencias entre las categorías de cada factor. A continuación aparece la tabla de análisis de la varianza, con algunas diferencias con la tabla que aparece en STATA. Por ejemplo, SPSS añade dos conceptos, de los que es mejor prescindir en este análisis, que son Intersección y Total, tachados en la salida de SPSS. Los resultados se interpretarían de la misma manera que en el apartado anterior.

**Estadísticos descriptivos**

**Variable dependiente: dif_peso**

| Ejercicio | Dieta | Media | Desv. típ. | N |
|---|---|---|---|---|
| Ejercicio | Dieta control | –1,80 | 3,033 | 5 |
| | Dieta hipocalórica | –12,40 | 3,050 | 5 |
| | Total | –7,10 | 6,280 | 10 |
| No ejercicio | Dieta control | 0,80 | 2,775 | 5 |
| | Dieta hipocalórica | –5,80 | 3,347 | 5 |
| | Total | –2,50 | 4,528 | 10 |
| Total | Dieta control | –0,50 | 3,064 | 10 |
| | Dieta hipocalórica | –9,10 | 4,606 | 10 |
| | Total | –4,80 | 5,827 | 20 |

| Pruebas de los efectos intersujetos | | | | | |
|---|---|---|---|---|---|
| Variable dependiente: dif_peso | | | | | |
| Fuente | Suma de cuadrados tipo III | gl | Media cuadrática | F | Significación |
| Modelo corregido | 495,600* | 3 | 165,200 | 17,668 | 0,000 |
| ~~Intersección~~ | ~~460,800~~ | ~~1~~ | ~~460,800~~ | ~~49,283~~ | ~~0,000~~ |
| Ejercicio | 105,800 | 1 | 105,800 | 11,316 | 0,004 |
| Dieta | 369,800 | 1 | 369,800 | 39,551 | 0,000 |
| Ejercicio + dieta | 20,000 | 1 | 20,000 | 2,139 | 0,163 |
| Error | 149,600 | 16 | 9,350 | | |
| ~~Total~~ | ~~1106,000~~ | ~~20~~ | | | |
| Total corregida | 645,200 | 19 | | | |

*R cuadrado = 0,768 (R cuadrado corregida = 0,725).

## 9.10. TEST DE FRIEDMAN CON SPSS Y CON STATA

Para aplicar el test de Friedman, tanto en SPSS como en STATA, se utilizará el ejemplo visto en el apartado 9.4, (puntuación de adhesión a la dieta mediterránea medida en tres tiempos diferentes en 10 sujetos).

### 9.10.1. Test de Friedman con SPSS

Para el test de Friedman, se procedería así:

- A través de sintaxis, se daría la siguiente orden:

```
NPAR TESTS

  /FRIEDMAN = Med_1 Med_2 Med_3

  /STAT DESCRIPTIVES QUARTILES.
```

- A través del menú: **Analizar → Pruebas no paramétricas → K muestras relacionadas → seleccionar todas las variables de medidas repetidas que queremos comparar y pasarlas al recuadro Contrastar variables → en el recuadro Tipo de prueba debe estar señalado Friedman → Aceptar**

En la salida de SPSS se obtendrían dos recuadros. En el primero aparecerán los rangos promedios de cada medición en los diferentes tiempos; el segundo recuadro muestra los estadísticos de contraste, con el número de observaciones, el valor de $\chi^2$, los grados de libertad y la significación estadística.

### 9.10.2. Test de Friedman con STATA

El test de Friedman con STATA es complicado, ya que este *software* no incluye este test por defecto. Por ello, es necesario instalarlo con antelación. También se requiere transponer los datos. De esta forma, las órdenes que es preciso dar para realizar un test de Friedman son las siguientes:

1. Buscar el paquete con la orden **findit.**

```
findit friedman
```

2. En la ventana que se abra habrá que encontrar el paquete con el nombre «**package snp2_1** from http://www.stata-journal.com/software/sj5-2».
3. Instalar el paquete.
4. Si los datos se han introducido como una columna para cada variable (medida repetida v1, v2, v3) y una fila por sujeto, entonces se deben transponer los datos mediante la siguiente orden, que convertirá a cada sujeto en una columna y les llamará v1, v2, v3, …, v25 (se asume que había 25 sujetos):

```
xpose, clear
```

5. Realizar el test de Friedman (se ha asumido que había 25 sujetos):

```
friedman v1 - v25
```

## 9.11. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Test | STATA | SPSS |
|---|---|---|
| ANOVA factorial | **anova *dif_peso dieta*** ///<br>***ejercicio dieta#ejercicio*** | **UNIANOVA dif_pesoBY ejercicio dieta /METHOD = SSTYPE(3) /INTERCEPT = INCLUDE /PRINT = DESCRIPTIVE/CRITERIA = ALPHA(.05) /DESIGN = ejercicio dieta ejercicio*dieta.** |
| ANOVA de medidas repetidas | **reshape long *med*,** ///<br>**i(*id*) j(*tiempo*)**<br>**anova *med id*** ///<br>***tiempo*, repeat(*tiempo*)** | |
| Test de Friedman* | **xpose, clear**<br>**friedman *v1-v25*** | **NPAR TESTS /FRIEDMAN = Med_1 Med_2 Med_3 /STATISTICS DESCRIPTIVES QUARTILES /MISSING LISTWISE.** |
| ANCOVA | **anova *imc tabaco c.mettotal***<br>**margins *tabaco***<br>**regress *imc i.tabaco mettotal*** | **UNIANOVA imc BY tabaco WITH mettotal /EMMEANS = TABLES(tabaco).** |
| Comparaciones intragrupo e intergrupos con medidas repetidas | **reshape long *oliva*,** ///<br>**i(*id*) j(*tiempo*)**<br>**anova *peso dieta* / *id*\|*dieta***<br>***///tiempo tiempo#dieta*,**<br>**/// repeat(*tiempo*)** | **GLM peso1 peso2 peso3 BY dieta /WSFACTOR = tiempo 3 Polynomial /METHOD = SSTYPE(3) /WSDESIGN = tiempo /DESIGN = dieta.** |

*No está por defecto en STATA, necesita ser descargado de internet (**findit**).

## REFERENCIAS


1. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ 1996;313:808.
2. Matthews JN, Altman DG. Interaction 3: How to examine heterogeneity. BMJ 1996;313:862.
3. Altman DG, Matthews JN. Statistics notes. Interaction 1: Heterogeneity of effects. BMJ 1996;313:486.
4. Lagakos SW. The challenge of subgroup analyses – reporting without distorting. N Engl J Med 2006;354:1667-9.
5. Schulz KF, Grimes DA. Multiplicity in randomized trials II: subgroup and interim analyses. Lancet 2005;365:1657-61.
6. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine –reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357(21):2189-94.
7. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomized controlled trials: systematic review. BMJ 2011;342:d1569.
8. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. BMJ 2003;326:219.
9. Twisk JWR. Applied longitudinal data analysis for epidemiology: a practical guide. 2nd ed. Cambridge: Cambridge University Press; 2013.
10. Zazpe I, Sánchez-Tainta A, Estruch R, Lamuela-Raventós RM, Schröder H, Salas-Salvadó J, et al. A large randomized individual and Group intervention conducted by registered dietitians increased adherence to Mediterranean type diets: The PREDIMED study. J Am Diet Assoc 2008;108:1134-44.
11. Martínez-González MA, Corella D, Salas-Salvadó J, Ros E, Covas MI, Fiol M, et al. Cohort profile: design and methods of the PREDIMED study. Int J Epidemiol 2012;41(2):377-85.
12. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-69.
13. Altman DG. Practical statistics for medical research. Londres: Chapman and Hall; 1991.

# CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE 10

*A. Sánchez-Villegas, N. Martín-Calvo, M. Á. Martínez-González*

## 10.1. INTRODUCCIÓN

Cuando se trata de asociar estadísticamente dos variables cuantitativas, puede transformarse una de las variables en policotómica o en ordinal (categorizar) mediante la subdivisión en intervalos con puntos de corte fijados *a priori* o creando grupos iguales (cuantiles), y aplicar después el análisis de la varianza. Otra posibilidad consiste en aplicar técnicas de *correlación o regresión*. Aunque correlación y regresión puedan parecer métodos similares y sus cálculos matemáticos sean parecidos, son dos procedimientos distintos, tanto conceptualmente como en sus aplicaciones prácticas.

## 10.2. CORRELACIÓN

La finalidad de la correlación es examinar la dirección y la magnitud de la asociación entre dos variables cuantitativas[1]. Así se conocerá la «intensidad» de la relación entre ellas (medir o cuantificar el grado de asociación que mantienen) y se sabrá si, al aumentar el valor de una variable, aumenta o disminuye el valor de la otra.

Cabe destacar que las variables utilizadas en la correlación son simétricas: en este caso *no* existirá una variable dependiente y otra independiente, sino que son mutuamente intercambiables. En este contexto se usan frecuentemente dos coeficientes de correlación: el de Pearson y el de Spearman.

### 10.2.1. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson *(r)* es el más utilizado, hasta el punto de que a veces se conoce simplemente con el nombre de coeficiente de correlación, sin más apellido. Se trata de un índice que mide lo bien que se ajustan los puntos a una línea recta ideal. Es un método estadístico paramétrico, ya que utiliza la media, la varianza, etc., y, por tanto, requiere criterios de normalidad para las variables analizadas. Puede tomar valores entre −1 y +1. Cuando los puntos forman una línea perfecta creciente (de izquierda a derecha), vale +1, y si forman una línea perfectamente recta, pero decreciente, su valor será −1. Este valor aumentará conforme se incremente la concentración de los puntos alrededor de la línea recta que mejor se ajuste a la información contenida en los puntos.

El valor de $r$ será grande cuando los puntos estén muy concentrados en torno a la recta, y pequeño cuando los puntos en el gráfico estén muy dispersos con respecto a la recta imaginaria que define la relación. Este coeficiente $r$ es una medida abstracta que no posee unidades, es adimensional. Cumple las siguientes propiedades:

- $r = 0$: no existe correlación.
- $r > 0$ (positivo): al aumentar una de las variables, también aumenta la otra (se dice que la asociación es directa o positiva).

1 A veces se puede concebir la correlación como la «fuerza de la asociación» entre dos variables. Esto es básicamente adecuado, ya que la escala en la que se mueve el coeficiente de correlación (desde −1 a +1) es una cuantificación del grado en que se asocian las dos variables, independientemente de cuáles sean sus unidades de medida. El coeficiente $r$ de Pearson mide, además, la proximidad de los puntos a una línea recta. Sin embargo, el coeficiente de correlación de Spearman (no paramétrico) no mide asociación lineal, sino asociación en general. No obstante, para preservar un uso tradicional es preferible reservar el concepto de «fuerza de la asociación» para las medidas de asociación y efecto clásicas en epidemiología, como son el riesgo relativo, la *odds ratio* o la *hazard ratio*.

- $r < 0$ (negativo): al aumentar una de las variables, disminuye la otra (se dice que la asociación es indirecta, inversa o negativa).
- $|r| < 0{,}30$: asociación débil.
- $0{,}30 < |r| < 0{,}70$: asociación moderada.
- $|r| > 0{,}70$: asociación fuerte.

El coeficiente $r$ será próximo a 1 (en valor absoluto) cuando las dos variables X e Y estén estrechamente relacionadas, es decir, cuando varíen casi enteramente al unísono (al aumentar lo hace la otra, y viceversa). Este concepto de variación al unísono se denomina *covarianza*. Se entiende con facilidad si se considera que, al relacionar una variable consigo misma, se obtiene el grado máximo de asociación (aunque esta idea no tenga aplicación práctica). En tal caso existe una correlación perfecta ($r = +1$), como, por ejemplo, entre el peso medido en libras y el peso medido en kilogramos (en realidad, es la misma variable expresada en dos unidades distintas).

Habrá correlación entre variables que miden lo mismo desde distinta óptica. Por ejemplo, los valores de una transaminasa (ALT) estarán correlacionados con los de la otra transaminasa (AST), pues las dos están midiendo la función hepática. También existirá correlación entre las horas que se dedican a estudiar bioestadística y la nota del examen. Imagine unos datos muy sencillos para este último ejemplo (tabla 10.1).

La *covarianza* de *xy* (SPxy) se calcula multiplicando para cada sujeto las diferencias de cada valor de X con respecto a su media por las diferencias de cada valor de Y con respecto a su media. A continuación se suman todas las cantidades y se divide por $n - 1$, siendo $n$ el número de individuos. Matemáticamente, se expresaría como:

$$\text{cov} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\text{SPxy}}{n-1}$$

$$\text{cov} = \frac{(30-80)(3-7)+(90-80)(6-7)+(100-80)(9-7)+(100-80)(10-7)}{4-1}$$

$$= \frac{200-10+40+60}{3} = \frac{290}{3} = 96{,}7$$

El numerador de la covarianza (290) se llama *suma de los productos xy* (SPxy).

La ecuación para calcular el coeficiente de correlación de Pearson *(r)* es el cociente entre SPxy y la raíz del producto de las sumas de cuadrados de ambas variables.

$$r = \frac{\text{SPxy}}{\sqrt{(\text{SCx}) \times (\text{SCy})}}$$

***Tabla 10.1 Horas dedicadas por cuatro alumnos al estudio y nota que obtuvieron***

| HORAS DEDICADAS | NOTA OBTENIDA |
|---|---|
| 30 | 3 |
| 90 | 6 |
| 100 | 9 |
| 100 | 10 |
| Media = 80 | Media = 7 |
| $s = 3{,}16$ | $s = 33{,}67$ |

Recuérdese que la suma de cuadrados de X es $\sum(x_i - \bar{x})^2$. Lo mismo puede decirse para la suma de cuadrados de Y:

$$r = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}}$$

Las sumas de cuadrados pueden obtenerse multiplicando la varianza por $n - 1$. Así, en el ejemplo de las notas: SCx = $3(3{,}16^2) = 30$ y SCy = $3(36{,}67^2) = 3.400$. En consecuencia:

$$r = \frac{290}{\sqrt{30 \times 3.400}} = +0{,}908$$

En este ejemplo, el coeficiente de correlación $r$ de Pearson valdrá +0,908. Cuando en un punto uno de los dos valores coincida exactamente con su media, el sumando para la covarianza de esa observación sería igual a 0. Cuando una observación esté compuesta por un valor de X que sea inferior a su media y el valor correspondiente de Y sea superior a la media de Y, el resultado será negativo. Si ambos puntos se alejan mucho de la media, el resultado de la multiplicación tendrá una gran magnitud y esa observación será muy influyente, porque aportará mucho al numerador de $r$. Se entiende que cuanto mayor sea el grado de variación al unísono, mayor valor absoluto tendrá la covarianza. Si la variación al unísono se produce en sentido inverso (aumenta Y cuando disminuye X), la covarianza tendrá signo negativo. Si no hay variación al unísono, la covarianza valdrá 0 (figs. 10.1 y 10.2).

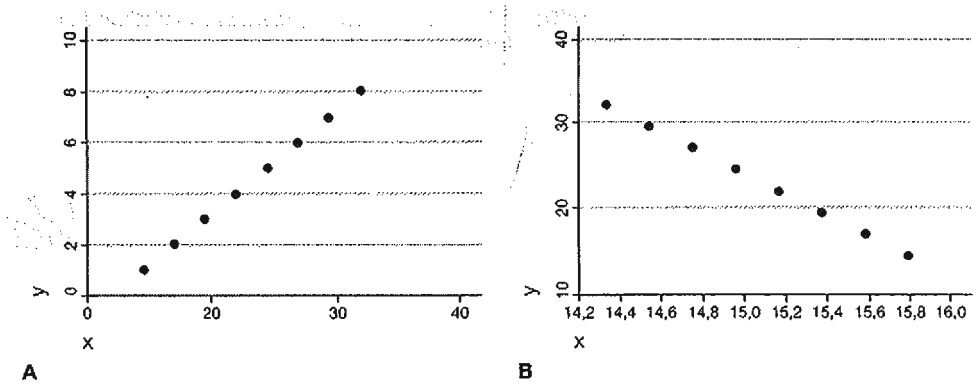



**Figura 10.1** Valores de la covarianza dependiendo de la variación al unísono de X e Y. **A.** Covarianza positiva > 0. **B.** Covarianza negativa < 0.

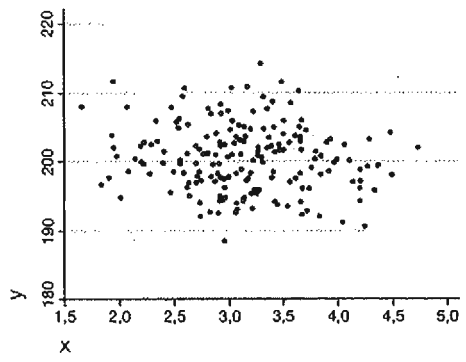



**Figura 10.2** Cuando no hay variación al unísono de X e Y, la covarianza vale 0.

En el *denominador* del coeficiente *r* de Pearson aparecen las sumas de cuadrados, que serán mayores al aumentar la dispersión de los valores de X y de Y. Por lo tanto, el coeficiente *r* de correlación será menor cuanta mayor dispersión exista. En el fondo, *r* es el cociente ya muchas veces visto entre efecto y error:

- La SPxy (numerador) sería el «efecto», que se coloca en el numerador.
- La raíz del producto de las sumas de cuadrados (denominador) corresponde al «ruido» o error de la variabilidad aleatoria.

$$r = \frac{SPxy}{\sqrt{(SCx) \times (SCy)}}$$

La expresión anterior es algebraicamente equivalente a otra expresión (**1**), que ofrece la ventaja de ser más fácil de recordar:

$$r = \frac{\sum (z_x \times z_y)}{n-1}$$

Como podría haberse supuesto, $z_x$ y $z_y$ son los valores tipificados de X e Y, mientras que $n$ es el número de observaciones o puntos (tamaño muestral). Es decir, cada dato se transforma en el número de desviaciones estándar que se aleja de la media.

Considérese el ejemplo de la tabla 10.2, que valora la relación entre el porcentaje de usuarios de cualquier producto de tabaco (incluidos cigarrillos, puros o pipa) y la incidencia de cáncer de pulmón en varones europeos[2]. En cada país existirá un par de valores (% de fumadores e incidencia estandarizada por edad de cáncer de pulmón). Se usará la notación X para el porcentaje de fumadores (variable *Fumad*) e Y para la incidencia de este tipo de cáncer (variable *Inccancer*).

Para valorar la asociación entre las dos variables, como primera aproximación suele prepararse un diagrama de dispersión (fig. 10.3). Al observar los datos de la tabla o la nube de puntos (v. fig. 10.3),

***Tabla 10.2 Comparación entre el porcentaje de usuarios de productos de tabaco y la incidencia de cáncer de pulmón estandarizada por edad en varones de diferentes países europeos***

| PAÍS | % DE USUARIOS VARONES *(FUMAD)* (X) | INCIDENCIA DE CÁNCER DE PULMÓN EN VARONES (CASOS/10.000 HAB.) *(INCCANCER)* (Y) |
|---|---|---|
| Alemania | 37 | 42,4 |
| Austria | 47 | 36,9 |
| Bélgica | 33 | 57,1 |
| Bulgaria | 49 | 53,7 |
| Croacia | 39 | 60,0 |
| Dinamarca | 35 | 43,3 |
| España | 37 | 53,3 |
| Finlandia | 33 | 31,2 |
| Francia | 36 | 47,7 |
| Grecia | 63 | 52,2 |
| Holanda | 33 | 47,4 |
| Hungría | 45 | 80,9 |
| Islandia | 29 | 31,6 |
| Irlanda | 34 | 37,9 |
| Italia | 34 | 45,4 |
| Polonia | 30 | 71,2 |
| Reino Unido | 26 | 38,2 |
| Rusia | 70 | 55,2 |
| Suecia | 17 | 18,2 |
| Suiza | 32 | 38,4 |

2 Los datos de consumo de tabaco corresponden a valores de 2006 obtenidos de la Organización Mundial de la Salud (OMS) (http://www.who.int/en/), y los datos de incidencia de 2008, a valores obtenidos de la Agencia Internacional del Cáncer (IARC) (http://globocan.iarc.fr/).

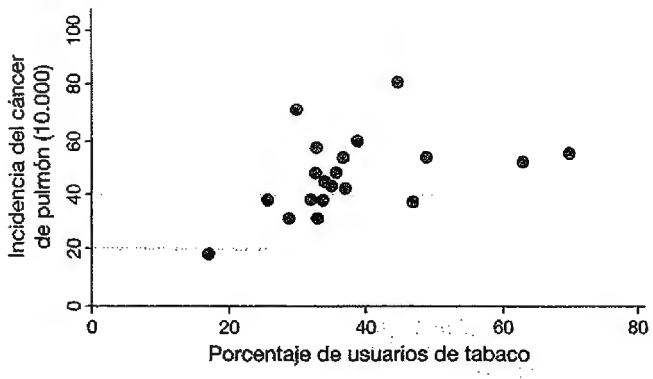



**Figura 10.3** Diagrama de dispersión que representa la relación entre el porcentaje de fumadores y la incidencia de cáncer de pulmón en varones europeos.

se aprecia que existe una tendencia a que la incidencia de cáncer de pulmón aumente a medida que se incrementa el porcentaje de fumadores varones en ese país. ¿Puede concretarse más esta relación? ¿Es posible caracterizarla cuantitativamente?

Puede calcularse ahora el coeficiente de correlación usando los valores *z*:

$$z = (\text{dato} - \text{media})/\text{desv. estándar}$$

En la tabla 10.3 se resumen los pasos necesarios para calcular el coeficiente de correlación *r* de Pearson siguiendo el procedimiento basado en valores *z*.

***Tabla 10.3 Pasos para calcular el coeficiente r de Pearson***

| | FUMAD (X) | INCCANCER (Y) | $z_x = \frac{x_i - 37,95}{12,1156}$ | $z_y = \frac{y_i - 47,11}{14,2810}$ | $z_x z_y$ |
|---|---|---|---|---|---|
| Alemania | 37 | 42,4 | −0,0784 | −0,3298 | 0,0259 |
| Austria | 47 | 36,9 | 0,7470 | −0,7149 | −0,5340 |
| Bélgica | 33 | 57,1 | −0,4086 | 0,6995 | −0,2858 |
| Bulgaria | 49 | 53,7 | 0,9121 | 0,4615 | 0,4209 |
| Croacia | 39 | 60 | 0,0867 | 0,9026 | 0,0782 |
| Dinamarca | 35 | 43,3 | −0,2435 | −0,2668 | 0,0650 |
| España | 37 | 53,3 | −0,0784 | 0,4334 | −0,0340 |
| Finlandia | 33 | 31,2 | −0,4086 | −1,1141 | 0,4552 |
| Francia | 36 | 47,7 | −0,1610 | 0,0413 | −0,0066 |
| Grecia | 63 | 52,2 | 2,0676 | 0,3564 | 0,7369 |
| Holanda | 33 | 47,4 | −0,4086 | 0,0203 | −0,0083 |
| Hungría | 45 | 80,9 | 0,5819 | 2,3661 | 1,3768 |
| Islandia | 29 | 31,6 | −0,7387 | −1,0861 | 0,8023 |
| Irlanda | 34 | 37,9 | −0,3260 | −0,6449 | 0,2103 |
| Italia | 34 | 45,4 | −0,3260 | −0,1197 | 0,0390 |
| Polonia | 30 | 71,2 | −0,6562 | 1,6869 | −1,1069 |
| Reino Unido | 26 | 38,2 | −0,9863 | −0,6239 | 0,6154 |
| Rusia | 70 | 55,2 | 2,6454 | 0,5665 | 1,4986 |
| Suecia | 17 | 18,2 | −1,7292 | −2,0244 | 3,5005 |
| Suiza | 32 | 38,4 | −0,4911 | −0,6099 | 0,2995 |
| Medias | 37,95 | 47,11 | $n = 20$ | SUMA = $(\Sigma z_x z_y)$= | 8,1487 |
| $s$ | 12,1156 | 14,281 | | | |

$$r = \frac{\sum(z_x \times z_y)}{n-1} = \frac{8{,}1487}{19} = +0{,}4289$$

***Interpretación:*** *el coeficiente r hallado es +0,43. Puede interpretarse desde dos puntos de vista, el de su signo y el de su magnitud:*

- *Como tiene signo positivo, se dice que, entre varones europeos, a medida que aumenta el porcentaje de fumadores en un país, crece también la incidencia de cáncer de pulmón.*
- *Como su magnitud absoluta es 0,43 (y el mínimo posible es 0 y el máximo 1), se dirá que la intensidad de la asociación entre ambas variables es el 43% de la máxima posible.*

### 10.2.2. Condiciones de aplicación de la correlación

Cabe señalar la existencia de varias condiciones, expresadas en virtud de los siguientes conceptos:

- *Variables cuantitativas.* Ambas variables examinadas han de ser cuantitativas. Para las variables ordinales, se puede usar el coeficiente de Spearman.
- *Normalidad.* La normalidad de ambas variables es un requisito en el coeficiente de correlación de Pearson, pero no en el de Spearman.
- *Independencia.* Las observaciones han de ser independientes, es decir, solo hay una observación de cada variable para cada individuo. No tendría sentido, por tanto, aplicar la correlación en un estudio que relacionase la ingesta diaria de sal y la tensión intraocular si se tomaran mediciones en los dos ojos de cada individuo. En este caso hay dos observaciones por paciente que están autocorrelacionadas, no son independientes; habría que considerar *n* como el número de pacientes, y no el de ojos, y usar métodos especiales para este tipo de diseños y otros que tienen *autocorrelación.* Se trata de casos en que la base de datos suele tener varias filas por paciente (v. apartado 19.8).

### 10.2.3. Coeficiente de correlación no paramétrico de Spearman[3] *(rho)*

El coeficiente de correlación de Spearman es un estimador no paramétrico que se utiliza en aquellos casos en los que las variables examinadas no cumplen necesariamente criterios de normalidad, o bien cuando las variables son ordinales.

Como sucede con otros métodos no paramétricos, el coeficiente de Spearman se basa en la sustitución de los valores originales de las dos variables por sus números de orden o rangos. La forma más sencilla de calcular este coeficiente consiste en usar los rangos en vez de los datos originales de las variables y calcular con los rangos un coeficiente de Pearson (tabla 10.4).

$$rho = \frac{315{,}5}{661{,}4} = +0{,}48$$

***Interpretación:*** *la asociación encontrada en el ejemplo presentado es rho = +0,48. Se aproxima bastante a la que se halló por el método paramétrico (r = +0,43). La conclusión es que, al aumentar el porcentaje de usuarios de tabaco, se incrementa también la incidencia de cáncer de pulmón en varones. Esta relación es moderada o intermedia, pues está en torno al 45% de la máxima posible.*

A diferencia del coeficiente de correlación de Pearson, este método *no* estima específicamente una asociación *lineal* entre las variables, sino solo una asociación en general. Por tanto, y en vista de que no todas las relaciones que se encuentran son lineales, debería usarse más (2). Otra ventaja es que no requiere supuestos previos sobre la distribución de los datos.

3 Se ha evitado el uso de la letra griega *ρ (rho),* que, en algunos textos, se usa para referirse a este coeficiente. Se considera más correcto reservar las letras griegas para parámetros poblacionales. Por eso usaremos simplemente la expresión *rho,* y no *ρ,* para referirnos al coeficiente de correlación de Spearman.

*Tabla 10.4 Pasos para calcular el coeficiente* **rho** *de Spearman*

| | FUMAD (X) | RANGO X | INC. CÁNCER PULMÓN (Y) | RANGO Y | $(rx_i-\overline{rx})\ (ry_i-\overline{ry})$ |
|---|---|---|---|---|---|
| Alemania | 37 | 13,5 | 42,4 | 8 | (13,5-10,5)(8-10,5) = –7,5 |
| Austria | 47 | 17 | 36,9 | 4 | Etc. = –42,25 |
| Bélgica | 33 | 7 | 57,1 | 17 | –22,75 |
| Bulgaria | 49 | 18 | 53,7 | 15 | 33,75 |
| Croacia | 39 | 15 | 60 | 18 | 33,75 |
| Dinamarca | 35 | 11 | 43,3 | 9 | –0,75 |
| España | 37 | 13,5 | 53,3 | 14 | 10,5 |
| Finlandia | 33 | 7 | 31,2 | 2 | 29,75 |
| Francia | 36 | 12 | 47,7 | 12 | 2,25 |
| Grecia | 63 | 19 | 52,2 | 13 | 21,25 |
| Holanda | 33 | 7 | 47,4 | 11 | –1,75 |
| Hungría | 45 | 16 | 80,9 | 20 | 52,25 |
| Islandia | 29 | 3 | 31,6 | 3 | 56,25 |
| Irlanda | 34 | 9,5 | 37,9 | 5 | 5,5 |
| Italia | 34 | 9,5 | 45,4 | 10 | 0,5 |
| Polonia | 30 | 4 | 71,2 | 19 | –55,25 |
| Reino Unido | 26 | 2 | 38,2 | 6 | 38,25 |
| Rusia | 70 | 20 | 55,2 | 16 | 52,25 |
| Suecia | 17 | 1 | 18,2 | 1 | 90,25 |
| Suiza | 32 | 5 | 38,4 | 7 | 19,25 |
| | | $\overline{rx}=10,5$ | | $\overline{ry}=10,5$ | SPxy = 315,5 |
| | | $s_{rx}=5,90$ | | $s_{ry}=5,92$ | $\sqrt{SCx \times SCy}=\sqrt{(19\times5,9^2)(19\times5,9^2)}$ * |
| | | | | | = 661,39 |

*Recuérdese que la suma de cuadrados puede obtenerse multiplicando el valor de la varianza de la variable por n – 1. $\overline{rx}$: rango medio de X; $\overline{ry}$: rango medio de Y; $s_{rx}$: desviación estándar de los rangos de X; $s_{ry}$: desviación estándar de los rangos de Y.

Existen otras fórmulas para el cálculo del coeficiente de Spearman, pero requieren corrección en caso de empates. En cambio, siempre se puede calcular un coeficiente de Spearman, como se ha hecho, siguiendo los pasos para calcular un coeficiente de Pearson, pero transformando antes los valores originales en sus rangos. El resultado ya está corregido para empates. En resumen, el coeficiente de Spearman ofrece dos ventajas importantes:

- Es un método no paramétrico y, a diferencia de los otros métodos no paramétricos que se han estudiado ya, *sí* permite construir intervalos de confianza, como se verá más adelante. El método para obtener el intervalo de confianza es el mismo que para la *r* de Pearson.
- Estima el grado de asociación de cualquier tipo, sea cual sea la función que relacione a las dos variables, sin exigir que sea lineal.

Existe otro coeficiente de correlación no paramétrico menos usado que el de Spearman, que se llama $\tau$ *(tau)* de Kendall. Está especialmente indicado con variables ordinales, pero siempre se puede usar también el de Spearman. Su interpretación es similar, aunque su cálculo es más complejo.

### 10.2.4. Errores de interpretación de la correlación

La correlación se aplica con el objetivo de medir el grado de asociación entre dos variables cuantitativas. Las variables en la correlación son *simétricas*, es decir, mutuamente intercambiables. En ningún momento se habla de que una de ellas podría ejercer el papel de «causa» (variable independiente) y la otra el de «efecto» (variable dependiente). Por tanto, no es relevante el eje que

ocupa cada variable. En el coeficiente de correlación no influyen las unidades de medida: siempre se mueve en el rango comprendido entre −1 y +1, independientemente de las unidades que se utilicen para medir las variables.

A diferencia del ejemplo presentado sobre tabaco y cáncer, ampliamente estudiado, es posible encontrar asociaciones fortuitas (debidas únicamente al azar). Por otra parte, puede presentarse un «factor no considerado» o «causa común» que aclare los hallazgos. Se trataría de una tercera variable en juego a la que se llama factor de confusión (**3-6**). Por ejemplo, tal vez resulte paradójico encontrar una correlación *inversa* entre la ingesta dietética total (calorías consumidas) y el peso corporal. ¿Cómo puede ser que los individuos que consumen más calorías estén más delgados? La explicación proviene de considerar una tercera variable en juego: el nivel de ejercicio físico practicado en el tiempo libre. Quienes más calorías consumen son los que más ejercicio físico realizan, y el ejercicio físico realizado en el tiempo libre es uno de los factores que más ayudan a mantener el peso ideal (7). Por eso, no basta que un coeficiente de correlación sea de gran magnitud para considerar que la asociación entre dos variables es causal: hay que mantener siempre una cierta prudencia y pensar en terceras variables que pudieran explicar la asociación encontrada. En resumen, cabe decir que los coeficientes de correlación miden la *asociación* entre dos variables, pero una asociación estadística no se debe confundir con una relación causa-efecto.

A veces se usa la correlación equivocadamente para estimar en qué grado concuerdan dos sistemas de medida de una misma variable. Por ejemplo, ¿qué *concordancia* existe entre el peso que un individuo declara tener y el que realmente aparece en la báscula cuando se le pesa? Los coeficientes de correlación estiman la *asociación,* pero no la *concordancia* (**8,9**). Si todos los sujetos estudiados descontasen sistemáticamente, por ejemplo, el 10% de su peso, la correlación sería perfecta, pero la concordancia entre los dos pesos sería muy mala, como puede apreciarse gráficamente en la figura 10.4.

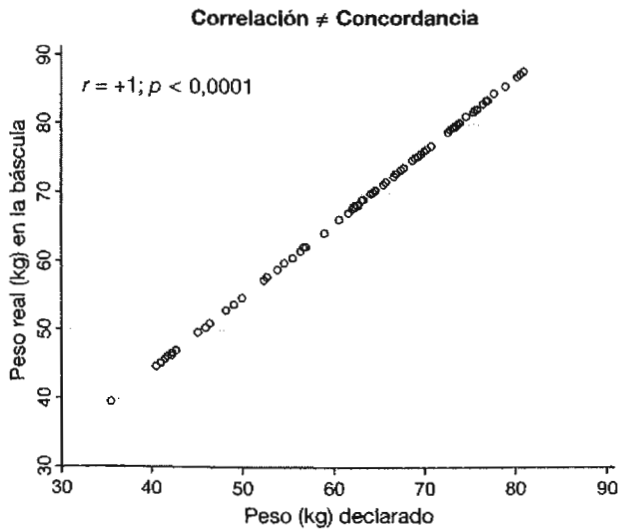



**Figura 10.4** La correlación resulta inadecuada para estimar la concordancia entre dos métodos de medida.

Se dispone de otros métodos específicos, diferentes al de la correlación, para estimar cuál es el grado de concordancia entre dos observaciones que pretenden medir lo mismo (**2,8-11**), tal como se verá en el capítulo 15.

También es fácil engañarse al encontrar que dos variables se correlacionan en medidas repetidas de las mismas efectuadas en el conjunto de un país o una región en el curso del tiempo. Los datos recogidos periódicamente en el tiempo (tendencia temporal) pueden manifestar tendencias cíclicas subyacentes y necesitan un tratamiento estadístico específico que desborda las pretensiones de este libro (**12**).

Existe el peligro de que las observaciones o puntos que se estudien sean solo una fracción *sesgada* de todas las observaciones posibles o existentes. Es decir, que no se seleccionen y analicen todas las posibles observaciones, sino solo aquellas que lleven al lector a una conclusión errónea. Este error sistemático se denomina *sesgo de selección* (**6,13**) y puede invalidar las conclusiones. Hay que ser lectores críticos y preguntarse siempre por qué se han elegido esos puntos y no otros. Por ejemplo, si un estudio presentase una correlación muy alta entre dos variables tomando un conjunto de países como unidades de observación, habría que preguntarse con qué criterio se seleccionaron esos países y no otros.

### 10.2.5. Test de hipótesis para la correlación

Para hallar la significación estadística del coeficiente de correlación en muestras grandes puede aplicarse un test que sigue la distribución *t* de *Student*. La hipótesis nula de este test será que el coeficiente de correlación vale 0 en la población.

$$H_0 \equiv r_{\text{población}} = 0$$

En este caso se ha evitado el uso de letras griegas para no confundir $\rho$ (coeficiente de Pearson *poblacional*) con *rho* (coeficiente no paramétrico de Spearman). Cuando el test resulte significativo, podemos rechazar la hipótesis nula de $r_{\text{población}} = 0$.

Se debe calcular el siguiente valor de *t*:

$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}$$

Después hay que comparar la *t* hallada con la tabulada para $n - 2$ grados de libertad. En el ejemplo de la correlación entre el consumo de tabaco y la incidencia de cáncer de pulmón existían 20 países. Por tanto, se tienen 18 grados de libertad y el valor de *t* sería:

$$t_{18} = 0,43\sqrt{\frac{18}{1-0,43^2}} = 2,02$$

Como el valor que aparece en la tabla para un error $\alpha$ de 0,05 a dos colas con 18 grados de libertad es $t_{18} = 2,101$, el valor de *t* obtenido a través de la fórmula se encuentra ligeramente por debajo del de la tabla, lo que se corresponde con un valor de la significación estadística *p* ligeramente por encima de 0,05. También se puede calcular con STATA mediante:

```
display 2*ttail(18, 2.02)
```

o bien con Excel, introduciendo =DISTRT(2,02;18;2) para obtener $p = 0,059$. Se concluye que la correlación encontrada está próxima al límite de la significación estadística. La interpretación más exacta sería que, si la muestra procediese de una población en la que el hábito tabáquico no guardase relación alguna con la incidencia de cáncer de pulmón ($r_{\text{población}} = 0$), la probabilidad de hallar en la muestra una correlación de 0,43 o más fuerte aún sería del 5,9%.

En este caso se ha desarrollado el cálculo para el coeficiente de correlación de Pearson, pero una expresión análoga también sería aplicable para el coeficiente de correlación no paramétrico de Spearman, si la muestra es grande ($n > 30$):

$$t_{n-2} = rho\sqrt{\frac{n-2}{1-(rho)^2}}$$

### 10.2.6. Intervalo de confianza para la correlación

En el estudio de la correlación es conveniente calcular los intervalos de confianza para el coeficiente de correlación. El cuadro 10.1 recoge los pasos que deben seguirse para calcularlo cuando la muestra sea grande.

En los casos en que el intervalo de confianza abarque el 0 (es decir, si el límite inferior resultase negativo y el superior positivo), se puede afirmar que no existe una correlación estadísticamente

**CUADRO 10.1 CÁLCULO DEL INTERVALO DE CONFIANZA PARA UN COEFICIENTE DE CORRELACIÓN**

1. *Transformar r en* $r_{trans}$
   La siguiente transformación facilita su tratamiento según una distribución normal:

$$r_{trans} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

   Para el ejemplo ($r$ = +0,4289):

$$r_{trans} = \frac{1}{2}\ln\left(\frac{1+0,4289}{1-0,4289}\right) = \frac{1}{2}\ln(0,9171) = 0,4585$$

2. *Calcular el error estándar de* $r_{trans}$

$$EE_{r_{trans}} = \frac{1}{\sqrt{n-3}}$$

   Para el ejemplo ($n$ = 20):

$$EE_{r_{trans}} = \frac{1}{\sqrt{20-3}} = 0,2425$$

3. *Sumar y restar «z» veces el error estándar a* $r_{trans}$
   Si el intervalo de confianza es al 95%, entonces $z$ = 1,96.

$$IC_{r_{trans}} = r_{trans} \pm z_{\alpha/2}\left(\frac{1}{\sqrt{n-3}}\right)$$

$$IC_{r_{trans}} = 0,4585 \pm (1,96)(0,2425) = -0,0168a + 0,9339$$

4. *Revertir la transformación*
   La transformación inversa a la realizada en el paso 1 es:

$$r = \frac{e^{2r_{trans}} - 1}{e^{2r_{trans}} + 1}$$

   Así se obtienen los límites inferior (LIC) y superior (LSC) de confianza al 95% para $r$:

$$r_{LIC} = \frac{e^{2\times -0,0168} - 1}{e^{2\times -0,0168} + 1} = -0,0168$$

$$r_{LSC} = \frac{e^{2\times 0,9339} - 1}{e^{2\times 0,9339} + 1} = +0,7324$$

significativa (para un error $\alpha$ bilateral del 5%) entre las dos variables. Cuando los dos límites de confianza sean negativos, será indicio de una asociación inversa; es decir, una variable disminuye cuando la otra aumenta.

Al interpretar un coeficiente de correlación se tendrán siempre en cuenta cuatro aspectos:

1. Signo.
2. Magnitud.
3. Significación estadística.
4. Intervalo de confianza (y su precisión).

***Interpretación:***

1. *El signo es positivo: esto supone que la asociación es directa, es decir, cuanto mayor es el consumo de tabaco, mayor es la incidencia de cáncer de pulmón.*
2. *Su magnitud es moderada, pues representa el 43% de la máxima posible.*
3. *La asociación no llega a ser estadísticamente significativa, aunque está cercana a la significación. La probabilidad de encontrar un coeficiente de correlación r muestral ≥ 0,43 si el r poblacional realmente fuese 0 es del 5,8% ($p_{2\ colas}$ = 0,058).*
4. *El intervalo de confianza es ancho y se tiene una confianza del 95% en que el r poblacional se encuentre en el rango –0,017 a +0,73. Este resultado denota gran imprecisión.*

## 10.3. COEFICIENTE DE CORRELACIÓN CON STATA

### 10.3.1. Coeficiente de correlación de Pearson, nivel de significación estadística y tamaño muestral

El coeficiente de correlación de Pearson puede obtenerse en STATA a través de dos menús:

**Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Correlations and covariances**

o

**Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Pairwise correlations**

o mediante las instrucciones:

```
correlate   v1 v2        vp
pwcorr      v1 v2        vp
```

siendo $v_1$-$v_p$ las variables cuantitativas implicadas en el análisis. Se realizarán todas las posibles comparaciones dos a dos entre variables.

La instrucción **pwcorr** presenta una serie de opciones (ausentes en la instrucción **correlate**) que permiten obtener, además del coeficiente de correlación, su significación estadística (valor *p* a dos colas) y el tamaño muestral de cada una de ellas. Las opciones son:

**obs:** permite visualizar el número de observaciones de cada comparación entre dos variables.

**sig:** presenta el nivel de significación estadística de cada comparación.

**print (#):** permite la visualización de *solo* aquellas correlaciones que alcanzan un determinado nivel (#) de significación estadística. Normalmente, este valor es 0,05 (**p(*.05*)**).

**star(#):** permite marcar con un *asterisco* las comparaciones con un nivel de significación estadística previamente especificado por el investigador. Por ejemplo, si este valor es 0,05: **st(*.05*)**.

En el ejemplo de la correlación entre el consumo de tabaco y la incidencia de cáncer de pulmón, se ha añadido también la variable consumo *per capita* de alcohol en el último año (l/año) *(Alcohol)*[4]. Se calcularán los coeficientes de correlación de Pearson para la comparación entre las tres variables incluidas en la base de datos.

```
. cor  Fumad Inccancer Alcohol
(obs=20)

             | Fumad Inccan~r  Alcohol
-------------+---------------------------
       Fumad |  1.0000
   Inccancer |  0.4289  1.0000
     Alcohol |  0.2870  0.3444  1.0000

. pwcorr  Fumad Inccancer  Alcohol, o sig

             | Fumad Inccan~r  Alcohol
-------------+---------------------------
       Fumad |  1.0000
             |      20
   Inccancer |  0.4289  1.0000
             |  0.0592
             |      20      20
     Alcohol |  0.2870  0.3444  1.0000
             |  0.2198  0.1371
             |      20      20      20

. pwcorr  Fumad Inccancer  Alcohol, p(.05) st(.05)

             | Fumad Inccan~r  Alcohol
-------------+---------------------------
       Fumad |  1.0000
   Inccancer |          1.0000
     Alcohol |                  1.0000
```

Se solicita a STATA el número de observaciones incluidas para el cálculo de los coeficientes de correlación, así como los niveles de significación estadística de dichos coeficientes (valores *p*)

Se solicita a STATA que muestre solo aquellos valores *p* de significación estadística que sean menores de 0,05 y que los marque con un asterisco

***Interpretación:*** *el primer análisis recoge, además del coeficiente de correlación de una variable consigo misma que siempre valdrá 1, el valor de los tres posibles coeficientes de correlación, uno para cada comparación entre un par de variables. El coeficiente de correlación de Pearson es: 0,4289 para la comparación Fumad-Inccancer (ya se había obtenido este mismo valor en el cálculo manual), 0,2870 para la comparación entre el consumo de tabaco y el consumo de alcohol (Fumad-Alcohol) y 0,3444 para la comparación entre el consumo de alcohol y la incidencia de cáncer de pulmón (Alcohol-Inccancer). El segundo análisis recoge, además, el valor p de significación estadística asociado a cada coeficiente, así como el número de observaciones incluidas en cada comparación (n = 20 países en todos los casos). Ninguno de los tres coeficientes r resultó estadísticamente significativo: p = 0,0592; p = 0,2198 y p = 0,1371, respectivamente). Cuando se solicita al programa que muestre los valores p asociados a los coeficientes r siempre y cuando estos valores p sean inferiores a 0,05, el programa no ofrece ningún dato, ya que ninguno de los coeficientes alcanza la significación estadística.*

### 10.3.2. Coeficiente de correlación de Spearman, nivel de significación estadística y tamaño muestral

Para calcular un coeficiente no paramétrico de Spearman debe acudirse al menú:

**Statistics → Nonparametric analysis → Tests of hypotheses → Spearman's Rank correlation**

4 El dato de consumo de alcohol en el último año se ha obtenido de datos procedentes de la OMS (http://www.who.int/en/). Puede accederse a la base de datos completa a través de nuestra página web: www.unav.es/preventiva.

o a la instrucción:

```
spearman   v1 v2          vp
```

Con esta instrucción, STATA calculará los coeficientes de correlación de Spearman, pero no el valor *p* de significación estadística asociado al mismo.

Al igual que en el ejemplo anterior, esta instrucción presenta algunas opciones, como **print** (#) y **star**(#) (antes mencionadas y explicadas).

En el ejemplo:

```
. spearman Fumad Inccancer Alcohol
(obs=20)
```

| | Fumad | Inccan~r | Alcohol |
|---|---|---|---|
| Fumad | 1.0000 | | |
| Inccancer | 0.4755 | 1.0000 | |
| Alcohol | 0.4299 | 0.1550 | 1.0000 |

```
. spearman Fumad Inccancer Alcohol, stats(rho p obs)
```

| Key |
|---|
| *rho* |
| *Number of obs* |
| *Sig. level* |

El cuadro *Key* presenta el orden de aparición de los valores en el listado STATA

| | Fumad | Inccan~r | Alcohol |
|---|---|---|---|
| Fumad | 1.0000 | | |
| | 20 | | |
| Inccancer | 0.4755 | 1.0000 | |
| | 20 | 20 | |
| | 0.0341 | | |
| Alcohol | 0.4299 | 0.1550 | 1.0000 |
| | 20 | 20 | 20 |
| | 0.0585 | 0.5141 | |

```
. spearman Fumad Inccancer Alcohol, p(.05) star(.05)
(obs=20)
```

| | Fumad | Inccan~r | Alcohol |
|---|---|---|---|
| Fumad | 1.0000 | | |
| Inccancer | 0.4755* | 1.0000 | |
| Alcohol | | | 1.0000 |

***Interpretación:*** *obsérvese que el coeficiente de correlación de Spearman para la comparación Fumad-Inccancer tiene un valor de 0,4755 y es estadísticamente significativo (p = 0,0341). De hecho, cuando se solicita a STATA la representación de únicamente los coeficientes significativos, el programa presenta este junto con un asterisco que señala que el coeficiente tiene un valor p <0,05.*

## 10.4. COEFICIENTE DE CORRELACIÓN CON OTROS PROGRAMAS

### 10.4.1. Coeficiente de correlación con Excel

Excel incluye entre sus funciones estadísticas el coeficiente de correlación. Se utilizarán los datos del ejemplo del uso de tabaco y la incidencia de cáncer de pulmón en 20 países (v. tabla 10.2). Estos datos han sido copiados en un archivo de Excel en el que se han situado, en la columna A, los valores de la variable X *Fumad* (desde la casilla A2 hasta la A21) y, en la columna B, los de la variable Y *Inc. cáncer pulmón* (desde la casilla B2 hasta la B21). Basta con escribir en una casilla la expresión =COEF.DE.CORREL(A2:A21;B2:B21) para que el programa devuelva el coeficiente de correlación *r* de Pearson.

B22 =COEF.DE.CORREL(A2:A21;B2:B21)

| | A | B | C |
|---|---|---|---|
| 1 | **Fumad (X)** | **Inc. cáncer pulmón (Y)** | |
| 2 | 37 | 42,4 | |
| 3 | 47 | 36,9 | |
| 4 | 33 | 57,1 | |
| 5 | 49 | 53,7 | |
| 6 | 39 | 60 | |
| 7 | 35 | 43,3 | |
| 8 | 37 | 53,3 | |
| 9 | 33 | 31,2 | |
| 10 | 36 | 47,7 | |
| 11 | 63 | 52,2 | |
| 12 | 33 | 47,4 | |
| 13 | 45 | 80,9 | |
| 14 | 29 | 31,6 | |
| 15 | 34 | 37,9 | |
| 16 | 34 | 45,4 | |
| 17 | 30 | 71,2 | |
| 18 | 26 | 38,2 | |
| 19 | 70 | 55,2 | |
| 20 | 17 | 18,2 | |
| 21 | 32 | 38,4 | |
| 22 | **Coeficiente correlación** | **0,428879497** | |

=COEF.DE.CORREL(A2:A21;B2:B21)

### 10.4.1.1. Intervalo de confianza para la correlación con Excel

Lamentablemente, es raro encontrar implementada la opción de intervalos de confianza para un coeficiente de correlación en el *software* habitual de bioestadística. Se recoge cómo se puede preparar un programa sencillo en Excel que calcule intervalos de confianza al coeficiente de correlación usando los datos del ejemplo.

| | | | |
|---|---|---|---|
| 1 | r muestral= | **0,4289** | Zona de introducción de los datos |
| 2 | n= | **20** | |
| 3 | IC al 95%= | **95** | |
| 4 | rtransf | 0,458547935 | = 0,5*(LN((1+B1)/(1-B1))) |
| 5 | EE(rtransf) | 0,242535625 | = 1/RAIZ(B2-3) |
| 6 | alfa medio (%) | 2,5 | = (100-B3)/2 |
| 7 | z | 1,959963985 | = -DISTR.NORM.ESTAND.INV(B6/100) |
| 8 | lic (rtransf) | -0,01681315 | = B4-(B7*B5) |
| 9 | lsc (rtransf) | 0,933909025 | = B4+(B7*B5) |
| 10 | LIC (95%)= | **-0,01681157** | = (EXP(2*B8)-1)/(EXP(2*B8)+1) |
| 11 | LSC (95%)= | **0,732411212** | = (EXP(2*B9)-1)/(EXP(2*B9)+1) |
| 12 | t | 2,014351706 | = B1*((B2-2)/(1-(B1^2)))^0,5 |
| 13 | p (2 colas)= | **0,059163757** | = DISTR.T(ABS(B12);B2-2;2) |

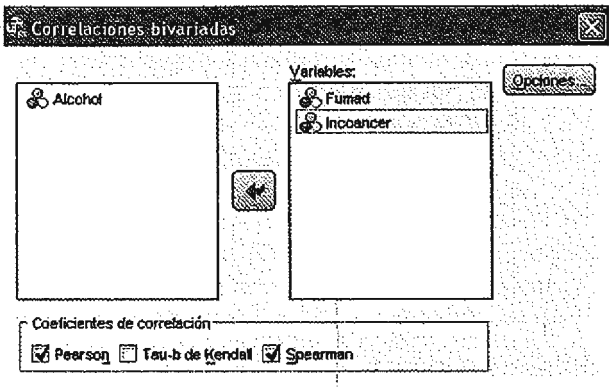
Las pequeñas diferencias con lo antes calculado se deben a los redondeos.

### 10.4.2. Coeficiente de correlación con SPSS

Se presenta el camino específico para calcular los coeficientes de correlación de Pearson y de Spearman en SPSS. De todas formas, el coeficiente de Pearson también aparece dentro de los resultados que proporciona este programa cuando se solicita la regresión lineal, ya que las dos técnicas (correlación y regresión) estiman relaciones lineales y están emparentadas matemáticamente.

Los pasos que han de seguirse son:

**Analizar → Correlaciones → Bivariadas...**




En el recuadro *Variables* se colocan las variables implicadas en el análisis. Se marcan *Coeficientes de correlación: Pearson, Spearman.*

Las salidas que proporciona SPSS para los coeficientes de Pearson y de Spearman son similares a las obtenidas con el programa STATA. SPSS presenta los coeficientes, sus valores *p* y el número de observaciones utilizadas para su cálculo.

### 10.4.3. Coeficiente de correlación con R/Splus

En R/Splus se puede importar una tabla de datos guardándola primero desde Excel (se usará la primera fila para los nombres de las variables). Se debe archivar como texto delimitado por tabulaciones y así se transformará en un fichero con extensión .txt. Después se importa ese fichero .txt desde R/Splus con la siguiente orden, en la que la nueva tabla de datos, ya en formato R/Splus, se denomina «dp»:

```
>dp< -read.table("c://d_precio.txt", header=T)
```

Se puede pedir a R/Splus que describa los datos que contiene «dp», simplemente escribiendo «dp». Se obtendrá el contenido de los datos:

```
> dp
```

| | tabaco | infl |
|---|---|---|
| 1 | 67 | 18 |
| 2 | 83 | 18 |
| 3 | 78 | 29 |
| 4 | 76 | 30 |
| 5 | 80 | 34 |
| 6 | 79 | 36 |
| 7 | 65 | 38 |
| 8 | 109 | 39 |
| 9 | 88 | 40 |
| 10 | 65 | 43 |
| 11 | 71 | 52 |
| 12 | 95 | 54 |
| 13 | 89 | 57 |
| 14 | 100 | 59 |
| 15 | 106 | 62 |

El coeficiente de correlación se puede solicitar con la orden:

```
> cor(dp$precio, dp$infl)

[1] 0.5096744
```

Antes de cada variable hay que indicar el nombre de la base de datos en que se encuentra dicha variable, separado por el signo del dólar ($). Así, dp$precio significa «la variable precio que está en la base de datos dp». Si se quiere obviar esto, se puede actuar en dos pasos:

```
> attach(dp)

> cor(precio, infl)

[1] 0.5096744
```

Para obtener la significación estadística es preciso recurrir a una función *homemade*, es decir, una función hecha en casa, que produzca el valor de *t* y el valor de *p* a dos colas. Se escribirá cada línea independientemente y se pulsará el cursor para pasar a la línea siguiente, esperando a que aparezca automáticamente el signo + antes de escribir la línea siguiente:

```
> cor.test<-function(x,y){
+ gl<-length(x)-2
+ r<-cor(x,y)
+ t<-r*((gl/(1-r^2))^0.5)
+ p<-2*(1-pt(t,gl))
+ c(t,p)}
> cor.test(precio,infl)
[1] 2.13589820 0.05228445
```

## 10.5. REGRESIÓN LINEAL SIMPLE

El análisis de ANOVA sirve para comprobar si una variable con más de dos categorías («factor» o *variable independiente*) tiene relación con una segunda variable cuantitativa (también llamada *respuesta* o *variable dependiente*). Sin embargo, existen dos problemas que no se pueden solucionar con el análisis de la varianza:

1. El ANOVA solo concluye indicando si existe o no asociación estadística entre dos variables, pero no define exactamente cuál es la magnitud de esa relación. ¿Cuánto aumenta la variable dependiente por cada unidad de incremento de la independiente?
2. Algunos factores tienen tantas categorías que parecen una variable cuantitativa, e incluso tal vez interese usar como *variable independiente* una variable cuantitativa.

La regresión viene a resolver estos dos problemas. En su forma más sencilla se llama *regresión lineal simple* y es una técnica estadística que analiza la relación entre dos variables cuantitativas, tratando de verificar si dicha relación es lineal. La regresión está dirigida a *describir* de una manera completa cómo se establece esta relación, de tal manera que incluso se puede *predecir* (siempre con cierto margen de error) cuál va a ser el valor de una variable una vez que se conoce el valor de la otra. Esta predicción puede resultar bastante imprecisa si la asociación entre ambas variables es débil, pero cuando la asociación es fuerte, la regresión ofrece un modelo estadístico que puede alcanzar finalidades predictivas. La *regresión* supone que hay una variable *fija,* controlada por el investigador (variable *independiente* o predictora, a veces considerada *«exposición»* o posible *«causa»*), y otra que no está controlada (variable de respuesta o *dependiente,* que ejerce el papel de *«respuesta»* o posible *«efecto»*)[5]. La variable dependiente ocupa siempre el eje de ordenadas (eje vertical o de la *y*); la independiente ocupa el eje de abscisas (eje horizontal o de la *x*). En la tabla 10.5 se presentan los distintos sinónimos y equivalencias que pueden tener las dos variables de una regresión.

La correlación y la regresión persiguen distintas finalidades y no siempre están indicadas al mismo tiempo (2). Ambas técnicas pueden confundirse erróneamente, ya que en las salidas de los programas de ordenador suelen aparecer mezcladas. Conceptualmente, la correlación está dirigida a medir el grado o fuerza de la asociación entre dos variables cuantitativas. En cambio, la regresión pretende definir la línea que mejor se ajusta a los puntos *(x,y)* para conseguir la mejor predicción de Y a partir de X. En la figura 10.5 se recogen las diferentes aplicaciones que puede tener cada método cuando se trata de relacionar dos variables cuantitativas.

La relación entre dos variables cuantitativas es bastante fácil de intuir. Por ejemplo, al ver crecer a un niño surge la pregunta acerca de si tendrá una estatura suficiente para su edad. Se espera que por cada incremento de edad (hasta los 25 años, después puede que suceda la contrario) se

***Tabla 10.5 Denominaciones y equivalencias de las dos variables que se usan en un modelo de regresión***

| VARIABLE INDEPENDIENTE (X) | VARIABLE DEPENDIENTE (Y) |
|---|---|
| Controlada por el investigador | Respuesta que no se controla |
| Información disponible | Información que se desea conseguir |
| Predictor | Desenlace predicho |
| Variable previa | Variable criterio |
| Exposición | Resultado |
| Posible «causa» | Posible «efecto» |
| Factor | Consecuencia |
| Regresor (Greenland, 1998) | Regresando (Greenland, 1998) |

5 Hay que tomar muchas reservas antes de hablar propiamente de causas y efectos, ya que, para realizar inferencias causales, se han de tener en cuenta otros criterios que no son estadísticos (14,15).

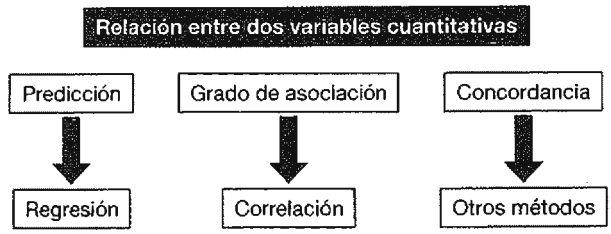



**Figura 10.5** Aplicaciones de la regresión y la correlación.

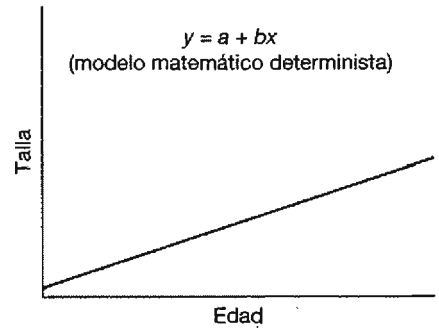



**Figura 10.6** Relación teórica entre la edad y la talla.

produce un incremento de talla (fig. 10.6). En el ejemplo, Y sería la talla y X la edad. El modelo responde a la sencilla ecuación de una recta *(y = a + bx),* en la que la talla sería igual a la suma de una constante *a* más la edad multiplicada por otra constante *b.* A la constante *a* se le llama la *ordenada en el origen,* que es el valor de Y cuando X vale 0 (sería la estatura de un recién nacido) y el coeficiente *b,* denominado *pendiente,* se interpreta como el incremento de Y por cada unidad de incremento de X, es decir, los centímetros que crece un niño al año.

Este modelo, que es fácil de entender, en principio puede considerarse determinista, en el que no hay errores ni variabilidad aleatoria; simplemente se dice que a tal edad le corresponde tal estatura. Pero en la realidad no sucede así. Nunca será posible realizar predicciones *perfectas* de la estatura que tendrá un niño una vez conocida su edad, ya que, aunque la edad tenga un efecto importante sobre la estatura, este efecto posee un cierto grado de variabilidad aleatoria (ruido o «error») y de ajuste inadecuado de los datos a la función que define el modelo, dado que existen otras muchas variables que influyen en la talla. En definitiva, se regresa al problema siempre presente en la estadística, la relación entre un «efecto» y un «error»:

$$\frac{\text{Efecto de la edad sobre la talla}}{\text{Error aleatorio}}$$

Una gráfica más realista se correspondería con la de la figura 10.7. En ella se recogen los puntos correspondientes a los valores de la talla (Y) y edad (X) para una grupo de niños. Se aprecia que no describen una línea recta perfecta, sino que existe un cierto grado de dispersión en torno a la línea recta imaginaria que los atravesaría por el centro. Ahora la ecuación incluye un nuevo término *(e)* que representa el error o residual (y también el desajuste de los datos con el modelo lineal). Es una cantidad variable de un sujeto a otro, puede ser positiva o negativa, y es la cantidad que

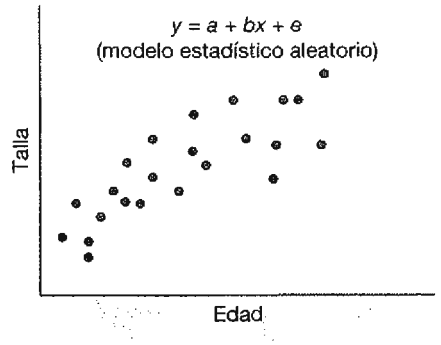



**Figura 10.7** Relación real entre edad y talla (incluyendo la variabilidad aleatoria).

habría que sumar o restar a la predicción realizada por el modelo para que coincida exactamente con lo observado en cada sujeto.

Para cada individuo existirá un valor predicho de la talla a partir de la edad ($y_{PRED} = a + bx$). A ese valor hay que sumarle o restarle una cantidad *e* para que coincida exactamente con la talla observada. El modelo real («imperfecto»), que se representa en la figura 10.7, es el que usa la estadística. Con él se podrá predecir la talla a partir de la edad, pero admitiendo un error *(e)*, que se llama *residual*, para la predicción de cada punto.

### 10.5.1. ANOVA de la regresión

El análisis de regresión comienza por un ANOVA. Si se usa la regresión para realizar la predicción de la talla (Y) a partir de la edad (X), el objetivo será trazar la línea recta que mejor se ajuste a los puntos. Esa recta establece una predicción de los valores que irá tomando Y (la talla) en función de X (la edad). La talla predicha por la recta en función de la edad ($y_{PRED}$) vendrá definida por la ecuación $y_{PRED} = a + bx$.

En la figura 10.8, además de la nube de puntos y de la recta de regresión, se ha dibujado la constante *a* u *ordenada en el origen* (valor de la talla cuando la edad vale 0) y el coeficiente *b* o *pendiente* de la recta (incremento de talla por cada año de edad). Se aprecia que la recta resume relativamente bien los puntos, pero casi ninguno de los puntos está exactamente sobre ella. La distancia entre cada

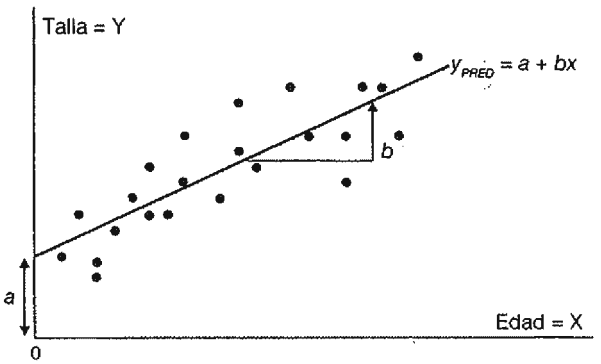



**Figura 10.8** Regresión lineal simple de la talla (variable dependiente, Y) y la edad (variable independiente, X).

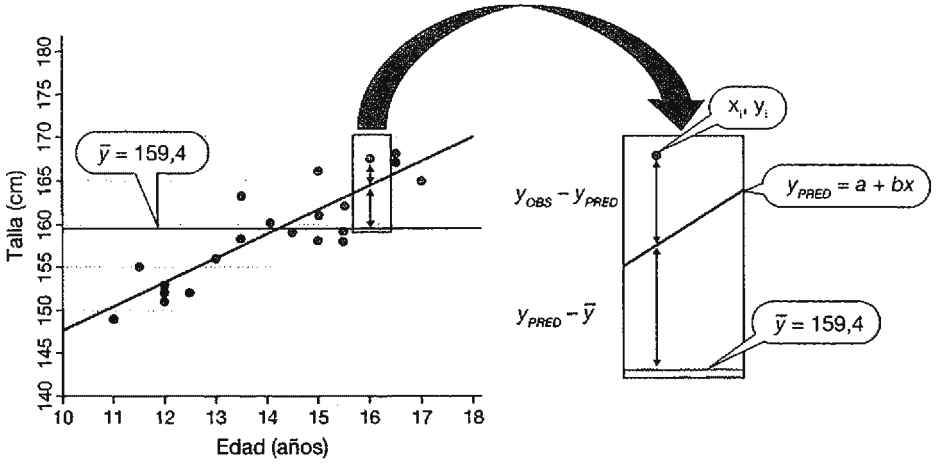



**Figura 10.9** Descomposición de la distancia desde la media de la variable dependiente ( $\bar{y}$ ) hasta cada punto. **Nota:** Se pone como ejemplo un punto que está por encima de lo predicho por la recta; esto sucederá para unos puntos, pero habrá otros que estén por debajo de lo predicho por la recta y su residual será negativo.

punto y la recta de regresión es el *residual (e; error aleatorio)* para cada punto. Esta distancia expresa el error aleatorio que existe en el modelo. ¿En qué medida es importante ese error? ¿Hasta qué punto es *más* importante el *efecto* de la edad sobre la talla que el *error* o variabilidad aleatoria? ¿Qué porcentaje de la variabilidad en la talla puede ser explicado por efecto de la edad y cuál no es explicado? Para responder a estas preguntas es preciso proceder a algo muy similar al análisis de la varianza.

La nube de puntos de la figura 10.9 recoge la relación entre edad y talla, en los datos reales de una muestra de chicas de 10 a 18 años (**16**). La talla observada para cada chica es la representada por cada punto y se denomina $y_{OBS}$. La media de la talla en esta muestra es 159,4 cm, y la media de la edad es 14,3 años. La recta de regresión que mejor resume la información contenida en esos puntos se ha representado con trazo grueso, mientras que la media de la variable talla ($\bar{y}$ = media de la talla) se ha marcado con trazo más fino. La recta de regresión siempre pasa por el punto correspondiente a la media de las X y la media de las Y. A este punto se le llama *centroide* o centro de gravedad ($\bar{x}, \bar{y}$).

Se parte de suponer que la edad no estuviese asociada con la talla y no sirviese para predecirla (hipótesis nula). En esta situación, la única predicción que podría hacerse de Y ya no dependería de X. ¿Cuál sería entonces la mejor predicción acerca del valor que va a tomar Y en un individuo concreto? Esa predicción se obtendrá simplemente a partir del valor más representativo de Y, que es su *media* ($\bar{y}$), sin tener en cuenta para nada el valor de la variable X en ese individuo. El análisis de la varianza que se hace en una regresión trata de contestar a esta pregunta: ¿es mejor usar X para predecir la variabilidad que existe en Y?, o ¿se puede conseguir la misma explicación de Y sin considerar los valores de X simplemente usando la media, $\bar{y}$? Se comparan dos predicciones:

- Predecir cada valor de Y teniendo en cuenta el valor de X en ese individuo ($H_1$).
- Predecir cada valor de Y aplicándole simplemente $\bar{y}$ ($H_0$).

La media de la talla ($\bar{y}$) es la hipótesis nula ($H_0$) y representa una predicción de la talla de inferior calidad por ser indiscriminada, ya que predice lo mismo para todos, sea cual sea su edad. La

hipótesis alternativa ($H_1$) propone que la edad es mejor que $\bar{y}$ para predecir la talla. Se comparan las dos predicciones, preguntándose si $H_1$ es significativamente mejor que $H_0$.

En la figura 10.9 puede observarse que la hipótesis nula ($H_0$) que corresponde a $\bar{y}$ es la línea horizontal fija en 159,4 cm. La pendiente de esta línea es 0.

Otra forma de expresar $H_0$ consistiría en decir que es la hipótesis que sostiene que en la población la recta tiene una pendiente $\beta$ igual a 0.

$$H_0 \equiv \beta = 0$$

La hipótesis alternativa ($H_1$) sostendría, en cambio, que la pendiente de la recta es significativamente diferente de 0.

$$H_1 \equiv \beta \neq 0$$

¿Hay que decantarse por $H_0$ o por $H_1$? Para responder a esta pregunta se debe calcular la suma de cuadrados total (SC total), que es la propia de la variable dependiente Y (*talla*, en el ejemplo).

$$\text{SC total} = \text{SC}y = \sum (y_i - \bar{y})^2$$

Esta cantidad es simplemente el numerador de la varianza de Y. Se puede calcular multiplicando la varianza de Y por sus grados de libertad ($n - 1$).

En el ANOVA de la regresión, la SC total (SC$y$) se descompone en dos porciones:

$$\text{SC total} = \text{SC}_{\text{hasta la recta}} + \text{SC}_{\text{desde la recta}}$$

En la figura 10.10 se ha enmarcado uno de los puntos, distinguiendo dos distancias: desde el punto a la recta y desde la recta hasta $\bar{y}$. Esta descomposición es la que se amplía a la derecha de la figura. La parte superior (distancia entre el punto observado y lo predicho por la recta) equivale al error o residual. La parte inferior se ha explicado ya por la regresión. En las sumas de cuadrados parciales ($\text{SC}_{\text{hasta la recta}}$ y $\text{SC}_{\text{desde la recta}}$) estas distancias se elevan al cuadrado.

Se obtendrán así las sumas de cuadrados (SC), como muestra la figura 10.10. Habrá una suma de cuadrados explicada por la regresión (SC regresión = $\text{SC}_{\text{hasta la recta}}$) y otra que no es explicada por la regresión y que se llama suma de cuadrados residual ($\text{SC}_{\text{desde la recta}}$). La suma de ambas será la suma de cuadrados total (SC Total). Con esto, tal como muestra la figura 10.10, se consigue descomponer la suma de cuadrados total en sus dos partes.

$$\text{SC total} = \text{SC regresión} + \text{SC residual}$$

$$\sum (y_{OBS} - \bar{y})^2 = \sum (y_{PRED} - \bar{y})^2 + \sum (y_{OBS} - y_{PRED})^2$$

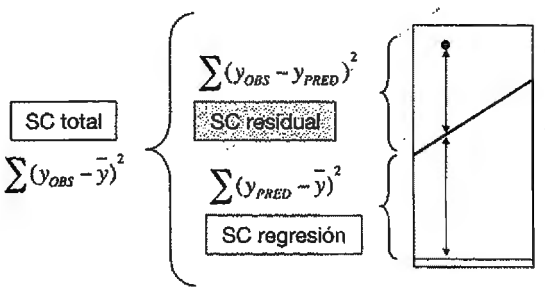



**Figura 10.10** Descomposición de la suma de cuadrados en una regresión.

lo que es equivalente a:

$$SC\ total = SC_{DESDE\ \bar{y}\ HASTA\ LA\ RECTA} + SC_{DESDE\ LA\ RECTA\ A\ CADA\ PUNTO}$$

En la parte derecha de la ecuación, el primer sumatorio corresponde a la explicación que aporta la recta de regresión y el segundo es la variabilidad residual no explicada por la recta. Por eso se les llama, respectivamente, suma de cuadrados de la regresión y suma de cuadrados residual.

Cuando se rechaza $H_0$, se dice que hay regresión de Y sobre X, ya que se puede explicar un porcentaje de los valores de Y a partir de los valores de X. Es decir, conocido el valor de X para un individuo, se predice mejor con la ecuación $(y_{PRED} = a + bx)$ que solamente con $\bar{y}$. Sin embargo, la predicción nunca es perfecta y queda algo sin explicar. Lo que queda sin explicar está expresado en la varianza residual que aparece en la tabla del ANOVA de regresión. La raíz cuadrada de la varianza residual es la desviación estándar residual. Una vez que se tiene en cuenta la variable X, el 95% de los residuales de Y se encontrarán aproximadamente en el intervalo de ±2 desviaciones estándar; esa desviación estándar es la desviación estándar residual (o error típico de la estimación).

Se dice que hay regresión de una variable Y sobre otra X cuando la segunda sirve para explicar la primera. Se nota en que la dispersión de los valores de los residuales de Y se reduce cuando se tiene en cuenta X (**17**). En la figura 10.11 se representan dos histogramas: el de la izquierda (v. fig. 10.11A) corresponde a la distribución del perímetro de la cintura (Y), y el de la derecha (v. fig. 10.11B) es el de los residuales de una regresión del perímetro de la cintura sobre el índice de masa corporal (X). Puede apreciarse que la dispersión se ha reducido. Una vez que se tiene en cuenta la variable X, la desviación típica se reduce desde 14,37 hasta 9,38 cm. Una desviación estándar inferior implica menor incertidumbre. Cuanto más se reduzca la desviación estándar, mejor capacidad de predicción tiene el modelo. Interesa, por tanto, comparar ambas desviaciones estándar, la de Y y la de los residuales. Si existe regresión, como en este caso, se reduce mucho la variabilidad de la distribución de los residuales con respecto a la variabilidad total de Y.

Para analizar los cálculos numéricos se usará un ejemplo muy sencillo. La tabla 10.6 recoge cinco puntos u observaciones, con sus correspondientes valores de X e Y.

Si se solicita a un ordenador que calcule la ecuación de la recta $(y_{PRED} = a + bx)$ para estos cinco puntos, la solución será: $y_{PRED} = 2{,}1 + x$. Es decir, la ordenada en el origen o constante $a$ vale +2,1 y el coeficiente de regresión o pendiente de la recta $b$ vale +1. Con esta ecuación se pueden predecir los valores de la variable Y $(y_{PRED})$ a partir de cada valor de la variable X. Así se ha hecho en la tabla 10.7. Existirá una diferencia entre cada valor de Y observado $(y_{OBS})$ y cada valor predicho $(y_{PRED})$. Esta diferencia es el *residual*. La suma de cuadrados residuales se obtiene elevando cada residual al cuadrado y sumándolos todos (v. tabla 10.7, última columna).

La suma de cuadrados residual (10,2) indica la magnitud de la variabilidad que queda sin explicar por la recta de regresión. Debe compararse con la suma de cuadrados total:

$$SC\ total = \sum (y_{OBS} - \bar{y})^2 = (6-8{,}1)^2 + (4-8{,}1)^2 + (7-8{,}1)^2 + (11-8{,}1)^2 + (12{,}5-8{,}1)^2 = 50{,}2$$

En la figura 10.12, se muestra que en esa SC total (50,2) hay una parte no explicada por la regresión, que es la SC residual (10,2). Lo restante (40) será lo que sí está explicado.

$$SC\ regresión = SC\ total - SC\ residual = 50{,}2 - 10{,}2 = 40$$

Para completar una tabla de ANOVA solo faltan los grados de libertad, que son $n - 1$ para el total, 1 para la regresión (pues hay una sola variable predictora X) y $n - 2$ para los residuales[6].

6 Si hubiese varios predictores $(p)$ o variables independientes ($X_1$, $X_2$, $X_3$, etc.), los grados de libertad de la regresión serían el número $p$ de predictores y los grados de libertad residuales serían $N - p - 1$.

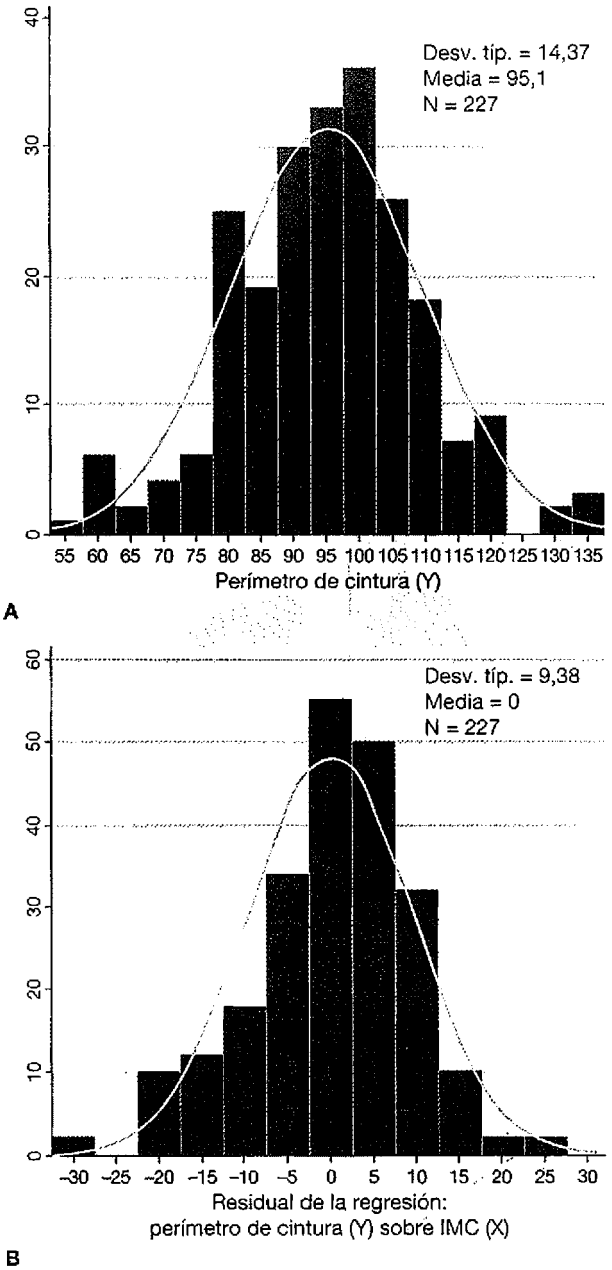



**Figura 10.11** **A.** Histograma del perímetro de la cintura (Y). **B.** Histograma de los residuales de la regresión de esta variable sobre el índice de masa corporal (IMC = X).

Como cualquier análisis de la varianza, el ANOVA de regresión concluye en un test $F$ que expresa el número de veces que es mayor la varianza explicada que la no explicada. En el ejemplo, ese test vale $F_{1,3} = 11{,}8$; los subíndices expresan que tiene un grado de libertad en el numerador y tres en el denominador. El valor $p$ de significación estadística que corresponde a una $F_{1,3} = 11{,}8$

*Tabla 10.6 Ejemplo sencillo de regresión lineal: datos para X e Y*

| X | Y |
|---|---|
| 2 | 6 |
| 4 | 4 |
| 6 | 7 |
| 8 | 11 |
| 10 | 12,5 |

*Tabla 10.7 Ejemplo sencillo de regresión lineal: predicciones de Y, residuales y residuales al cuadrado*

| X | $Y_{OBS}$ | $Y_{PREB}$ | RESIDUALES | RESIDUALES$^2$ |
|---|---|---|---|---|
| | | $y = 2,1 + x$ | $y_{OBS} - y_{PREO}$ | $(y_{OBS} - y_{PREO})^2$ |
| 2 | 6 | 4,1 | +1,9 | 3,61 |
| 4 | 4 | 6,1 | −2,1 | 4,41 |
| 6 | 7 | 8,1 | −1,1 | 1,21 |
| 8 | 11 | 10,1 | +0,9 | 0,81 |
| 10 | 12,5 | 12,1 | +0,4 | 0,16 |
| $\bar{x} = 6$ | $\bar{y} = 8,1$ | | | Suma = 10,2 |

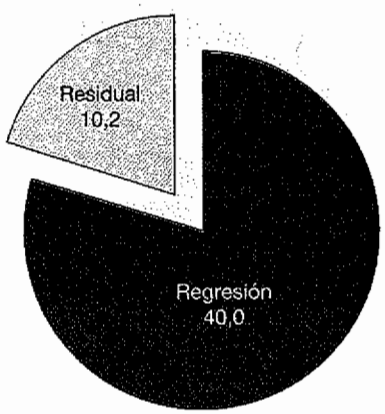



**Figura 10.12** Descomposición de la suma de cuadrados en un ANOVA de regresión.

se puede encontrar en las tablas o con Excel, donde =DISTR.F(11,8;1;3) devuelve $p = 0,041$. Por tanto, se puede rechazar la hipótesis nula de que estos cinco puntos provengan de una población con una pendiente $\beta$ que valga 0, y se concluirá que existe una asociación estadísticamente significativa entre X e Y, es decir, existe regresión de Y sobre X.

La tabla 10.8 presenta la descomposición de la variabilidad total en el ANOVA de una regresión lineal simple aplicada a los datos de las tablas 10.6 y 10.7.

### 10.5.2. Coeficiente de determinación, $R^2$

En el ejemplo anterior, la suma de cuadrados total valía 50,2 y expresaba el total de la variabilidad de Y. De esta cantidad, una porción es explicada linealmente por X. Esa porción explicada es 40 y corresponde a la suma de cuadrados de regresión. Se comprende entonces que pueda decirse que X es capaz de explicar el 80% de la variabilidad total de Y, ya que 40 es el 80% de 50,2. Este concepto corresponde a un coeficiente importante en regresión, que es el que compara lo explicado por la regresión con la variabilidad total de Y. Dicho coeficiente se llama $R^2$ o *coeficiente de determinación*:

*Tabla 10.8 Descomposición de la variabilidad total en el ANOVA de una regresión lineal simple (una sola variable X)*

| FUENTE | SUMA DE CUADRADOS | gl | VARIANZA | F |
|---|---|---|---|---|
| Regresión | $\sum(y_{PRED} - \bar{y})^2 = 40$ | 1 | $\frac{\text{SC regres.}}{1} = 40$ | $\frac{40}{3,4} = 11,8$ |
| Residual | $\sum(y_{OBS} - y_{PRED})^2 = 10,2$ | $n - 2 = 3$ | $\frac{\text{SC residual}}{n-2} = \frac{10,2}{3} = 3,4$ | |
| Total | $\sum(y_{OBS} - \bar{y})^2 = 50,2$ | $n - 1 = 4$ | | |

$$R^2 = \frac{\text{SC regresión}}{\text{SC total}}$$

Se interpreta como el *porcentaje de la variabilidad total de la variable dependiente (Y) que es explicado por la variable independiente (X)*. Los posible valores para $R^2$ van desde 1, que es el máximo (la recta daría una explicación perfecta, lo que supone que los valores de Y están totalmente determinados por X), a 0, que es el mínimo (la recta no explica nada, no existe asociación entre X e Y). Cuanto más próximo a 1 sea $R^2$, mayor es la fuerza de la asociación entre ambas variables.

La raíz cuadrada de $R^2$ equivale al *coeficiente de correlación* de Pearson[7], que se trató al principio de este capítulo:

$$\sqrt{R^2} = \pm r$$

A diferencia de $R^2$, que siempre es positivo, $r$ puede ser positivo o negativo (tendrá el mismo signo que la pendiente de la recta que hemos llamado $b$).

En el ejemplo utilizado, el coeficiente de determinación será:

$$R^2 = \frac{40,0}{50,2} = 0,797$$

A su vez, el coeficiente de correlación de Pearson será:

$$r = \sqrt{R^2} = \sqrt{0,797} = 0,893$$

Se sabe que su signo es positivo porque en la ecuación de la recta, $b$ tenía signo positivo ($b = +1$), y porque en la nube de puntos se apreciaba claramente que cuando crecía X, también lo hacía Y. Es importante hacer notar que, salvo en los casos extremos en que $R^2$ vale 0 o 1, la magnitud de $r$ es siempre superior a la de $R^2$. Este hecho tiene sus repercusiones prácticas. Una correlación puede parecer impresionante, por ejemplo $r = 0,7$, y, sin embargo, el modelo lineal explicaría menos del 50% de lo observado ($R^2 = 0,49$).

La relación entre $r$ y $R^2$ permite aplicar un método más sencillo para calcular $r$ que el visto al principio en este capítulo. Para calcular el coeficiente $r$ de correlación de Pearson, basta con extraer la raíz cuadrada del coeficiente de determinación ($R^2$). Si se conoce el valor de $R^2$, el coeficiente de Pearson se puede calcular simplemente como:

$$r = \pm\sqrt{R^2}$$

Es importante tener en cuenta que $R^2$ es siempre positivo, mientras que $r$ puede ser positivo o negativo. Si se aplica el método simple de cálculo de $r$, debe tomarse la precaución de asignar a

7 Solo se escribe con mayúsculas R cuando hay varias variables independientes, y entonces se llama coeficiente de correlación múltiple o R múltiple.

*r* el signo que tenga la pendiente *b* de la recta, ya que el signo de *r* debe coincidir siempre con el del coeficiente de regresión o pendiente de la recta.

### 10.5.3. Ajuste de una recta por mínimos cuadrados

Hasta ahora se ha hablado de predicciones a partir de una ecuación de regresión, pero el lector se estará preguntando por el modo de saber cuáles son los coeficientes *a* y *b* que definen la recta que mejor se ajusta a la nube de puntos. Esto supone conocer el valor de la ordenada en el origen *a* y de la pendiente de la recta *b* para obtener lo predicho por la recta:

$$y_{pred} = a + bx$$

Para calcular *a* y *b*, se usa el método de «mínimos cuadrados». Según este método, se demuestra que las ecuaciones son:

$$b = \frac{SPxy}{SCx}$$

$$a = \bar{y} - b\bar{x}$$

Puede apreciarse la semejanza de la ecuación de la pendiente de la recta *b* con la del coeficiente de correlación de Pearson *r*. El numerador es el mismo: la suma de productos de *xy* (SPxy), pero el denominador de *b* tiene en cuenta solo la suma de cuadrados de X (SCx). En cambio, en el coeficiente de correlación el denominador era la raíz cuadrada del producto de las dos sumas de cuadrados ($\sqrt{(SCx) \times (SCy)}$).

De hecho, cuando se conoce *r*, se puede calcular directamente *b* a partir de *r* con las desviaciones típicas ($s_y$, $s_x$) de las dos variables, según la expresión:

$$b = r \times \frac{s_y}{s_x}$$

En el presente ejemplo, se sabe que $r = 0{,}893$, la desviación estándar de X es 3,16 y la de Y es 3,54. Por lo tanto:

$$b = 0{,}893 \times \frac{3{,}54}{3{,}16} = 1{,}00$$

También se cumple lo contrario: si se conoce *b*, puede calcularse *r*, como muestra la figura 10.13.

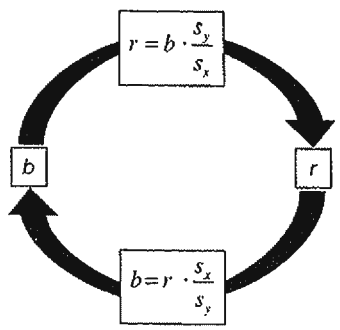



**Figura 10.13** Relación entre el coeficiente de correlación de Pearson *r* y la pendiente de la recta o coeficiente de regresión *b*.

Los coeficientes *a* y *b* se pueden obtener con muchas calculadoras y diversos programas de estadística. Si se desea calcular *b* sin conocer el valor de *r*, bastaría aplicar la primera fórmula explicada[8]:

$$b = \frac{\text{SPxy}}{\text{SCx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

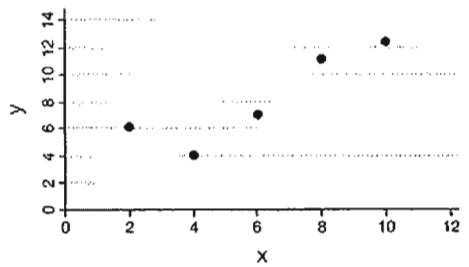
El objetivo principal de un análisis de regresión lineal suele ser el cálculo del valor de la *pendiente de la recta*. Esta pendiente, a la que se asigna la letra *b* en la ecuación, también se conoce como *coeficiente de regresión*. Es muy importante, porque mide el cambio de la variable Y por cada unidad de cambio de X. Su magnitud sirve para predecir cuánto aumentará Y cada vez que X se incremente en una unidad. A diferencia del coeficiente de correlación, que es simétrico, el coeficiente de regresión *no lo es*. Si X e Y se intercambiasen, cambiaría *b*. Otra gran diferencia entre el coeficiente de regresión y el de correlación es que el primero, *b*, se mide en las unidades de Y y de X. Es más, su magnitud *depende totalmente de las unidades de medida de Y y de X*. Por su parte, el coeficiente de correlación *r* no depende en absoluto de las unidades de medida.

El signo de *b* puede ser positivo o negativo. Si es positivo, a cada aumento de X le corresponde un incremento de Y; si es negativo, Y decrece a medida que aumenta X. Para este coeficiente de regresión o pendiente de la recta también se puede calcular un valor *p* de significación estadística. La hipótesis nula de este valor *p* es que la pendiente de la recta (en la población, $\beta$) vale 0.

$$H_0 = \beta = 0$$

Es decir, la hipótesis nula coincide con la del ANOVA de la regresión y mantiene que la variable X no predice nada de la variable Y. Por lo tanto, si se encontrase un valor $p = 0{,}04$, debería interpretarse así: si en la población de la que procede la muestra X no predijese nada de Y (o, lo que es lo mismo, Y no dependiese en absoluto de X), la probabilidad de encontrar una pendiente como la hallada o una todavía mayor en una muestra como esta sería del 4%. La significación estadística del coeficiente de regresión coincide exactamente con la del coeficiente de correlación y con la del ANOVA de la regresión.

Para aplicar estos conceptos al ejemplo anterior, se empieza por preparar un diagrama de dispersión (fig. 10.14). El ajuste por mínimos cuadrados consiste en buscar la recta que mejor resuma la información contenida en estos puntos, que será la que tenga el menor valor de los residuales al cuadrado. Se empezará por calcular la ordenada en el origen *a* y el coeficiente de regresión *b*. Se consigue con los cálculos intermedios detallados en la tabla 10.9.




| X | Y |
|---|---|
| 2 | 6 |
| 4 | 4 |
| 6 | 7 |
| 8 | 11 |
| 10 | 12,5 |

**Figura 10.14** Ejemplo de regresión: diagrama de dispersión o nube de puntos.

8 Se cometen menos errores de redondeo, pero es más difícil de memorizar la siguiente ecuación, que es equivalente:

$$b = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

*Tabla 10.9 Ejemplo sencillo de regresión lineal: cálculo de los coeficientes de la recta*

| X | Y | $(x_i-\bar{x})$ | $(x_i-\bar{x})^2$ | $(y_i-\bar{y})$ | $(x_i-\bar{x})(y_i-\bar{y})$ |
|---|---|---|---|---|---|
| 2 | 6 | –4 | 16 | –2,1 | 8,4 |
| 4 | 4 | –2 | 4 | –4,1 | 8,2 |
| 6 | 7 | 0 | 0 | –1,1 | 0 |
| 8 | 11 | 2 | 4 | 2,9 | 5,8 |
| 10 | 12,5 | 4 | 16 | 4,4 | 17,6 |
| $\bar{x}=6$ | $\bar{y}=8,1$ | | **Suma = 40** | | **Suma = 40** |

$$b=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=\frac{40,0}{40,0}=+1$$

$$a=\bar{y}-b\bar{x}=8,1-6=+2,1$$

La ecuación de la recta será:

$$y_{PRED}=2,1+x$$

Una vez calculada la ecuación de la recta, se realizará el ANOVA de la regresión (v. tabla 10.8)[9].

### 10.5.4. Error estándar del coeficiente *b* de regresión (o pendiente)

Además del test *F* de significación estadística, es especialmente importante la varianza residual. En el ejemplo de la tabla 10.8, la varianza residual valía 3,4. La raíz cuadrada de la *varResid* es la desviación estándar residual ($s_{resid}$ = 1,84, en el ejemplo). La varianza residual cumple un papel muy importante para calcular el error estándar de la pendiente ($EE_b$), ya que este error corresponde a la raíz cuadrada del cociente entre esta varianza residual y la suma de cuadrados de X:

$$EE_b=\sqrt{\frac{VarResid}{SCx}}$$

En el ejemplo anterior, el error estándar de la pendiente ($EE_b$) valdría:

$$EE_b=\sqrt{\frac{3,4}{40}}=0,292$$

Se puede tratar una pendiente *b* de manera análoga a la media de una muestra. Así, si se conoce el error estándar de *b* ($EE_b$ = 0,292), se pueden seguir dos pasos interesantes:

1. Estimar el intervalo de confianza de la pendiente, sumando y restando el error estándar a la pendiente encontrada, una vez que este error se multiplica por el valor correspondiente de la distribución *t* de *Student* (con los grados de libertad residuales, es decir, $n-2$ y el error $\alpha$ que se vaya a asumir).

$$IC(1-\alpha) \text{ de } b=b\pm t_{\alpha/2,n-2}\times EE_b$$

En el ejemplo, para un intervalo de confianza al 95%:

$$IC(95\%)=+1\pm t_{0,025,3}(0,292)=+1\pm(3,18\times 0,292)=+0,07 \text{ a } +1,93$$

***Interpretación:*** *por cada unidad de aumento de la variable X, la variable Y también experimenta un incremento de una unidad. Tenemos una confianza del 95% de que, en la población, esta variación de Y por unidad de X estará entre +0,07 y +1,93.*

9 Un «atajo» para calcular el ANOVA es la equivalencia SC regresión = $(SCx)b^2 = 40 \times 1^2 = 40$.

2. Llevar a cabo un test de hipótesis dividiendo el valor de la pendiente por su error estándar. Este test de hipótesis sigue una distribución *t* de *Student* y, a veces, se le llama test de Wald.

$$t_{n-2} = \frac{b}{EE_b}$$

Sustituyendo los valores del ejemplo, se obtendrá:

$$t_3 = \frac{1}{0{,}292} = 3{,}4 \rightarrow p_{2colas} = 0{,}04$$

Como el valor *p* a dos colas ($t = 3{,}4$ con tres grados de libertad) es 0,04, se rechazará la hipótesis nula y puede afirmarse que la pendiente es significativamente distinta de 0, es decir, hay evidencia de que X e Y están asociadas entre sí. El valor *p* del test de Wald es exactamente equivalente al del ANOVA de regresión ($F_{1,3} = 11{,}8$, como se recordará).

$$t_3 = \sqrt{F_{1,3}} = \sqrt{11{,}8} = 3{,}4$$

Resulta lógico, porque la hipótesis nula de ambos test es exactamente la misma.

El valor de *t* también coincide exactamente con el de la prueba de significación estadística del coeficiente de correlación *r* de Pearson. En el ejemplo, recordemos que *r* valía 0,893. Este valor, una vez calculado *b*, se puede obtener a través de las desviaciones estándar de las dos variables (v. fig. 10.13) por la siguiente expresión:

$$r = b \times \frac{s_x}{s_y} = (+1) \times \frac{3{,}16}{3{,}54} = +0{,}893$$

El test de significación estadística de la correlación (v. apartado 10.2.5) será:

$$t_{n-2} = r \times \sqrt{\frac{n-2}{1-r^2}} = 0{,}893 \times \sqrt{\frac{3}{1-0{,}797}} = 3{,}4$$

De esta forma, el error estándar de *a* es:

$$EE_a = \sqrt{\text{varResid} \times \left(\frac{(\bar{x})^2}{SCx} + \frac{1}{n}\right)} = \sqrt{3{,}4 \times \left(\frac{6^2}{40} + \frac{1}{5}\right)} = 1{,}93$$

### 10.5.5. Error estándar de la predicción e intervalo de confianza para la predicción media

La varianza residual que se obtenía en el ANOVA resultaba útil para estimar el error estándar de la pendiente. También tiene otra finalidad interesante: obtener intervalos de confianza para las predicciones a partir de la recta de regresión.

Para un valor dado de $X_i$, la recta predice una media para Y. Esta media puede representarse como la media de Y condicionada a que X tenga un valor concreto ($\bar{y} \mid x_i$). Esa predicción es un valor puntual y habrá que estimar su intervalo de confianza, partiendo, como siempre, de un error estándar. Si, por ejemplo, se desea obtener la predicción de cuál será el valor *medio* de Y cuando X vale 8 ($x_i = 8$), el error estándar vendrá dado por:

$$EE_{\bar{y}|x_i} = \sqrt{\text{varResid} \times \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SCx}\right)} = \sqrt{3{,}4 \times \left(\frac{1}{5} + \frac{(8-6)^2}{40}\right)} = 1{,}01$$

El intervalo de confianza sería

$$\overline{y} \mid x_i = a + bx \pm \left( t_{\alpha/2,n-2} \times \mathrm{EE}_{\overline{y}|x_i} \right)$$

$$(\overline{y} \mid x = 8) = 2{,}1 + [1 \times (8)] \pm (3{,}18 \times 1{,}01) = 10{,}1 \pm 3{,}21 = +6{,}89 \text{ a } +13{,}31$$

Podría pensarse en ir representando gráficamente todos los intervalos de confianza de $\overline{y} \mid x_i$ para todos los posibles valores que pueda tomar $X_i$. Se crearían así unas bandas de confianza alrededor de la recta de regresión que proporcionan una buena imagen de la capacidad predictiva de un modelo de regresión. Esta opción está implementada en la mayoría de los paquetes estadísticos y resulta muy sencilla de aplicar.

### 10.5.6. Intervalos de predicción individuales

Lo visto anteriormente se refería al rango de valores creíbles para la media ( $\overline{y}$ ) de la variable Y que se espera que posean aquellos que tienen un valor concreto $X_i$. También podría interesar predecir los valores que pueda tomar Y para un solo *individuo* ($y_i$) con un valor $x_i = 8$. Esta predicción será mucho más variable que la de la media, pero el procedimiento es bastante parecido al anterior:

$$\overline{y} \mid x_i = a + bx \pm \left( t_{\alpha/2,n-2} \times \sqrt{\mathrm{varResid}\left( 1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\mathrm{SCx}} \right)} \right)$$

Lo único que ha cambiado es que se ha añadido +1 dentro del paréntesis.

$$(\overline{y} \mid x_i = 8) = 2{,}1 + 8 \pm \left( 3{,}18 \times \sqrt{3{,}4 \times \left( 1 + \frac{1}{5} + \frac{(8-6)^2}{40} \right)} \right) = +3{,}42 \text{ a } +16{,}78$$

Como puede apreciarse, el intervalo de confianza para la predicción del valor individual ($y_i$) es considerablemente más amplio que el intervalo para la predicción de la media ($\overline{y}$). También se pueden obtener gráficas con bandas de predicción individual en los diferentes paquetes estadísticos.

## 10.6. USOS E INTERPRETACIÓN DE UNA REGRESIÓN LINEAL

Una de las utilidades más interesantes de la regresión es que puede definir la recta que idealmente iría uniendo las diferentes medias que toma Y para cada grupo de valores de X. Es decir, se calculará la media de Y a medida que cambia X.

En los siguientes ejemplos se deja de lado la parte del cálculo para centrarse en las consecuencias prácticas de la visualización de la recta de regresión. La figura 10.15 muestra un ejemplo procedente de una muestra representativa de la población adulta de 15 países miembros de la Unión Europea (7). Se relacionó el índice de masa corporal (IMC, en inglés, *body mass index*, abreviado como BMI)[10] con el número de horas por semana que, en su tiempo libre, la persona permanecía sentada. Se ha representado solo la recta correspondiente a las mujeres.

Se aprecia que, al aumentar las horas en posición sentada durante el tiempo libre, lo hacía también el índice de masa corporal. La pendiente de la recta suele ser el punto de mayor interés

10 El IMC es el indicador que se usa habitualmente en los estudios epidemiológicos de sobrepeso y obesidad, y se calcula dividiendo el peso en kg entre la talla en m elevada al cuadrado:

$$\mathrm{IMC} = \frac{\mathrm{P(kg)}}{\mathrm{T^2(m)}}$$

en cualquier análisis de regresión, ya que cuantifica la asociación entre X e Y. Sus unidades serán las de Y (kg/m$^2$) divididas por las de X (horas). En este ejemplo, las unidades de medida son distintas para las dos variables. La pendiente mide en cuántos kg/m$^2$ cambia el IMC por cada hora *más* de estar sentado a la semana. Se diría que por cada hora más que una mujer permanece sentada durante la semana, el IMC se incrementa en 0,03 kg/m$^2$. Este resultado se entiende mejor con unos breves cálculos. Por ejemplo, una mujer que mide 1,65 m y pesa 70 kg tiene un IMC de 70/(1,65$^2$) = 25,712. Para una mujer de estas características, cada hora más sentada supondría pasar de 70 kg a 70,08 kg. Cada 10 horas más que permanezca sentada, supondrán unos 800 g *más* de peso. Sin embargo, el coeficiente de determinación ($R^2$) es muy reducido, lo cual se interpreta diciendo que solo con la información sobre las horas de permanencia sentado se puede explicar muy poca variabilidad en el IMC (apenas el 0,6%). Los autores de este trabajo tuvieron en cuenta muchas otras variables para explicar el IMC, mediante el uso de regresión lineal múltiple.

En el ejemplo anterior (v. fig. 10.15) no se han representado los puntos, ya que se incluyó a miles de participantes y una nube con miles de puntos resultaría poco informativa. Pero, cuando el número de puntos es más reducido (<100), es muy recomendable presentar los resultados de una regresión acompañados de los puntos reales que originan la recta. También es conveniente presentar la varianza residual (o su raíz cuadrada, la desviación estándar residual). No se debe nunca prolongar la línea más allá del rango de los datos observados, ya que sería una extrapolación inválida por no sustentarse en los datos.

En el siguiente ejemplo (fig. 10.16) se compara la resistencia a la insulina medida mediante el modelo de homeostasis (índice HOMA) con el índice de masa corporal (IMC) en un subgrupo de participantes en el nodo de Navarra del ensayo PREDIMED (**18**). Se trata de 34 sujetos de alto riesgo cardiovascular, por lo cual ambos índices son elevados.

La presentación de los puntos junto con la recta de regresión permite, además de ofrecer los datos reales, hacerse una idea intuitiva de la variabilidad de la variable de respuesta que puede ser atribuida a la variable independiente y de la variabilidad que queda sin explicar. Este análisis

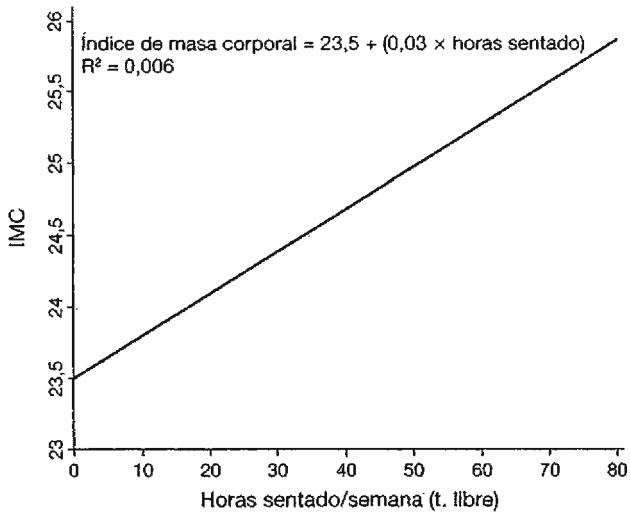



**Figura 10.15** Relación entre el índice de masa corporal (IMC = Y) y las horas sentado a la semana en el tiempo libre (X).

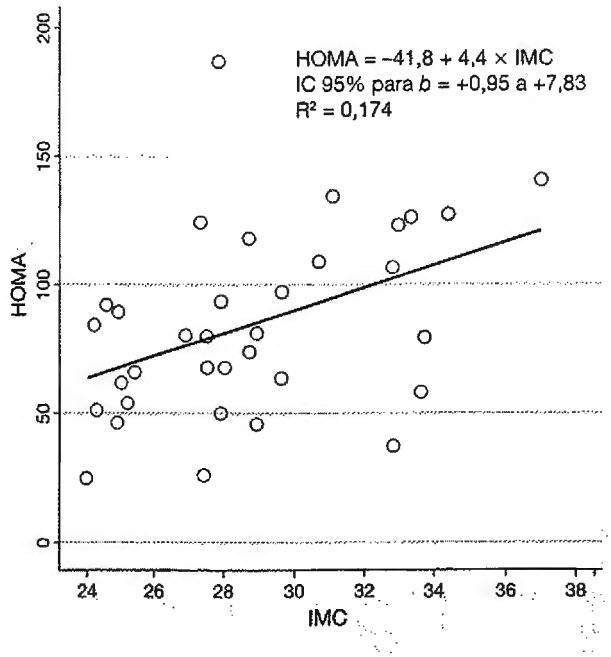



**Figura 10.16** Relación entre la resistencia a la insulina según el modelo de homeostasis (HOMA) (Y) y el índice de masa corporal (IMC) (X).

se intuye por lo separados que quedan los puntos de la recta (2). Mostrar la desviación estándar residual también ayuda a este propósito: mejora la presentación al ofrecer el intervalo de confianza de *b* (pendiente), como se ha hecho en el ejemplo de la figura 10.16.

## 10.7. SUPUESTOS DEL MODELO DE REGRESIÓN

Los supuestos que deben asumirse para realizar una regresión lineal son:

- *Normalidad* de la distribución *condicional* de la variable Y (variable dependiente). Se refiere no solo a que la variable Y siga una distribución normal, sino a que, además, para cada valor de X, la distribución de posibles valores de Y también siga una normal.
- *Linealidad.* Se determina que exista una relación lineal subyacente entre la variable X y las medias de la variable Y condicionadas a cada valor de X. Se supone que esta relación existe en la población de la que procede la muestra.
- *Homogeneidad de varianzas («homoscedasticidad»).* Las varianzas de la distribución de Y condicionadas a cada valor de X han de ser homogéneas.
- *Independencia* de las observaciones $y_i$. Cada observación de la variable Y debe ser independiente de las demás. Por ejemplo, en un estudio en el que Y representase el número de lesiones cutáneas en un brazo y hubiese pacientes en los que se han estudiado los dos brazos, habría dos observaciones por paciente autocorrelacionadas entre sí, no independientes; habría que considerar como *n* el número de pacientes, y no el número de brazos.

Otro ejemplo importante de transgresión del supuesto de independencia se da cuando se usa el tiempo como variable independiente, ya que un determinado valor suele condicionar el siguiente valor en el tiempo[11] (12).

Los primeros tres supuestos se consideran cumplidos si los *residuales* ($y_{pred} - y_i$) siguen una distribución normal. Se comprueba extrayendo los residuales para cada observación y aplicando después los procedimientos habituales para examinar la normalidad de una variable que incluyen test estadísticos y representaciones gráficas (v. apartado 10.8). Aunque existen métodos estadísticos para analizar el cuarto supuesto, es muy importante también la consideración del diseño empleado.

## 10.8. REPRESENTACIÓN GRÁFICA DE LOS RESIDUALES EN UNA REGRESIÓN LINEAL

Los residuales recogen aquella información que está presente en unos datos y que el modelo no ha sido capaz de resumir. Por este motivo, el estudio de los residuales es interesante en cualquier análisis de regresión.

Para comprobar la adecuación de unos datos a los supuestos de la regresión lineal pueden usarse varias representaciones gráficas de los residuales. Lo más habitual es recurrir a gráficos de normalidad. A modo de ejemplo, puede apreciarse el contraste entre los dos gráficos de la figura 10.17. Ambos gráficos analizan los residuales de modelos de regresión lineal simple. El de la izquierda (v. fig. 10.17A) es un gráfico *Q-Q* normal. Se observa que los residuales de este modelo se apartan totalmente de la normalidad y la aproximación a un modelo lineal no sería adecuada. Requeriría probar transformaciones de la variable independiente o introducir nuevos predictores o $X^{12}$. En cambio, el gráfico de la derecha (v. fig. 10.17B) (gráfico *P-P* normal) muestra que los puntos están situados casi exactamente en la diagonal, lo que conduce a concluir que el modelo lineal es sustancialmente correcto.

Además de los gráficos *P-P* y *Q-Q* de probabilidad normal, otra representación que resulta interesante para comprobar lo apropiado del modelo lineal es un diagrama de dispersión representando, como de costumbre, la variable independiente X en el eje de abscisas. Ahora bien, en vez de situar la variable dependiente Y en el eje de ordenadas, lo que se representa en ordenadas son los residuales. Es mejor presentar los residuales tipificados o normalizados (valores *z* de los residuales), lo que facilita la visualización de un patrón homogéneo sin la interferencia de las unidades de medida (fig. 10.18A). Se trata de comprobar si los residuales normalizados presentan una dispersión constante (homogeneidad de varianzas u homoscedasticidad) a lo largo del espectro de todos los posibles valores de la variable independiente (1-23). Si se apreciase que, a medida que crece la variable X, progresivamente la nube de puntos de los residuales se va «abriendo» (forma de embudo), entonces se tendría un indicio fundado de que la varianza no es constante, sino que existe disparidad de varianzas a lo largo de los posibles valores de X. En esta situación se dirá que los residuales son heteroscedásticos, o que hay heteroscedasticidad. Diferentes programas estadísticos realizan, además de estas gráficas de dispersión (eje de

11 Por ejemplo, el número de casos de meningitis que ocurren en 1 año puede influir mucho en el número de casos que existirán al año siguiente (por ejemplo, porque se puede decidir vacunar a todos los niños precisamente por haber observado muchos casos el año anterior; con otras enfermedades infecciosas será porque habrá más oportunidades de contagiarse, etc.). Es decir, el segundo valor de la variable comparada no es independiente, sino que está condicionado por el primero, el tercero por el segundo, y así sucesivamente. Esta situación, que se llama *autocorrelación*, exige aplicar técnicas de regresión especiales que se agrupan bajo el concepto de análisis de series temporales o modelos autorregresivos tipo ARIMA *(autoregressive moving average)*. Estas técnicas son muy usadas en los análisis estadísticos propios de la economía, pero hasta ahora han tenido escasas aplicaciones en epidemiología y medicina (19-21).

12 Podrían llevarse a cabo modelos cuadráticos, cúbicos, etc. Estos modelos no siguen la ecuación de la recta, sino otras ecuaciones del tipo $y = a + b_1x_1 + b_2x_1^2$; $y = a + b_1x_1 + b_2x_1^2 + b_3x_1^3$ o $y = a + b_1\frac{1}{x_1}$. También podrían incluirse nuevas variables independienres ($X_2, X_3 \ldots X_p$) en el modelo de regresión lineal (que se transformaría en un modelo de regresión lineal múltiple).

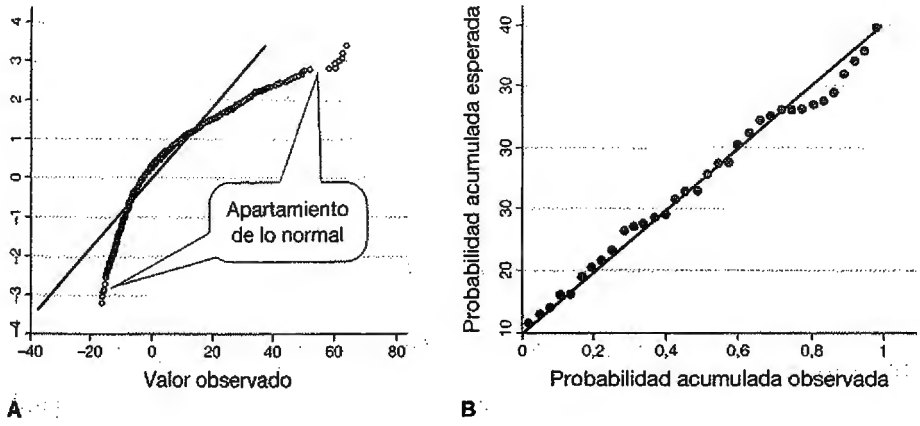



**Figura 10.17** Representación gráfica de los residuales de dos modelos distintos de regresión lineal simple. **A.** Gráfico *Q-Q* normal. **B.** Gráfico *P-P* normal.

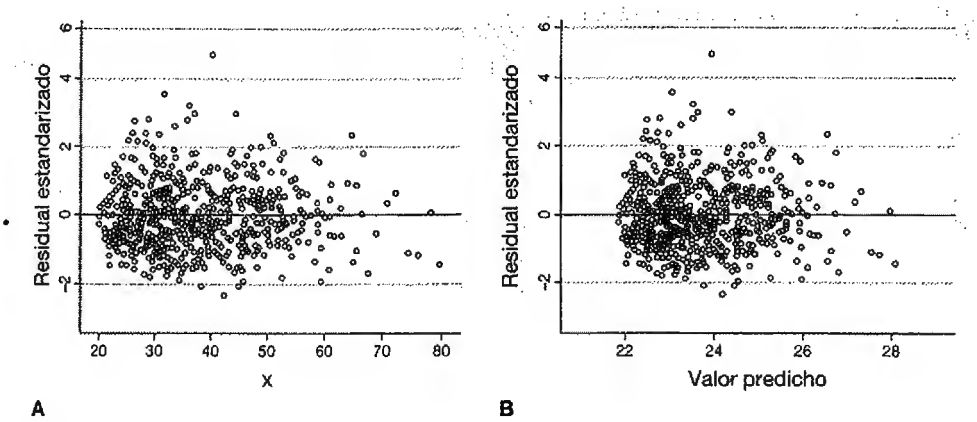



**Figura 10.18** Representación gráfica de los residuales para valorar si existe heteroscedasticidad. **A.** Valor residual tipificado *vs.* valor X. **B.** Valor residual tipificado *vs.* valor Y predicho.

ordenadas: residual, y eje de abscisas: valores de X), otro tipo de gráfico para valorar la heteroscedasticidad. Se trata de un gráfico de dispersión que muestra los valores residuales en el eje de ordenadas y el valor Y predicho por el modelo en el de abscisas (fig. 10.18B). La figura 10.19 recoge una distribución de los residuales que resulta heteroscedástica. Se ha valorado si la ingesta de grasa monoinsaturada en la dieta (g/día) depende de las calorías totales (kcal/día) ingeridas. En la parte izquierda de la figura (v. fig. 10.19A) se recoge el diagrama de dispersión que resume los resultados de la regresión, y en el de la derecha se valora si la varianza de los residuales es homogénea (v. fig. 10.19B). Se llega a la conclusión de que no lo es, sino que aumenta a medida que lo hacen las calorías consumidas. En la gráfica de la derecha puede apreciarse la forma de embudo de los residuales.

En esta situación (heteroscedasticidad) no se cumplirán bien los supuestos del modelo de regresión, salvo que la muestra sea muy grande. Una solución consiste, por tanto, en ampliar

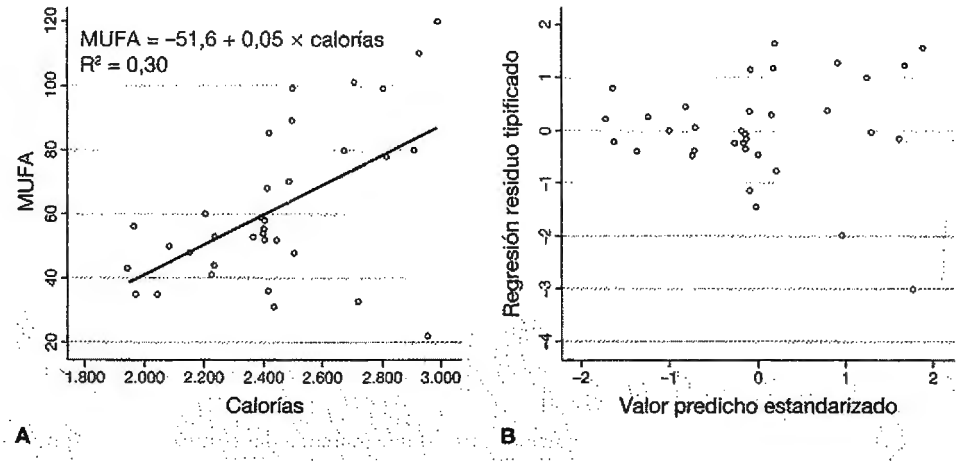



**Figura 10.19** Relación entre la ingesta de grasa monoinsaturada (MUFA) (Y) y las calorías totales ingeridas (X). **A.** Valor Y *vs.* valor X. **B.** Valor residual tipificado *vs.* valor Y predicho tipificado.

la muestra (**22**). Otra idea sería dividir la muestra en varios subgrupos según las calorías consumidas y ajustar una regresión distinta para cada subgrupo. Así, dentro de cada subgrupo, las varianzas de los residuales serían más homogéneas y no se violaría el supuesto de homoscedasticidad.

Puede recurrirse a otros textos (**23-27**) para profundizar más en los extensos procedimientos que se han desarrollado para el análisis de los residuales, los diagnósticos de regresión o la identificación de *outliers* y puntos influyentes. Algunos de estos procedimientos se encuentran disponibles en SPSS y STATA. También se han ofrecido interesantes revisiones que presentan ejemplos relevantes de análisis de regresión lineal en las publicaciones médicas (**28**).

## 10.9. CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LINEAL CON STATA E INSTRUCCIONES POSTESTIMACIÓN

Para llevar a cabo un análisis de regresión lineal simple con STATA, se utilizará una base datos en la que se ha recogido el peso de un total de 10 recién nacidos (*peso_rn*, en gramos) y la edad de sus madres (*edad*, años). Se desea valorar si la edad de las madres (X) guarda relación con el peso de sus descendientes (Y).

### 10.9.1. Realizar un gráfico de dispersión

El primer paso debe consistir siempre en solicitar un gráfico de dispersión para apreciar visualmente si se puede asumir un modelo lineal entre ambas variables.

Se ajustará una regresión cuando la nube de puntos apunte a que existe una relación *lineal*. Una nube de puntos puede sugerir que no existe ninguna relación (fig. 10.20A). En ausencia de relación, la pendiente *b* será igual a 0. También puede obtenerse una pendiente de 0 por un motivo distinto: que exista relación, pero no lineal (**26**), sino que dicha relación siga una curva u otro tipo de función (alejamiento de la linealidad)[13] (fig. 10.20B).

13 Otras posibles relaciones son la cuadrática, la cúbica, la racional, etc. Estos modelos no siguen la ecuación de la recta, sino otras ecuaciones del tipo $y = a + b_1x_1 + b_2x_1^2$; $y = a + b_1x_1 + b_2x_1^2 + b_3x_1^3$ o $y = a + b_1\frac{1}{x_1}$

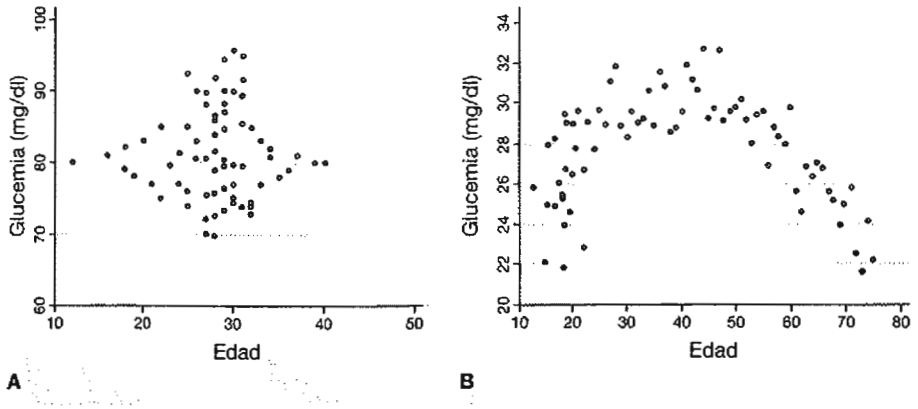
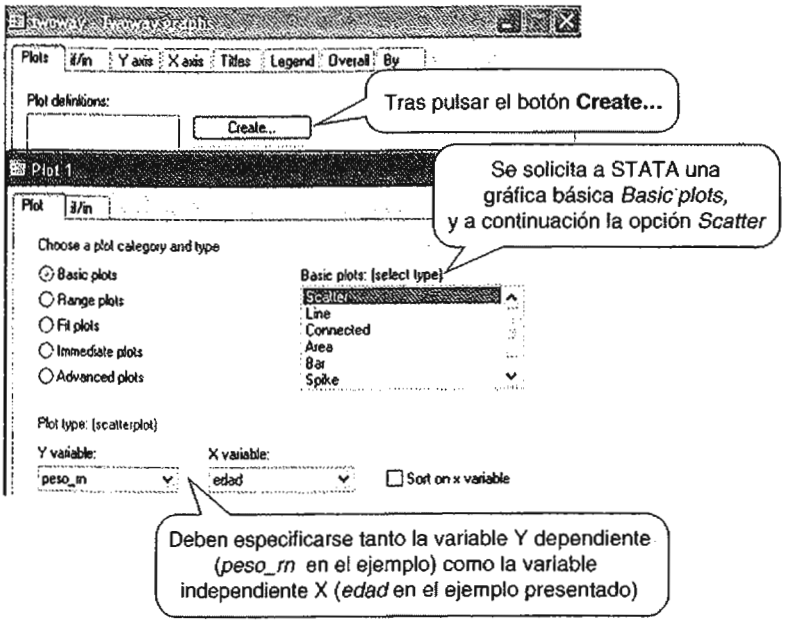



**Figura 10.20** Dos nubes de puntos con pendiente igual a 0 que muestran dos situaciones muy distintas. **A.** No existe relación. **B.** Relación no lineal.

En STATA, los gráficos de dispersión pueden producirse a través del menú:
**Graphics → Twoway graph (scatter, line, etc.)**




o mediante instrucciones[14]:

```
twoway scatter y x
```

14 Otras instrucciones que conducen a la misma representación gráfica serían:

```
graph twoway scatter y x
twoway (scatter y x)
scatter y x
```

Junto con esta instrucción, existen otras muy útiles que pueden aplicarse sobre la nube de puntos del gráfico de dispersión y que complementan este primer análisis gráfico de los datos.

1. Por ejemplo, STATA puede trazar la línea de predicción de Y a partir de los datos de X ajustando el modelo lineal:

```
twoway lfit y x
```

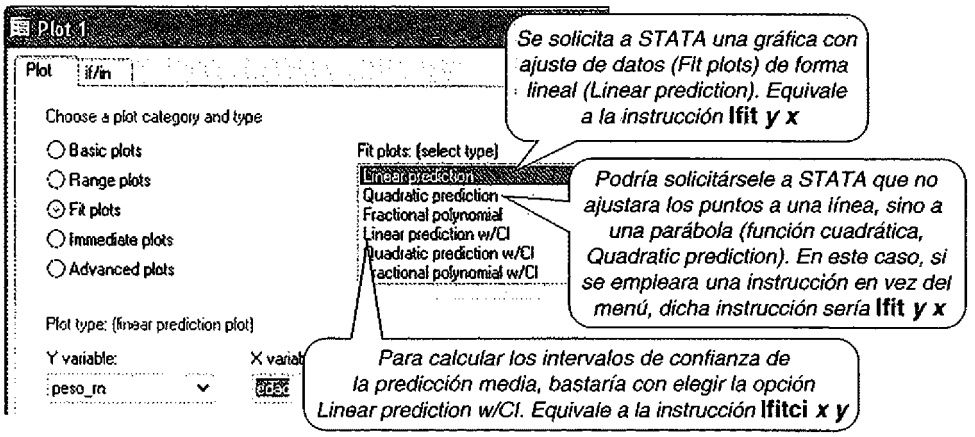
2. El intervalo de confianza para la predicción media (v. apartado 10.5.5) se obtiene a través de la instrucción:

```
twoway lfitci y x
```

3. Finalmente, STATA traza los intervalos de confianza de las predicciones individuales, es decir, para cada observación predicha $y_i$ (v. apartado 10.5.6), a través de la instrucción:

```
twoway lfitci y x, stdf
```

Estos mismos procedimientos pueden ser solicitados a través del menú del programa:
**Graphics → Twoway graph (scatter, line, etc.)**

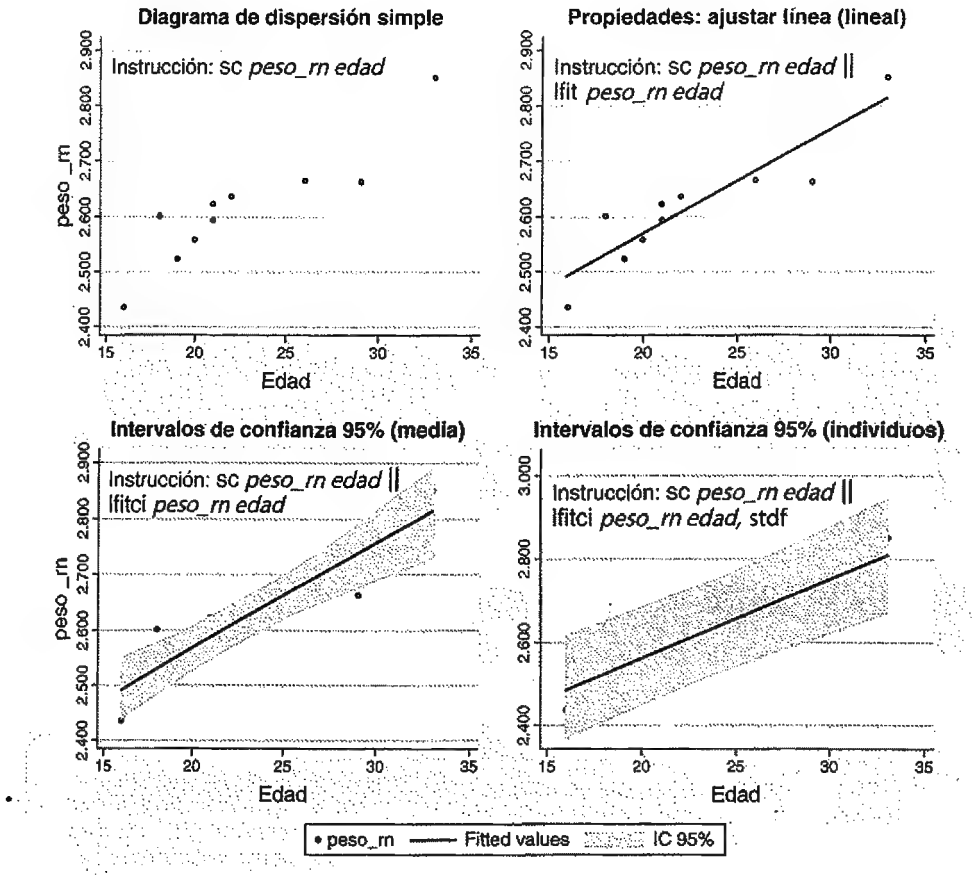



En el ejemplo concreto de la predicción del peso del recién nacido a partir de la edad de la madre[15]:

```
scatter peso_rn edad || lfit peso_rn edad
scatter peso_rn edad || lfitci peso_rn edad
scatter peso_rn edad || lfitci peso_rn edad, stdf
```

15 Podrían utilizarse las expresiones:

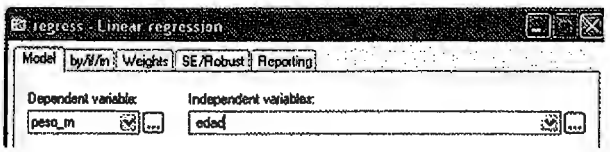
```
tw (sc peso_rn edad) (lfit peso_rn edad)
tw (sc peso_rn edad) (lfitci peso_rn edad)
tw (sc peso_rn edad) (lfitci peso_rn edad, stdf)
```




### 10.9.2. Ajustar el modelo de regresión lineal simple

La regresión lineal simple puede llevarse a cabo con STATA utilizando el menú o a través de instrucciones. Con el menú:

**Statistics → Linear models and related → Linear regression**




Utilizando instrucciones:

```
regress y x
```

siendo *y* la variable dependiente y *x* la variable independiente.

En el ejemplo de la predicción del peso del recién nacido a partir de la edad de la madre:

```
. regress peso_rn edad
```

| Source | SS | df | MS | Number of obs = | 10 |
|---|---|---|---|---|---|
| | | | | F( 1, 8) = | 37.77 |
| Model | 87024.1956 | 1 | 87024.1956 | Prob > F = | 0.0003 |
| Residual | 18430.2044 | 8 | 2303.77555 | R-squared = | 0.8252 |
| | | | | Adj R-squared = | 0.8034 |
| Total | 105454.4 | 9 | 11717.1556 | Root MSE = | 47.998 |

| peso_rn | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edad | 18.63872 | 3.032608 | 6.15 | 0.000 | 11.64552 | 25.63193 |
| _cons | 2195.229 | 69.90144 | 31.40 | 0.000 | 2034.036 | 2356.422 |

***Interpretación:*** *en primer lugar, STATA muestra el ANOVA de la regresión (v. apartado 10.5.1). Las iniciales SS equivalen a la «suma de cuadrados» en inglés (Sum of Squares), df a los «grados de libertad» (Degree of Freedom) y MS al término «media cuadrática» (Mean Square) o, lo que es lo mismo, varianza. Es importante destacar la varianza residual (MS Residual = 2.303,78), que, además de otras utilidades, permitirá calcular el error estándar de la pendiente (EEb).*

*El valor del estadístico F es obtenido a partir del cociente entre la varianza del modelo y la varianza residual. En este ejemplo, su valor es de 37,77 (87.024,20/2.303,78). Debe destacarse, por su importancia, la significación estadística de este test F (p = 0,0003), que, en este ejemplo, permite rechazar la hipótesis nula de pendiente igual a 0. Se puede afirmar que hay evidencia de una asociación estadísticamente significativa (p = 0,0003) entre las variables X e Y. Después aparece el coeficiente de determinación*[16] *($R^2$)(R-squared) = 0,8252. Puede decirse que el 82,5% de la variabilidad del peso del recién nacido puede explicarse por la edad de la madre, es decir, por el modelo.*

*Por último, STATA ofrece los valores de la recta de regresión (v. apartado 10.5.3). En la columna Coef. se sitúa el valor de la pendiente de la recta o coeficiente de regresión (b = 18,64) asociado a la variable independiente (edad, en este ejemplo) y el valor de la ordenada en el origen o constante del modelo matemático (a = 2.195,23). Por cada año más de edad de la madre, por término medio, el recién nacido pesa 18,64 g más. Dados a y b, se podría predecir el peso de un recién nacido a partir de la edad de la madre (peso_rn = 2.195,23 + 18,64 × edad). En la siguiente columna se presentan los respectivos errores estándar (Std. Err) de a y b. Aunque el error estándar de la pendiente (EEb = 3,03) es el que reviste mayor interés, STATA también calcula el error estándar de la ordenada en el origen (EEa = 69,9). En la siguiente columna aparece el valor de t calculado al dividir cada coeficiente por su respectivo error estándar (test de Wald). Cada valor de t corresponde a un valor p de significación estadística (a dos colas), que aparece en la columna adyacente (P > | t |). La hipótesis nula para este test es que el coeficiente respectivo vale 0. En este ejemplo, se puede rechazar la hipótesis nula de que la recta corte al eje de coordenadas (ordenada en el origen poblacional = 0, p ≤ 0,001) y de que la pendiente poblacional sea 0 (p ≤ 0,001). Por último, en las dos últimas columnas aparecen los intervalos de confianza al 95% para la ordenada en el origen y para la pendiente de la recta. Cuando el intervalo de confianza al 95% incluya el valor nulo (0), no existirán diferencias estadísticamente significativas (p > 0,05). No sucede así en este ejemplo. El intervalo de confianza solo presenta valores positivos que determinan que hay que rechazar la hipótesis nula. La edad de la madre predice el peso de su recién nacido; al aumentar la edad de la madre, su descendencia tendrá siempre más peso (con un nivel de confianza del 95% se puede afirmar que, por cada año más de edad, este incremento puede oscilar entre 11,6 y 25,6 g).*

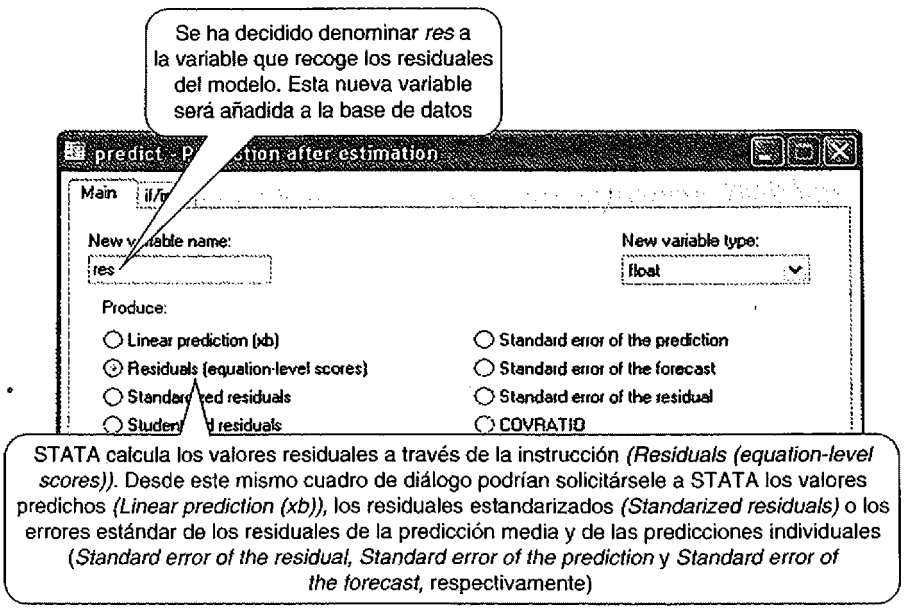
16 Junto con el coeficiente de determinación, aparece el denominado R cuadrado corregido *(Adj R-squared)* = 0,8034, que siempre será menor que R cuadrado y que tiene utilidad para comparar entre sí la capacidad predictiva de varios modelos con *distinto número de predictores (modelos de regresión lineal múltiple)*. Cuando se quieran comparar modelos con distintos número de variables independientes, se debe usar la $R^2$ corregida de los modelos que se comparan.

### 10.9.3. Guardar valores predichos y residuales y sus correspondientes errores estándar

A través del método de mínimos cuadrados, STATA calcula los coeficientes *a* y *b*. Por tanto, a través de la ecuación de la recta pueden calcularse los valores pronosticados o predichos de Y ($y_{PRED}$). Los valores residuales se calcularán a través de la fórmula: $y_{OBS} - y_{PRED}$. Por último, en ocasiones se utilizan los valores residuales tipificados o estandarizados, que son calculados restando a cada valor residual su valor medio y dividiendo a continuación este valor por su desviación estándar (como se hace en los valores *z* de la distribución normal). Otro procedimiento habitual consiste en calcular los errores estándar de la predicción media, de los valores predichos individuales y de los valores residuales.

STATA es capaz de realizar estos cálculos y guardar estos valores en forma de nuevas columnas en la base de datos a través del siguiente menú:

**Statistics → Postestimation → Predictions, residuals, etc.**




O con las instrucciones:

```
predict ypred, xb
```

Se consigue lo mismo solo con:

```
predict ypred
```

(Se calculan los valores predichos por el modelo que son introducidos como una nueva variable a la que se ha decidido llamar *ypred* en este ejemplo. STATA calculará los valores predichos sin necesidad de introducir la subinstrucción **xb**).

```
predict res, residuals
```

(Se calculan los valores residuales. La nueva variable recibirá el nombre *res*. STATA guarda los valores residuales igualmente con las subinstrucciones **resid** o **score**).

```
predict zres, rstandard
```

(Se calculan los valores residuales estandarizados. La nueva variable recibirá el nombre *zres*).

```
predict eeres, stdr
predict eep, stdp
predict eeind, stdf
```

(Se calculan los valores de los errores estándar de los residuales, de la predicción (*y* media) y de cada valor individual ($y_i$). Las nuevas variables han sido etiquetadas con los nombres *eeres*, *eep* y *eeind*, respectivamente.)

A continuación se muestra el aspecto final de la base de datos de la predicción del peso del recién nacido según la edad de la madre:

| | edad | peso_rn | ypred | res | zres | eep | eeres | eeind |
|---|---|---|---|---|---|---|---|---|
| 1 | 19 | 2523 | 2549.365 | -26.36447 | -.5954006 | 18.52127 | 44.28022 | 51.44719 |
| 2 | 33 | 2850 | 2810.307 | 39.69341 | 1.219484 | 35.27485 | 32.54936 | 59.56585 |
| 3 | 20 | 2557 | 2568.003 | -11.00319 | -.2450656 | 16.96635 | 44.89898 | 50.90808 |
| 4 | 21 | 2594 | 2586.642 | 7.358084 | .1624058 | 15.84519 | 45.30679 | 50.54548 |
| 5 | 18 | 2600 | 2530.726 | 69.27425 | 1.594656 | 20.41105 | 43.44151 | 52.15733 |
| 6 | 21 | 2622 | 2586.642 | 35.35809 | .7804146 | 15.84519 | 45.30679 | 50.54548 |
| 7 | 22 | 2637 | 2605.281 | 31.71936 | .696986 | 15.25375 | 45.50933 | 50.3632 |
| 8 | 16 | 2435 | 2493.448 | -58.4483 | -1.423944 | 24.87848 | 41.04677 | 54.06213 |
| 9 | 29 | 2663 | 2735.752 | -72.75169 | -1.77241 | 24.87848 | 41.04677 | 54.06213 |
| 10 | 26 | 2665 | 2679.835 | -14.83553 | -.3350374 | 18.52127 | 44.28022 | 51.44719 |

### 10.9.4. Comprobar la normalidad de los residuales del modelo

La normalidad de los residuales (variable creada y guardada por el programa según procedimiento del apartado anterior) puede comprobarse a través de pruebas de normalidad y/o con representaciones gráficas del tipo *Q-Q* y *P-P*.

Existen diferentes pruebas de normalidad, como el test de Shapiro-Wilk, el test de Shapiro-Francia o el test de la asimetría y curtosis *(skewness and kurtosis)*. Todas estas pruebas pueden realizarse a través del menú del programa:

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Shapiro-Wilk normality test**

**Shapiro-Francia normality test**

**Skewness and kurtosis normality test**

Sus correspondientes instrucciones son:

```
swilk res
sfrancia res
sktest res
```

(Siempre y cuando la nueva variable que recoge los residuales del modelo haya sido denominada *res*).

Los resultados del ejemplo del peso del recién nacido indican que no existen indicios para rechazar la hipótesis nula de normalidad de los valores residuales[17].

17 Sin embargo, en este ejemplo concreto con tan pocos datos ($n$ = 10) debe darse poco valor a los resultados de las pruebas de normalidad, ya que es casi imposible que resulte significativo cuando $n$ es tan pequeño.

```
. swilk res

                    Shapiro-Wilk W test for normal data

    Variable |   Obs        W          V          z       Prob>z
-------------+--------------------------------------------------
         res |    10    0.96630    0.519     -1.056    0.85459

. sfrancia res

                    Shapiro-Francia W' test for normal data

    Variable |   Obs        W'         V'         z       Prob>z
-------------+--------------------------------------------------
         res |    10    0.97448    0.427     -1.301    0.90344

. sktest res

                    Skewness/Kurtosis tests for Normality
                                                      ------ joint ------
    Variable |   Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)  Prob>chi2
-------------+-------------------------------------------------------------
         res |    10         0.7583         0.6069         0.35     0.8391
```

Los valores de probabilidad son $p = 0{,}855$, $p = 0{,}9034$ y $p = 0{,}839$ para las pruebas de Shapiro-Wilk, Shapiro-Francia y de asimetría y curtosis, respectivamente

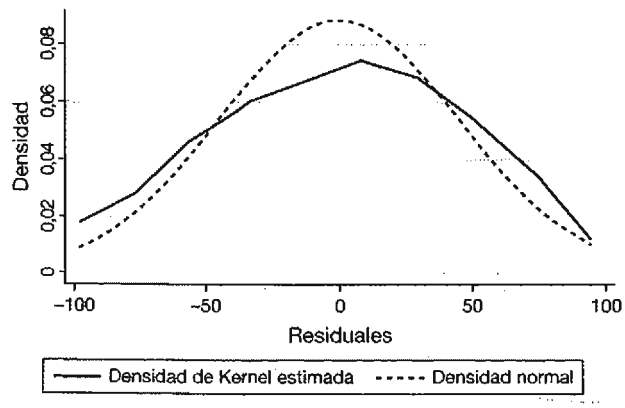
Los residuales pueden representarse a través de los gráficos *Q-Q* y *P-P* en el menú:
**Statistics → Summaries, tables, and tests → Distributional plots and tests →**
Para crear un gráfico *Q-Q*: **Normal quantile plot**
Para un gráfico *P-P*: **Normal probability plot, standardized**
Estos menús corresponden a las instrucciones:

```
qnorm res
pnorm res
```

Existe un interesante procedimiento no paramétrico en STATA, que es la denominada *función de densidad de Kernel*. Esta técnica, que utiliza procedimientos de «suavizado» o «alisamiento» de la distribución, constituye un afinamiento de los clásicos histogramas empleados para la visualización de la distribución de los residuales de un modelo. Puede obtenerse a través del menú:
**Statistics → Nonparametric analysis → Kernel density estimation**
o de la instrucción:

```
kdensity res, normal
```

(Al incluir la subinstrucción **normal,** a la gráfica de densidad de Kernel, se superpone una nueva gráfica que presenta la distribución normal esperada de los valores residuales.)



***Interpretación:*** *la gráfica presenta la distribución teórica normal de los valores residuales del modelo y la función de densidad de probabilidad estimada según este procedimiento no paramétrico. Pueden observarse dos distribuciones similares con media de 0 y desviación típica de 1.*

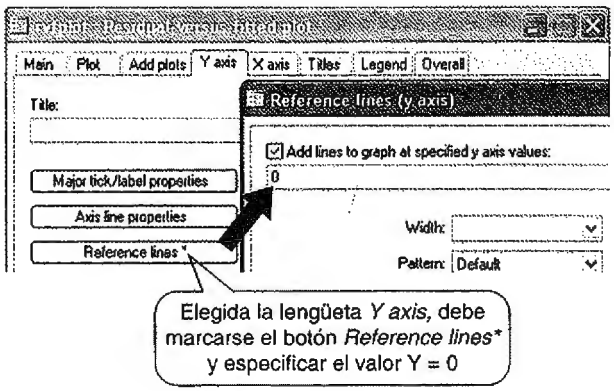
### 10.9.5. Representaciones gráficas

Además de las gráficas *Q-Q* y *P-P* para comprobar la normalidad de los residuales y, por tanto, la idoneidad del modelo de regresión, STATA realiza otras representaciones gráficas muy útiles dirigidas a determinar la homogeneidad de varianzas[18].

STATA crea una gráfica en la que se representan los valores residuales en el eje de ordenadas frente a los valores predichos de Y según el modelo de regresión en el de abscisas *(Residual vs. fitted plot)*. Para que exista homogeneidad de varianzas, la distribución de puntos por encima y por debajo de la línea Y = 0 debe ser similar sin que exista un patrón determinado. La instrucción del menú empleada para la obtención de esta gráfica es:

**Statistics → Linear models and related → Regression diagnostics → Residual-versus-fitted plot**

Puede solicitársele a STATA que trace una línea recta sobre el valor Y = 0 para una mejor visualización e interpretación de la gráfica de dispersión. El siguiente cuadro de diálogo muestra el procedimiento que se aplica.




```
rvfplot, yline(0)
```

(La subinstrucción **yline(0)** permite trazar una línea recta sobre el valor Y = 0.)

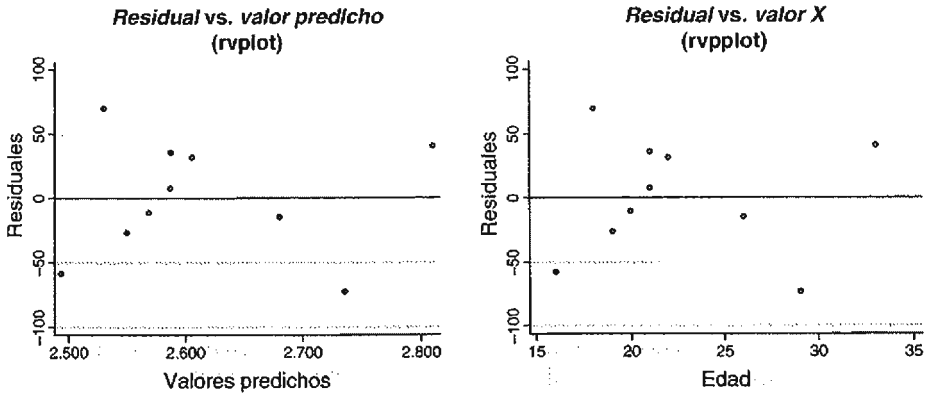
STATA también realiza un diagrama de dispersión entre los valores residuales del modelo (eje de ordenadas) y la variable independiente X (eje de abscisas) *(Residual vs. predicted plot)*. La interpretación es similar al caso anterior.

**Statistics → Linear models and related → Regression diagnostics → Residual-versus-predictor plot**

```
rvpplot x, yline(0)
```

(En este caso hay que especificar el nombre de la variable independiente X, que será representada en el eje de abscisas. En el ejemplo del capítulo: **rvpplot *edad***).

18 STATA utiliza de igual manera la instrucción **estat hettest, normal** o su forma reducida **hettest** para valorar la heterogeneidad a través del test de Breusch-Pagan/Cook-Weisberg. La hipótesis nula es la homogeneidad de varianzas, por lo que un resultado significativo indicará heterogeneidad.




***Interpretación:*** *la visualización de las gráficas* ***rvfplot*** *y* ***rvpplot*** *no parece hacer sospechar la presencia de heteroscedasticidad. La distribución de los puntos a lo largo de la línea de referencia (Y = 0) no sigue un patrón concreto.*

Si se desea representar los valores residuales tipificados en vez de los residuales, se debe solicitar un gráfico de dispersión a STATA, empleando las variables creadas y guardadas según el apartado 10.9.4.

En el ejemplo del peso del recién nacido según la edad de la madre:

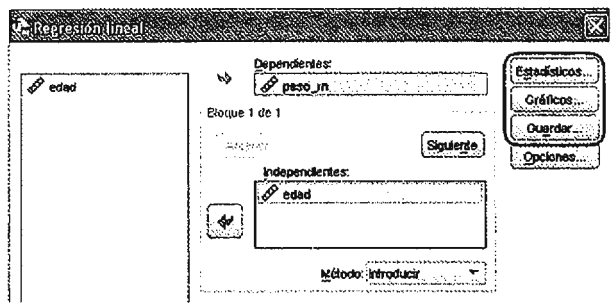
```
sc zres ypred, yli (0)
sc zres edad, yli (0)
```

## 10.10. REGRESIÓN LINEAL CON OTROS PROGRAMAS INFORMÁTICOS

### 10.10.1. Regresión lineal con SPSS

El menú de SPSS para realizar un modelo de regresión lineal es el siguiente:

**Analizar → Regresión → Lineales**[19]

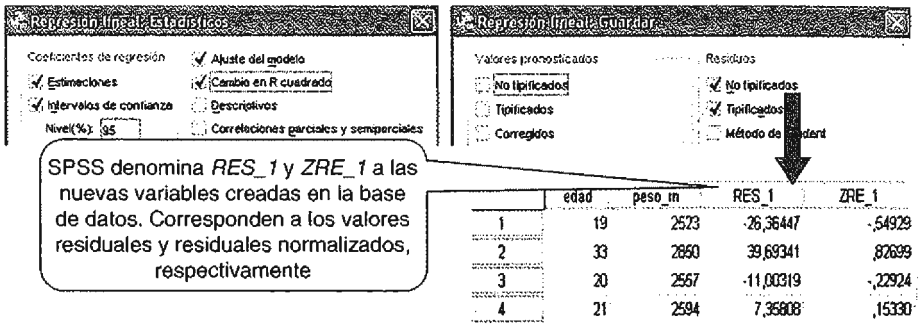



Deben destacarse las especificaciones de tres de los botones del cuadro de diálogo de SPSS.

A través del botón ESTADÍSTICOS, en la opción *Coeficientes de regresión* pueden obtenerse diferentes indicadores, como los propios de la regresión *(Estimaciones, Ajuste de modelo),* los intervalos de confianza de los coeficientes de regresión *(Intervalo de confianza)* o los coeficientes de determinación ($R^2$) *(Cambio en R cuadrado).*

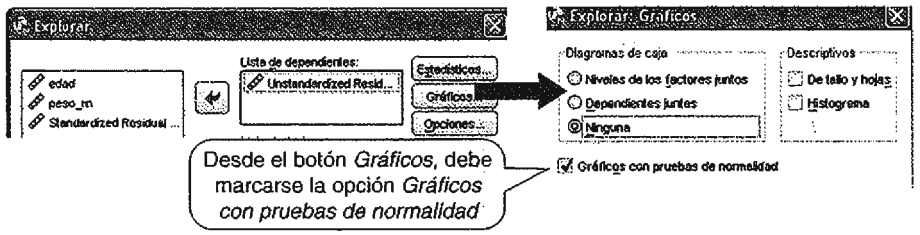
La creación de residuales puede realizarse con el botón GUARDAR. Pueden marcarse las opciones *Residuos No tipificados* o *Tipificados,* según cómo se desee la normalización de los valores residuales.

19 Lineal... en versiones anteriores a SPSS 19.0.

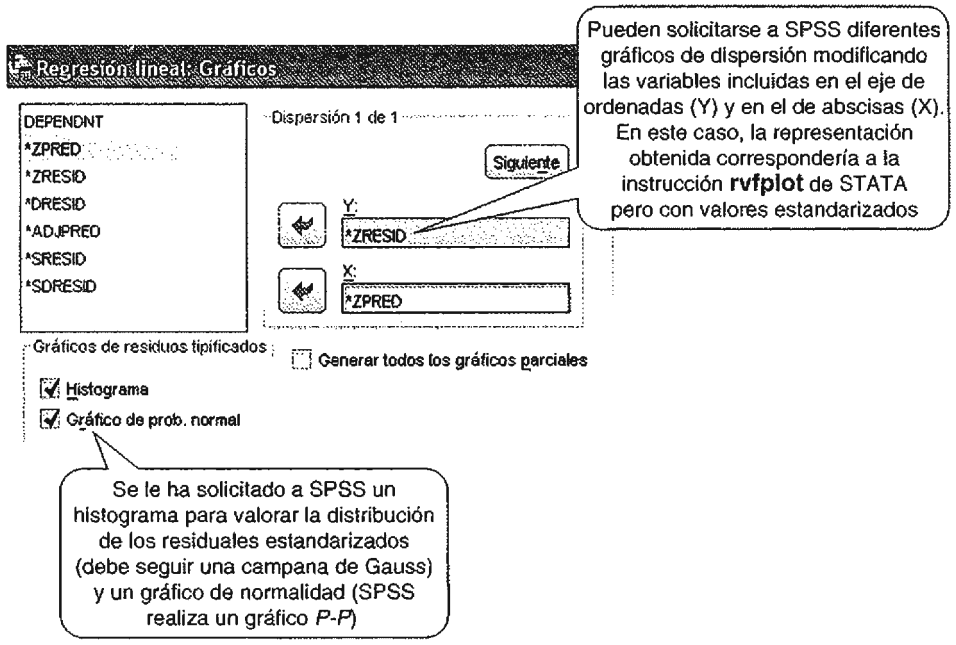



Para comprobar si los residuales del modelo siguen una distribución normal, pueden emplearse pruebas de normalidad. SPSS calcula la prueba de normalidad de Kolmogorov-Smirnov y la prueba de Shapiro-Wilk. La instrucción es[20]:

**Analizar → Estadísticos descriptivos → Explorar**




El botón GRÁFICOS permite la creación de gráficas de normalidad *(P-P)*, histogramas y otras representaciones como gráficos de dispersión, que permiten comprobar gráficamente la adecuación del modelo de regresión y el cumplimiento de la homoscedasticidad.



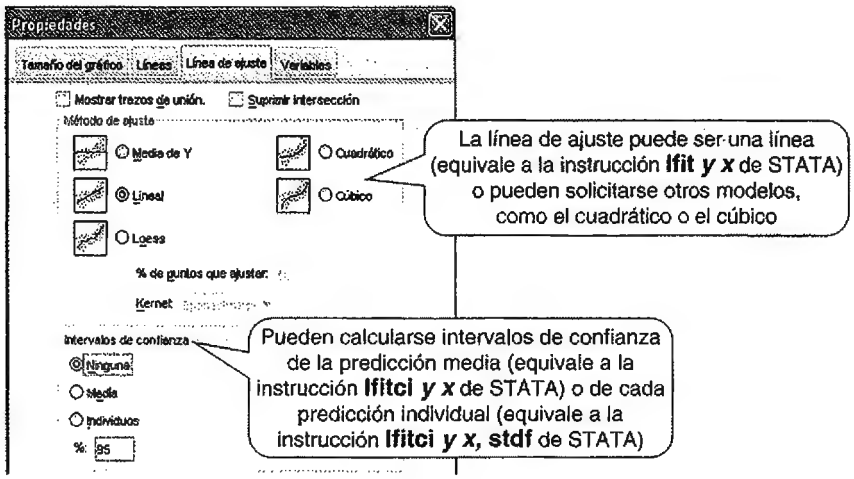

20 Con esta instrucción, SPSS también realiza un gráfico *Q-Q*.

SPSS permite obtener también gráficos de dispersión entre las variables X e Y y ajustar líneas de predicción, así como calcular los intervalos de confianza de la predicción media y de los valores individuales.

Para ello debe acudirse al menú:

**Gráficos → Cuadros de diálogo antiguos → Dispersión/Puntos...**[21]


Para ajustar una línea de predicción o representar los intervalos de confianza, debe editarse la gráfica creada (a través de un doble *clic* con el ratón) y seguir la instrucción del menú **Editor de gráficos: Elementos → Línea de ajuste total**




## 10.10.2. Regresión lineal con Excel

Pueden descargarse todos los cálculos realizados con el programa Excel para llevar a cabo un análisis de regresión con la variable edad y peso del recién nacido a través de: http://www.unav.es/departamento/preventiva/recursos_bioestadistica.

21 Las últimas versiones de SPSS incorporan cambios en los menús que dificultan la creación de gráficas al investigador, pero permiten emplear instrucciones de versiones anteriores mucho más sencillas.

### 10.10.3. Regresión lineal con R/Splus

Es posible usar los mismos datos antes presentados en Excel (incluidos los nombres EDAD y PESO_RN como cabecera o *header* de las dos columnas), pero se guardarán como archivo de texto delimitado por tabulaciones (puede elegirse, por ejemplo, el nombre dr.txt para el archivo). Una vez archivados, se leen desde R/Splus con la siguiente orden, que crea una base de datos, ya en formato R/Splus, que se llama simplemente «d»:

```
> d< -read.table("c://dr.txt", header=T)
> d
```

| | EDAD | PESO_RN |
|---|---|---|
| 1 | 19 | 2523 |
| 2 | 33 | 2850 |
| 3 | 20 | 2557 |
| 4 | 21 | 2594 |
| 5 | 18 | 2600 |
| 6 | 21 | 2622 |
| 7 | 22 | 2637 |
| 8 | 16 | 2435 |
| 9 | 29 | 2663 |
| 10 | 26 | 2665 |

Para solicitar la regresión lineal, se indicará al programa que use la base de datos «d» (orden *attach*) y después se escribirá **lm(PESO_RN~EDAD)**. La expresión «lm» significa *linear model*. El signo ~ se obtiene pulsando simultáneamente las teclas AltGr y 4, y, después, la barra espaciadora. Es importante no olvidar que R/Splus considera distintas las letras mayúsculas y minúsculas (es *«case sensitive»*). Aquí los nombres de las variables se habían escrito en mayúsculas en Excel y R/Splus no los reconocería si se escribiesen en minúsculas. Es preferible crear un archivo temporal en el que se guarden todos los resultados del modelo lineal. En nuestro ejemplo llamaremos a este archivo «temp». La secuencia de órdenes sería entonces:

```
> attach(d)
> temp<-lm(PESO_RN~EDAD)
```

Se ha creado ya un fichero temporal (de nombre «temp») en el que se almacena la información del análisis de regresión.

Se pide el contenido de «temp» con dos opciones:

```
> coef(temp)
(Intercept)        EDAD
2195.22874   18.63872
> summary(temp)

Call:
lm(formula = PESO_RN ~ EDAD)

Residuals:
```

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -72.752 | -23.482 | -1.823 | 34.448 | 69.274 |

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2195.229 | 69.901 | 31.405 | 1.15e-09 *** |
| EDAD | 18.639 | 3.033 | 6.146 | 0.000275 *** |

```
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
Residual standard error: 48 on 8 degrees of freedom
Multiple R-Squared: 0.8252, Adjusted Rsquared: 0.8034
F-statistic: 37.77 on 1 and 8 DF,  p-value: 0.0002752
```

Pueden identificarse de nuevo las diversas cantidades calculadas. R/Splus siempre añade una pequeña descriptiva de los residuales. Las representaciones gráficas se pueden pedir del siguiente modo:

```
> attach(d)
> plot(EDAD, PESO_RN)
> abline(lsfit(EDAD,PESO_RN)$coef)
```

## 10.11. RELACIÓN ENTRE REGRESIÓN, ANOVA Y *T* DE STUDENT

La relación entre la regresión, el análisis de la varianza y la *t* de Student se mostrará a través de un ejemplo.

Tras un programa educativo se comparó la mejoría en conocimientos *(nota)* en los tres grupos asignados (*grupo;* 1 = A (grupal); 2 = B (individual); 3 = C (grupal + ind.)), según muestra la tabla 10.10. ¿Hay diferencias entre los grupos?

*Tabla 10.10 Valoración de los conocimientos después de un programa educativo (nota sobre 10)*

| | GRUPO A: EDUCACIÓN GRUPAL | GRUPO B: EDUCACIÓN INDIVIDUAL | GRUPO C: GRUPAL + INDIVIDUAL |
|---|---|---|---|
| | 0 | 0 | 4 |
| | 1 | 2 | 4 |
| | 1 | 3 | 4 |
| | 1 | 3 | 5 |
| | 2 | 4 | 6 |
| Medias = | 1,0 | 2,4 | 4,6 |
| $s$ = | 0,71 | 1,52 | 0,89 |
| $n$ = | 5 | 5 | 5 |
| | $s_{TOTAL}$ = 1,839 | | |

Podría pensarse que este problema solo se puede resolver con un ANOVA de un factor. Sin embargo, a continuación se demuestra que el ANOVA de una vía es solo un caso particular de la regresión. Para resolverlo por ANOVA se obtendrían las siguientes sumas de cuadrados:

$$\text{SC resid.} = 4(0{,}71^2) + 4(1{,}52^2) + 4(0{,}89^2) = 14{,}4$$
$$\text{SC total} = 14(1{,}839^2) = 47{,}3$$

Puede obtenerse el listado a través del menú:
**Statistics → Linear models and related → ANOVA/MANOVA → One-way ANOVA**

```
. oneway nota grupo

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      32.9333333      2   16.4666667      13.72     0.0008
 Within groups          14.4      12          1.2
------------------------------------------------------------------------
    Total           47.3333333     14   3.38095238

Bartlett's test for equal variances:  chi2(2) =   2.2692  Prob>chi2 = 0.322
```

A idéntica conclusión se llega mediante una regresión. Se comienza por decidir qué «suelo» o categoría se elegirá para situarla como referencia y comparar frente a ella las otras dos categorías o grupos. Por ejemplo, parece lógico que el grupo de referencia o «suelo» sea el grupo A (educación grupal) y los otros dos se compararán con respecto a él. Seguidamente se crean dos variables para los grupos B y C, que se llaman indicadoras o *dummy* (*dummy_B* y *dummy_C*). Una variable *dummy* siempre es dicotómica y toma los códigos 0 y 1. Se caracteriza porque su valor es 0 para todas las categorías salvo la propia. La tabla 10.11 recoge las características de las variables *dummy* creadas en el ejemplo de la nota:

*Tabla 10.11* ***Dos variables dummy sustituyen a una variable con tres categorías***

| | VARIABLE ORIGINAL *(GRUPO)* | | |
|---|---|---|---|
| | A = 1 | B = 2 | C = 3 |
| *dummy_B* | 0 | 1 | 0 |
| *dummy_C* | 0 | 0 | 1 |

STATA puede crear las dos variables *dummy* a partir de las siguientes instrucciones:

```
generate dummy_B= grupo==2
generate dummy_C= grupo==3
```

A continuación se introducen ambas variables *dummy* como independientes en el análisis de regresión lineal, de tal manera que se obtendrá el siguiente modelo:

$$Nota = a + b_1 * dummy\_B + b_2 * dummy\_C$$

Este análisis realizado en STATA presenta el siguiente aspecto:

```
. regress  nota dummy_B dummy_C

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =   13.72
       Model |  32.9333333     2  16.4666667            Prob > F      =  0.0008
    Residual |      14.4      12        1.2            R-squared     =  0.6958
-------------+------------------------------           Adj R-squared =  0.6451
       Total |  47.3333333     14 3.38095238            Root MSE      =  1.0954

------------------------------------------------------------------------------
        nota |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dummy_B |         1.4   .6928203     2.02   0.066    -.1095258    2.909526
     dummy_C |         3.6   .6928203     5.20   0.000     2.090474    5.109526
       _cons |           1   .4898979     2.04   0.064    -.0673959    2.067396
------------------------------------------------------------------------------
```

***Interpretación:*** *puede observarse que el programa devuelve un listado de salida del ANOVA de regresión que coincide exactamente con el listado obtenido al llevar a cabo un análisis de ANOVA de un factor. Además, se acompaña de un nuevo listado que muestra los coeficientes del modelo de regresión. Con estos coeficientes se pueden estimar los valores predichos para cada uno de los tres grupos. De aquí se deducen fácilmente las diferencias de los grupos B y C con respecto al A, considerado como referencia.*

*Se parte del modelo general:*

$$Nota = a + b_1 * dummy\_B + b_2 * dummy\_C$$

*Se sustituyen los coeficientes y se obtienen tres ecuaciones, una por grupo. Para el grupo A, las dos variables dummy valen 0, por lo que la media de la nota será la constante; para el B, la media será la constante más el coeficiente de la dummy_B, y para el C, la media de la nota será la constante más el coeficiente de la dummy_C.*

***Grupo A:*** $Nota = a + b_1*dummy\_B + b_2*dummy\_C = 1 + (1,4*0) + (3,6*0) = 1,000$

***Grupo B:*** $Nota = a + b_1*dummy\_B + b_2*dummy\_C = 1 + (1,4*1) + (3,6*0) = 2,4$

***Grupo C:*** $Nota = a + b_1*dummy\_B + b_2*dummy\_C = 1 + (1,4*0) + (3,6*1) = 4,6$

*Como puede verse, las predicciones coinciden con las medias de la tabla 10.10. Los coeficientes de las dos variables dummy (+1,4 para B y +3,6 para C) constituyen, por tanto, una estimación de la diferencia de medias entre el grupo B y el A y entre el grupo C y el A respectivamente.*

$$\bar{y}_B - \bar{y}_A = +1,4 \rightarrow \textit{coeficiente de la variable dummy\_B}$$
$$\bar{y}_C - \bar{y}_A = +3,6 \rightarrow \textit{coeficiente de la variable dummy\_C}$$

*Los dos valores p que acompañan a cada uno de estos dos coeficientes en la tabla de la regresión (p = 0,066 y p < 0,001) valoran si la diferencia de medias es estadísticamente significativa para la comparación entre el grupo A y B y para la comparación entre el A y C, respectivamente. Solo la segunda de estas comparaciones ($\bar{y}_C - \bar{y}_A = +3,6$) es estadísticamente significativa. Si se quisiese hacer un test de hipótesis para comparar el grupo B con el C, se deberían usar otras variables dummy (dummy_A y dummy_C, dejando B como «suelo» o categoría de referencia).*

Con estos procedimientos se han comparado las medias entre dos grupos. Así se actuaba con la *t* de Student (y con los contrastes tras el ANOVA), con la diferencia de que, en este ejemplo, se usa la regresión. Como se puede apreciar es equivalente, ya que la *t* de Student es solo un *caso particular* de la regresión lineal. Los resultados de una *t* de Student coincidirán exactamente con los del listado de coeficientes de la regresión. Como única precaución debe usarse la raíz cuadrada de la varianza residual ($\sqrt{1,2} = 1,095$) para calcular el error estándar y tener en cuenta que los grados de libertad son los residuales (*gl* = 12). Así se obtiene una diferencia de medias = +3,6 y un error estándar = 0,693:

$$t_{gl=12} = \frac{\bar{y}_C - \bar{y}_A}{s_{resid}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{4,6-1}{1,095\sqrt{\frac{1}{5}+\frac{1}{5}}} = \frac{+3,6}{0,693} = 5,196$$

El resultado es idéntico al obtenido en la regresión. Puede llamar la atención que se esté utilizando una variable categórica (grupo = A, B o C) como independiente en la regresión, que es un procedimiento pensado para variables cuantitativas. Es posible, y, como se ha visto en el ejemplo, los resultados que produce son válidos, tienen pleno sentido y son aprovechables para sustituir a los del ANOVA y la *t* de Student.

## 10.12. USO DE LA REGRESIÓN PARA SUSTITUIR AL ANOVA FACTORIAL

En la tabla 10.12 se recoge una extensión del ejemplo presentado en la tabla 10.10. Ahora se ha tenido en cuenta si las personas que recibieron el programa de aprendizaje fueron sometidas a pequeños exámenes sorpresa con *feedback* rápido (variable *exam*; 0 = no; 1 = sí).

Si con estos datos se lleva a cabo un ANOVA factorial (de dos factores), se obtendrá:

Puede obtenerse el listado a través del menú: **Statistics → Linear models and related → ANOVA/MANOVA → Analysis of variance and covariance**

```
. anova  nota grupo exam grupo#exam
```

```
                  Number of obs =        30     R-squared     =  0.5835
                  Root MSE      =  2.22111     Adj R-squared =  0.4968
```

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 165.9 | 5 | 33.18 | 6.73 | 0.0005 |
| grupo | 37.8 | 2 | 18.9 | 3.83 | 0.0360 |
| exam | 124.033333 | 1 | 124.033333 | 25.14 | 0.0000 |
| grupo#exam | 4.06666667 | 2 | 2.03333333 | 0.41 | 0.6668 |
| Residual | 118.4 | 24 | 4.93333333 | | |
| Total | 284.3 | 29 | 9.80344828 | | |

*Tabla 10.12 Conocimientos (nota) según método educativo y exámenes*

| MÉTODO A: EDUCACIÓN GRUPAL | MÉTODO B: EDUCACIÓN INDIVIDUAL | MÉTODO C: GRUPAL + INDIVIDUAL |
|---|---|---|
| **NO sometidos a exámenes sorpresa** (*exam* = 0) | | |
| 0 | 0 | 4 |
| 1 | 2 | 4 |
| 1 | 3 | 4 |
| 1 | 3 | 5 |
| 2 | 4 | 6 |
| Medias = **1,0** | **2,4** | **4,6** |
| *s* = **0,71** | **1,52** | **0,89** |
| *n* = **5** | **5** | **5** |
| **SÍ sometidos a exámenes sorpresa** (*exam* = 1) | | |
| 2 | 1 | 4 |
| 6 | 5 | 8 |
| 6 | 7 | 8 |
| 6 | 9 | 9 |
| 10 | 10 | 10 |
| Medias = **6,0** | **6,4** | **7,8** |
| *s* = **2,83** | **3,58** | **2,28** |
| *n* = **5** | **5** | **5** |

***Interpretación:*** *gracias a la instrucción* ***anova,*** *STATA lleva a cabo un ANOVA de dos factores. El listado obtenido muestra los valores de las sumas de cuadrados, grados de libertad y varianzas entre grupos (model); el residual; cada una de las variables independientes (grupo y exam en el ejemplo) y la variable interacción (creada a través de la instrucción grupo#exam). Puede verse que no hay interacción entre ambos factores (grupo y examen), ya que el valor p asociado a la variable interacción vale 0,6668. Sin embargo, cada uno de estos factores resulta significativo (p para grupo = 0,036 y p para exam ≤ 0,001). Se concluirá que hay diferencias entre los métodos y que, al añadir los exámenes sorpresa frecuentes, los resultados mejoran significativamente.*

Se logra incluso más con un análisis de regresión lineal mediante el siguiente modelo:

$$Nota = a + b_1{*}dummy\_B + b_2{*}dummy\_C + b_3{*}exam + b_4{*}(exam{*}dummy\_B) + b_5{*}(exam{*}dummy\_C)$$

Este modelo de regresión lineal ofrece la ventaja con respecto al factorial de que los coeficientes son interpretables como diferencias de medias y se valoran con más detalle las posibles interacciones (mediante términos de producto), no solo desde el punto de vista de su significación estadística, sino de su magnitud. Se han introducido dos términos de producto (*exam*dummy_B* y *exam*dummy_C*) que valoran la interacción. Dicha interacción podría formularse del modo siguiente: ¿al hacer exámenes sorpresa con *feedback* rápido aumentan las diferencias entre los métodos? O bien: ¿son mayores las diferencias entre hacer exámenes sorpresa con *feedback* rápido y no hacerlos, según cuál sea el método docente? Como hay dos métodos docentes (B y C) que se comparan con un mismo «suelo» o categoría de referencia (A), se requerirá valorar dos interacciones. Para obtener términos de producto en STATA, tras haber creado las variables *dummy,* se puede usar la instrucción:

```
g ex_dumB = exam* dummy_B
g ex_dumC = exam* dummy_C
```

Una vez creadas estas variables, al aplicar el programa de regresión se obtiene el siguiente resultado:

```
. regress  nota  dummy_B dummy_C exam ex_dumB ex_dumC

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  5,    24) =    6.73
       Model |       165.9     5         33.18          Prob > F      =  0.0005
    Residual |       118.4    24  4.93333333           R-squared     =  0.5835
-------------+------------------------------           Adj R-squared =  0.4968
       Total |       284.3    29  9.80344828           Root MSE      =  2.2211

------------------------------------------------------------------------------
        nota |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     dummy_B |         1.4   1.404754     1.00   0.329    -1.499269    4.299269
     dummy_C |         3.6   1.404754     2.56   0.017     .7007306    6.499269
        exam |           5   1.404754     3.56   0.002     2.100731    7.899269
     ex_dumB |          -1   1.986622    -0.50   0.619    -5.100186    3.100186
     ex_dumC |        -1.8   1.986622    -0.91   0.374    -5.900186    2.300186
       _cons |           1    .993311     1.01   0.324    -1.050093    3.050093
------------------------------------------------------------------------------
```

***Interpretación:*** *puede apreciarse que la suma de cuadrados de regresión (165,9) y la varianza de regresión (33,18) corresponden a la suma de cuadrados y varianza entre grupos (model) en el ANOVA de dos factores. Lo mismo se produce para la fuente de variabilidad residual con la suma de cuadrados (118,4) y la varianza (4,933). Al aplicar los coeficientes se obtendrán las medias:*

- *Sin exámenes sorpresa:*
  - ***Método A:*** *Nota = a = 1.*
  - ***Método B:*** *Nota = a + $b_1$*dummy_B = 1 + 1,4 = 2,4.*
  - ***Método C:*** *Nota = a + $b_2$*dummy_C = 1 + 3,6 = 3,6.*
- *Con exámenes sorpresa:*
  - ***Método A:*** *Nota = a + $b_3$*exam = 1 + 5 = 6.*
  - ***Método B:*** *Nota = a+ $b_1$*dummy_B+ $b_3$*exam+ $b_4$*ex_dumB = 1 + 1,4 + 5 − 1 = 6,4.*
  - ***Método C:*** *Nota = a+ $b_2$*dummy_C+ $b_3$*exam+ $b_5$*ex_dumC = 1 + 3,6 + 5 − 1,8 = 7,8.*

De nuevo, el modelo de regresión sirve para estimar la media de cada posible grupo construido al combinar las categorías de los factores. Lo interesante es que la regresión proporciona valores *p* para cada una de las cinco comparaciones específicas (cinco contrastes de hipótesis que consumen los cinco grados de libertad). En este ejemplo, las únicas comparaciones que son estadísticamente significativas son las referentes al efecto del examen ($p = 0{,}002$, que mejora en 5 puntos el rendimiento) y el efecto de añadir la educación individual a la grupal (método C frente a A, $p = 0{,}017$, con una magnitud de efecto de 3,6 puntos).

Otra gran ventaja de resolver estos problemas por métodos de regresión en vez de ANOVA es que se pueden calcular fácilmente intervalos de confianza para las diferencias de medias entre grupos. Se usará la *t* de Student con los grados de libertad residuales (gl = 24 en el ejemplo) ($t_{0{,}025{,}24} = 2{,}0639$) para multiplicar el error estándar. Este producto se suma y se resta al coeficiente y así se obtiene un intervalo de confianza para la diferencia. Estos intervalos de confianza se obtienen directamente en STATA. Por ejemplo, para comparar la diferencia de medias entre quienes se sometieron a exámenes y quienes no se sometieron (dentro del grupo A, que siguió un método de enseñanza individual), se hallaría:

$$\text{IC}95\% = 5 \pm (2{,}0639 \times 1{,}405) = +2{,}101 \text{ a } +7{,}899$$

Esta diferencia (*dif* = 5; IC 95%: 2,1 a 7,9) entre los que tuvieron exámenes sorpresa y quienes no los tuvieron solo es aplicable al grupo A. Si se desea obtener la magnitud del efecto de tener el

examen para todos los grupos, se deben eliminar los términos de interacción del modelo. Es lógico, puesto que ninguno de ellos resultó significativo y, por tanto, se puede mantener la hipótesis nula de que el efecto de los exámenes es el mismo sea cual sea el método de enseñanza. Esto equivaldría a hacer un ANOVA factorial personalizado sin incluir el término de interacción y simplificar el modelo, que ahora queda así:

$$Nota = a + b_1 * dummy\_B + b_2 * dummy\_C + b_3 * exam$$

Al ajustarlo por regresión en STATA, se obtienen los siguientes coeficientes:

```
. regress nota dummy_B dummy_C
```

| Source | SS | df | MS | Number of obs = | 15 |
|---|---|---|---|---|---|
| Model | 32.9333333 | 2 | 16.4666667 | F( 2, 12) = | 13.72 |
| Residual | 14.4 | 12 | 1.2 | Prob > F = | 0.0008 |
| | | | | R-squared = | 0.6958 |
| | | | | Adj R-squared = | 0.6451 |
| Total | 47.3333333 | 14 | 3.38095238 | Root MSE = | 1.0954 |

| nota | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| dummy_B | 1.4 | .6928203 | 2.02 | 0.066 | -.1095258 | 2.909526 |
| dummy_C | 3.6 | .6928203 | 5.20 | 0.000 | 2.090474 | 5.109526 |
| _cons | 1 | .4898979 | 2.04 | 0.064 | -.0673959 | 2.067396 |

La *t* de Student para *exam* ($t_{26}$ = 5,13) es exactamente la raíz cuadrada de la *F* que resultaría en un ANOVA de dos factores sin interacción ($F_{1,26}$ = 26,333).

```
. anova nota grupo exam
```

| | Number of obs = | 30 | R-squared = | 0.5692 |
|---|---|---|---|---|
| | Root MSE = | 2.17031 | Adj R-squared = | 0.5195 |

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 161.833333 | 3 | 53.9444444 | 11.45 | 0.0001 |
| grupo | 37.8 | 2 | 18.9 | 4.01 | 0.0303 |
| exam | 124.033333 | 1 | 124.033333 | 26.33 | 0.0000 |
| Residual | 122.466667 | 26 | 4.71025641 | | |
| Total | 284.3 | 29 | 9.80344828 | | |

El coeficiente para la variable *exam* ahora vale 4,067, que es la diferencia entre quienes tuvieron exámenes sorpresa y quienes no los tuvieron, pero promediada entre los tres métodos (A, B y C).

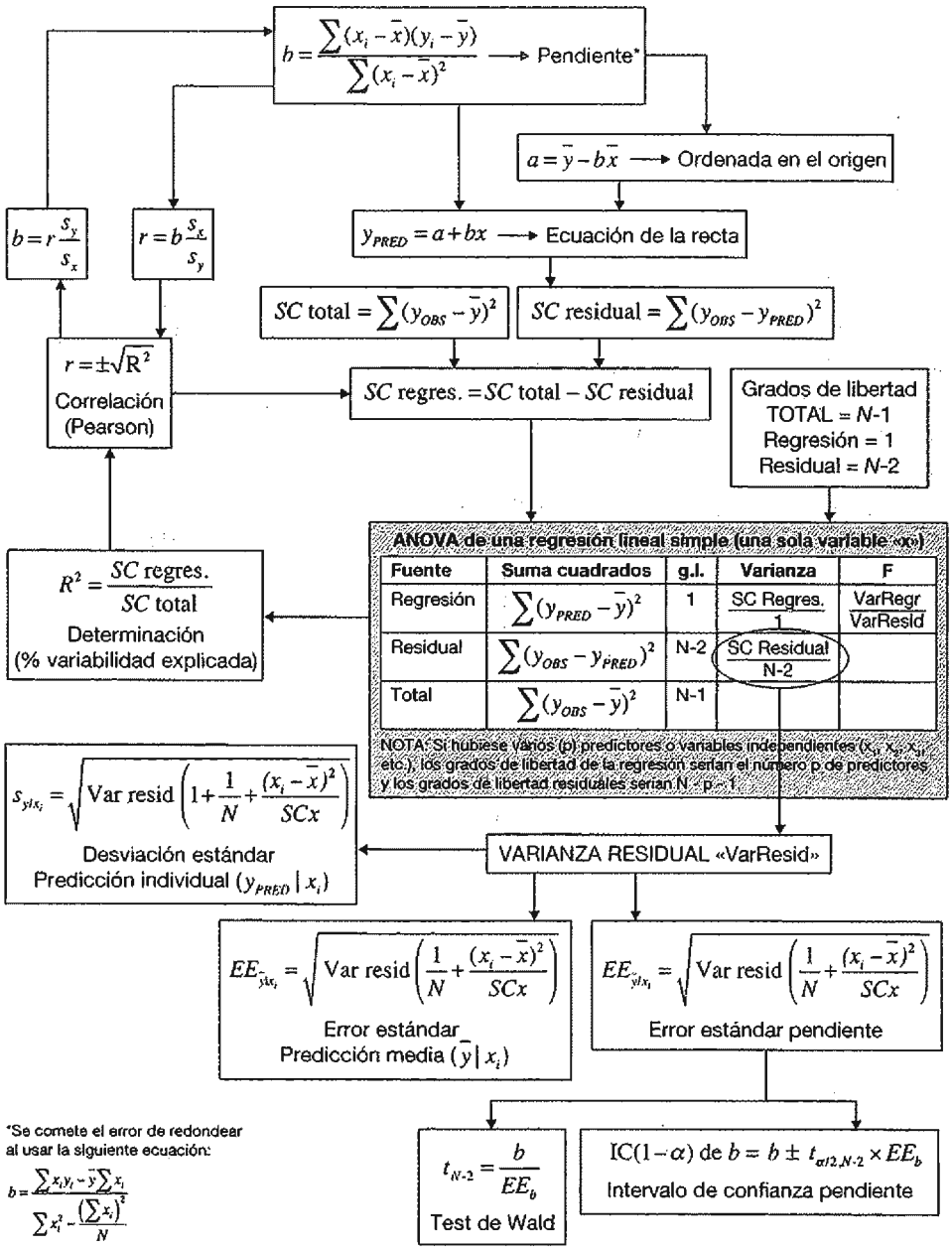
$$Dif_{\text{exam-no exam}} = \frac{(6-1)+(6{,}4-2{,}4)+(7{,}8-4{,}6)}{3} = 4{,}067$$

El intervalo de confianza al 95% para esta diferencia se hace teniendo en cuenta que ahora la varianza residual tiene 26 grados de libertad y, por tanto, $t_{0{,}025{,}26}$ = 2,0555:

$$IC\,95\% = 5 \pm (2{,}0555 \times 0{,}792) = +2{,}439\ \text{a}\ +5{,}695$$

Ahora el intervalo de confianza es más estrecho, pues se ha eliminado del error estándar el «ruido» que introducían los dos términos de interacción no significativos. Esta simplicidad siempre es preferible, porque incrementa la precisión. Este principio que defiende simplificar los modelos siempre que se pueda se suele denominar «parsimonia».

## 10.13. RESUMEN DE CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE




*Se comete el error de redondear al usar la siguiente ecuación:

$$b = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}$$

## 10.14. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| **Cálculo de coeficiente de correlación** | | |
| Pearson | correlate $v_1$ $v_2$<br>pwcorr $v_1$ $v_2$ | CORRELATIONS<br>/VARIABLES= $v_1$ $v_2$. |
| Spearman | spearman $v_1$ $v_2$ | NONPAR CORR<br>/VARIABLES= $v_1$ $v_2$. |
| **Cálculo de coeficiente de regresión** | regress $y$ $x$ | REGRESSION<br>/STATISTICS COEFF<br>/DEPENDENT $y$<br>/METHOD=ENTER $x$. |
| **Intervalos de confianza** | p.d. | /STATISTICS CI(95) |
| **Cálculo de coeficientes de determinación** | p.d. | /STATISTICS R |
| **Creación de:** | | |
| Valores predichos | predict *ypred*[1] | /SAVE PRED |
| Residuales | predict *res*[1], residuals | /SAVE RESID |
| Residuales estandarizados | predict *zres*[1], rstandard | /SAVE ZRESID |
| Error estándar de la predicción | predict *eep*[1], stdp | /SAVE SEPRED |
| Error estándar de los residuales | predict *eeres*[1], stdr | — |
| **Comprobación de normalidad de residuales (gráfica)** | | |
| *Q-Q* | qnorm *res* | EXAMINE<br>VARIABLES=*RES_1*[2]<br>/PLOT NPPLOT. |
| *P-P* | pnorm *res* | /RESIDUALS<br>NORMPROB(ZRESID)[3] |
| Densidad de Kernel con distribución normal | kdensity *res*, normal | ---------- |
| Histograma | histogram *res* | /RESIDUALS<br>HISTOGRAM(ZRESID)[3] |
| **Comprobación de normalidad de residuales (test)** | swilk *res*<br>sfrancia *res*<br>sktest *res* | EXAMINE<br>VARIABLES=*RES_1*<br>/PLOT NPPLOT.[4] |
| **Gráficos de dispersión** | | |
| Nube de puntos | twoway scatter $y$ $x$ | GRAPH<br>/SCATTERPLOT(BIVAR)=<br>$x$ WITH $y$. |
| Ajuste de línea de predicción | twoway lfit $y$ $x$ | Únicamente con las opciones de gráficas del menú |
| Intervalo de confianza de la predicción media | twoway lfitci $y$ $x$ | Únicamente con las opciones de gráficas del menú |
| Intervalo de confianza de las predicciones individuales | twoway lfitci $y$ $x$, stdf | Únicamente con las opciones de gráficas del menú |

| Procedimiento | STATA | SPSS |
|---|---|---|
| **Comprobación de homoscedasticidad de residuales:** | | |
| Residual-valor predicho | **rvfplot** | **GRAPH** |
| Residual-valor X | **rvpplot *x*** | **/SCATTERPLOT(BIVAR)= *RES_1* WITH *x*.** |

p.d., por defecto (el programa calcula el parámetro sin añadir instrucciones); *u*, variable cuantitativa; *x*, variable independiente; *y*, variable dependiente.

1 Nombre que el investigador decide dar a la nueva variable creada. En este ejemplo se han elegido los nombres: *ypred*, *res*, *zres*, *eep* y *eeres*.

2 Nombre que, por defecto, otorga SPSS a la variable residual creada por el programa. Se mantiene el mismo nombre en todo el ejemplo.

3 Subinstrucciones del menú de regresión que SPSS siempre realiza por defecto sobre variables estandarizadas (ZPRED: valores predichos estandarizados; ZRESID: valores residuales estandarizados). Podrían solicitarse estas gráficas al programa con valores no tipificados especificando RESID y PRED, respectivamente, en la instrucción.

4 Con esta instrucción, SPSS realiza el test de normalidad de Kolmogorov-Smirnov y el test de Shapiro-Wilk, y crea una gráfica *Q-Q*.

## REFERENCIAS


1. Motulsky H. Intuitive Biostatistics. New York: Oxford University Press; 1995.
2. Altman DG. Practical statistics for medical research. London: Chapman and Hall; 1991.
3. Greenland S, Morgenstern H. Confounding in health research. Annu Rev Public Health 2001;22:189-212.
4. De Irala J, Martínez-González MA, Guillén-Grima F. ¿Qué es un factor de confusión? Med Clin (Barc) 2001;117:377-85. Fe erratas: Med Clin (Barc) 2001; 117:775.
5. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. Am J Epidemiol 2002;155(2):176-84.
6. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel Ciencias Médicas; 2008.
7. Martínez-González MA, Martínez JA, Hu FB, Gibney MJ, Kearney J. Physical inactivity, sedentary lifestyle and obesity in the European Union. Int J Obes 1999;23(11):1192-201.
8. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346(8982):1085-7.
9. Shoukri MM. Measurement of agreement. En: Armitage P, Colton T, editors. Encyclopaedia of biostatistics. Chichester: John Wiley & Sons, 1999. p. 103-17.
10. Luiz RR, Leal-Costa AJ, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol 2003;56(10):963-7.
11. Llorca J, Delgado-Rodríguez M. Survival analytical techniques used to assess agreement of a quantitative variable. J Clin Epidemiol 2005;58(3):314-5.
12. Carrasco JL, Hernán MA, Martín-Hortelano C. 6.ª ed. El método estadístico en la investigación médica, Madrid: Ciencia 3, 1995.
13. Delgado-Rodríguez M, Llorca J. Bias. J Epidemiol Community Health 2004;58(8):635-41.

14. Rothman KJ. Causal inference. Chesnut Hill: Epidemiologic Resources; 1988.

15. Rothman KJ. Causes [reproducción del artículo clásico de 1976]. Am J Epidemiol 1995;141(2):90-5.

16. Pérez-Gaspar M, Gual P, De Irala-Estévez J, Martínez-González MA, Lahortiga F, Cervera S. Prevalencia de Trastornos de la Conducta Alimentaria (TCA) en las adolescentes navarras. Med Clin (Barc) 2000;114(13):481-6.

17. Healy MJR. Statistics from the inside. 15. Multiple regression (1). Arch Dis Child 1995;73(2):177-81.

18. Martínez-González MA, Corella D, Salas-Salvadó J, Ros E, Covas MI, Fiol M, for the PREDIMED Study Investigators. et al. Cohort Profile: design and methods of the PREDIMED study. Int J Epidemiol 2012;41:377-85.

19. Ríos M, García JM, Cubedo M, Pérez D. Análisis de series temporales en la epidemiología de la fiebre tifoidea en España. Med Clin (Barc) 1996;106(18):686-9.

20. Shibuya K, Inoue M, Lopez AD. Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. Int J Cancer 2005;117(3):476-85.

21. Kis M. Analysis of the time series for some causes of death. Stud Health Technol Inform 2002;90:439-43.

22. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-69.

23. Draper NR, Harry Smith H. Applied Regression Analysis. New York: Wiley; 1980.

24. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and other multivariable methods. 3rd ed. Boston: Duxbury Press; 1997.

25. Belsley DA, Kuh E, Welsh RE. Regression diagnostics: identifying influential data and sources of collinearity. New York: Wiley; 1980.

26. Belsley DA. Conditioning Diagnostics: Collinearity and Weak Data in Regression. New York: John Wiley & Sons; 1991.

27. Cook RD, Weisberg S. Residuals and influence in regression. New York: Chapman and Hall; 1982.

28. Godfrey K. Simple linear regression in medical research. En: Bailar JC III, Mosteller F, editors. Medical uses of statistics. 2nd ed. Boston: NEJM Books; 1992:201-32.

29. Sánchez-Cantalejo Ramírez E, Ocaña-Riola R. Actualizaciones en regresión: suavizando las relaciones. Gac Sanit 1997;11:24-32.

# INTRODUCCIÓN AL ANÁLISIS DE SUPERVIVENCIA

11

*E. Toledo, F. J. Basterra-Gortari, M. García-López, M. Á. Martínez-González*

## 11.1. INTRODUCCIÓN

Cuando interesa estudiar fenómenos como:

- el tiempo que tarda en producirse una defunción, o
- el lapso transcurrido hasta que se manifiesta un síntoma determinado, o
- el tiempo que transcurre para que se produzca la recidiva de una determinada enfermedad, o
- el tiempo que tarda en estropearse una prótesis,

hay que considerar el manejo de datos sobre el tiempo transcurrido hasta que se produce un evento *(time-to-event data)*, y deben aplicarse los métodos que se conocen generalmente como *análisis de supervivencia* (**1-8**). En estos casos, la variable de interés ya no es cuantitativa ni cualitativa, sino que toma la forma de *tiempo transcurrido hasta un suceso (time to event)*, lo que lleva a utilizar la combinación de dos elementos:

1. Si se produjo o no el desenlace (muerte, manifestación de un síntoma, recidiva, etc.).
2. Cuánto tiempo ha tardado en producirse ese desenlace o evento.

El primer componente es *dicotómico*; el segundo, *cuantitativo*.

Aunque se llame análisis de supervivencia, el desenlace analizado no tiene que ser la muerte. Aun así, debe tratarse de un acontecimiento que, como la muerte, solo pueda ocurrir una vez y que marque un punto de no retorno. Los efectos de esta índole suelen tener una característica que los hace inadecuados para otros análisis estadísticos: la existencia de información *truncada* o individuos *censurados (censored)* con tiempos de observación *incompletos*. Un sujeto ofrece una información truncada (está *censurado*) cuando para él termina el período de seguimiento por un motivo distinto a la ocurrencia del evento estudiado. Afortunadamente no todos mueren o desarrollan el evento durante un estudio. No se sabe entonces cuánto tardarían en desarrollarlo. Estos son los censurados. Quienes no mueren (o no desarrollan el evento en cuestión) durante el tiempo de observación serán censurados y, para ellos, se desconocerá el tiempo de supervivencia. Solo se sabe que superará al tiempo durante el cual fueron observados. También son censurados los sujetos que abandonan el estudio por su voluntad (abandonos, o pérdidas, *lost to follow-up*) o los que son retirados por los investigadores (retiradas, *withdrawals*).

Por tanto, los sujetos pueden terminar el estudio por una de las siguientes razones:

1. Se ha producido el evento.
2. Se ha terminado el estudio sin que presenten el evento.
3. Se retiran, abandonan el estudio, se pierden, … o presentan el evento por una causa competitiva (por ejemplo, si interesa estudiar como evento la mortalidad por cáncer, se considerará que una paciente que fallezca por enfermedad cardiovascular tuvo una causa competitiva de muerte).

Los individuos incluidos en los apartados 2), 3) están censurados y plantean un problema. Los métodos que se expondrán suponen que, si hubiesen seguido siendo observados, se habrían

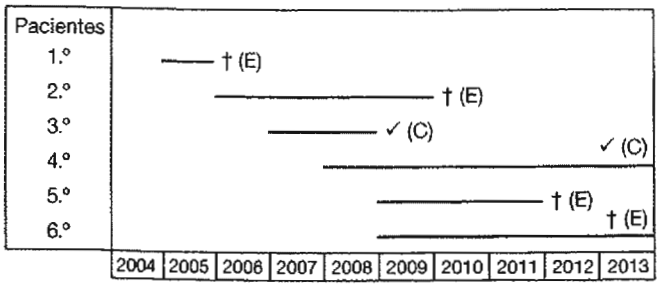



**Figura 11.1** Supervivencia de 6 pacientes objeto de seguimiento entre 2004 y 2013.

comportado del mismo modo que los que sí pudieron ser objeto de seguimiento hasta la ocurrencia del evento.

En ausencia de información censurada, es decir, si todos los sujetos fuesen seguidos completamente durante el mismo período de tiempo hasta que se produjera su muerte o el acontecimiento estudiado, se podrían usar otros métodos más sencillos. Sin embargo, los sujetos suelen ser objeto de seguimiento durante distintos períodos de tiempo. Además, no todos inician el estudio al mismo tiempo, sino que se van incorporando durante un período de meses o años. Un ejemplo de las distintas formas en que los sujetos pueden entrar, formar parte y salir del estudio es el que recoge la figura 11.1.

En la figura 11.1 se representa el tiempo de seguimiento de cada paciente mediante una línea. Se contemplan dos situaciones posibles: sujetos cuyo período de seguimiento acaba porque tienen el evento (E), que en este ejemplo sería su fallecimiento, o sujetos que estaban vivos cuando dejaron de ser observados (son retirados, abandonan el estudio, se pierde el seguimiento o están vivos al final del estudio), que son los censurados (C).

El paciente 1 empezó a ser estudiado a finales de 2004 y, tras un año en observación, falleció. El número 2 permaneció 4 años en observación (desde finales de 2005 a principios de 2010) antes de fallecer. El número 3 entró en el estudio a final de 2006 y abandonó el estudio estando vivo a final de 2008 (completó 2 años de seguimiento y luego se perdió). El paciente número 4 inició el estudio a finales de 2007 y tras 6 años de seguimiento, al término del estudio, seguía vivo. El quinto paciente falleció tras haber estado 3 años en el estudio y el último paciente falleció 5 años después de iniciarlo; su muerte coincidió con el final del estudio. El primer problema que se plantea es que cada sujeto entra en el estudio en una fecha de calendario distinta, lo cual se resuelve fácilmente, tal como se muestra en la figura 11.2.

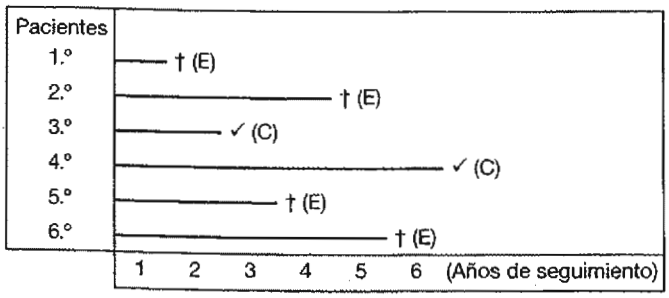



**Figura 11.2** Supervivencia de los 6 pacientes de la figura anterior, considerando solo la cantidad de tiempo durante el que han sido observados.

*Tabla 11.1 Datos para realizar un análisis de supervivencia*

| PACIENTE | AÑOS | MUERTE |
|---|---|---|
| 1.° | 1 | 1 |
| 2.° | 4 | 1 |
| 3.° | 2 | 0 |
| 4.° | 6 | 0 |
| 5.° | 3 | 1 |
| 6.° | 5 | 1 |

El cambio realizado consiste en considerar solo la cantidad de tiempo en que cada sujeto ha sido observado, lo que implica, en cierto modo, asumir que todos los participantes iniciaron el estudio en la misma fecha. Más en concreto, presupone asumir que se trata de pacientes *homogéneos,* es decir, que los criterios de entrada en el estudio fueron establecidos de una manera bien definida e igual para todos y que no se han producido cambios importantes en los métodos diagnósticos y terapéuticos, y, por tanto, tampoco en el pronóstico a lo largo del período de incorporación de pacientes al estudio (2004-2013). Si se hubiesen producido cambios en el modo de clasificar a los pacientes durante el período de estudio, se produciría el llamado fenómeno de Will Rogers[1].

Los datos para hacer un análisis de supervivencia se muestran en la tabla 11.1.

La variable ***MUERTE*** se ha codificado de la siguiente manera:

1. Fallecido.
0. Vivo.

La variable ***AÑOS***[2] recoge el tiempo transcurrido desde que cada paciente se incorporó al estudio hasta que murió. Si el paciente no ha fallecido, se indican los años totales de observación.

## 11.2. DESCRIPCIÓN DE LA SUPERVIVENCIA: MÉTODO DE KAPLAN-MEIER

Para estimar la probabilidad de supervivencia individual acumulada a lo largo del tiempo suele utilizarse el *método de Kaplan-Meier* (**9-12**). Se trata de un método no paramétrico, no presupone que los datos tengan una distribución particular. El único supuesto importante es que la *censura no sea informativa.* Lo más importante es que se asume que los sujetos censurados se habrían comportado del mismo modo que los que han sido objeto de seguimiento hasta el evento, si se

1 Este fenómeno suele ocurrir en estudios cuyo reclutamiento (admisión de nuevos participantes) se prolonga durante varios años coincidiendo con la modificación de los criterios diagnósticos de esa enfermedad o con cambios de sensibilidad de los aparatos diagnósticos. Podría ocurrir, por ejemplo, que al cabo de unos años se contara con mejores recursos para diagnosticar a un paciente. Esto haría que se identificasen ciertas lesiones que se habrían pasado por alto con los métodos habitualmente utilizados anteriormente (p. ej., ganglios afectados en oncología); así, los pacientes incorporados al estudio en los años más recientes (p. ej., después del 2005) terminarían por adscribirse a un estadio más avanzado de la enfermedad del que les correspondería si se les hubiese reclutado en fechas anteriores (p. ej., inicios de la década de 2000). Este concepto se conoce como «migración diagnóstica». Evidentemente, estos pacientes tienen una forma de enfermedad más grave que el conjunto del grupo en el que se les habría clasificado una década antes, aunque, por otra parte, son los de menor gravedad en el grupo en que se les sitúa ahora con los avances diagnósticos. Ocurre algo aparentemente paradójico: la mortalidad es menor en *ambos* grupos objeto de la migración diagnóstica: la mortalidad del estadio inferior disminuirá al haber menos pacientes graves y la del estadio superior también descenderá, porque ahora hay personas con una gravedad de enfermedad menor que la que habitualmente correspondía a ese grupo. Este fenómeno podría compararse con el símil de la altura: si la persona más alta de un grupo de gente baja pasa a pertenecer a otro grupo de personas con mayor altura, ambos grupos verán disminuir su altura media (**10**). Will Rogers afirmó que, cuando los oriundos de Oklahoma dejaron California, consiguieron que subiese la inteligencia media de ambos estados, lo cual no deja en buen lugar a la inteligencia media del estado de Oklahoma.

2 Muchos programas estadísticos no aceptan la letra ñ. Para las variables es preferible usar un nombre que no contenga la ñ. En cambio, en su mayoría sí suelen admitir el uso de la ñ para la etiqueta.

***Tabla 11.2 Reordenación de los datos de la tabla 11.1 Se han ordenado los tiempos (años) y se han marcado en negrita los sujetos censurados***

| AÑOS | PACIENTE | MUERTE |
|---|---|---|
| 1 | 1.° | 1 |
| 2 | 3.° | **0** |
| 3 | 5.° | 1 |
| 4 | 2.° | 1 |
| 5 | 6.° | 1 |
| 6 | 4.° | **0** |

hubiesen podido observar en tiempos completos. Esto supone admitir que es muy verosímil que no se trate de sujetos peculiares, sino que son representativos y no hay motivos para pensar que tengan mejor ni peor pronóstico que el resto de sujetos observados hasta ese momento. Este supuesto básico se denomina *censura no informativa,* ya que saber que un sujeto ha sido censurado no ofrece información adicional sobre su pronóstico. Desde el punto de vista práctico, se requiere suponer que quienes fueron censurados precozmente no son sujetos peculiares. En cambio, si el hecho de saber que un paciente se retira antes de tiempo (es censurado) indirectamente proporcionase información acerca de su pronóstico, se diría que la censura es *informativa.* En caso de censura informativa, surgirán dudas sobre la validez del procedimiento. No importa que existan muchos sujetos censurados, sino que la censura no esté relacionada con el pronóstico potencial, es decir, que no sea informativa.

Los datos del ejemplo antes comentado (6 pacientes) se repiten otra vez en la tabla 11.2, aunque se han resaltado en negrita los 2 pacientes censurados, que son aquellos de los que se desconoce el tiempo de supervivencia, ya que seguían vivos al final del seguimiento. Si se les excluyese y se utilizasen solo los casos de defunciones comprobadas (pacientes 1, 2, 5 y 6), podría calcularse fácilmente la supervivencia. No obstante, esta actuación sería *errónea,* ya que los participantes censurados aportan información valiosa.

Olvidando a los censurados, podría pensarse erróneamente que el valor de la supervivencia será:

| | |
|---|---|
| Al año: ha fallecido 1 y sobreviven 3 | Supervivencia = 3/4 = 0,75 |
| A los 2 años: ha fallecido 1 y sobreviven 3 | Supervivencia = 3/4 = 0,75 |
| A los 3 años: han fallecido 2 y sobreviven 2 | Supervivencia = 2/4 = 0,50 |
| A los 4 años: han fallecido 3 y sobrevive 1 | Supervivencia = 1/4 = 0,25 |
| A los 5 años: han fallecido todos | Supervivencia = 0 |

Esta aproximación lleva, por un lado, a desaprovechar la información que proporcionan los sujetos censurados (3.° y 4.°) y, por otro, a algo más importante: no es verdad, por ejemplo, que a los 5 años la supervivencia sea de 0, ya que el cuarto paciente ha sobrevivido más de 6 años. Lo correcto es aprovechar los datos censurados, como los de los pacientes 3 y 4, de los que se posee una información incompleta. Así trabaja el método de Kaplan-Meier.

Para obtener una descripción de la supervivencia por el método de Kaplan-Meier se debe disponer al menos de 2 columnas (dos variables por paciente) (v. tabla 11.1). La primera indicará el *tiempo* durante el cual se ha observado a cada paciente. La segunda señalará el *estado del paciente al final* del seguimiento. Se suele asignar un 1 a los que fallecieron (su tiempo de seguimiento equivale a su supervivencia) y un 0 a los que seguían vivos al final del seguimiento (censurados). Los datos se ordenarán según el tiempo de observación en orden ascendente.

Es posible así estimar la probabilidad de la supervivencia para un período dado. El *método de Kaplan-Meier* no supone que los datos tengan una distribución particular ni se basa en utilizar parámetros de resumen (media, desviación estándar, etc.). La supervivencia en el tiempo $t$ ($S_t$) se define como:

$$S_t = \prod \frac{s_i}{n_i}$$

donde la letra pi mayúscula (∏) es el «multiplicatorio», es decir, un símbolo análogo al sumatorio (Σ), que, en vez de expresar «sumar todo», quiere decir «multiplicar todo»; $s_i$ son los supervivientes en el tiempo $t_i$, y $n_i$ son los que están en riesgo de fallecer al inicio del tiempo $t_i$.

Este estimador de Kaplan-Meier expresa una *función* que variará a lo largo del tiempo, y no una única *cantidad*. El subíndice t indica que la supervivencia será distinta para uno u otro tiempo. En la tabla se representa, paso a paso, cómo se calcula el estimador de Kaplan-Meier. En cada período de tiempo se van multiplicando los cocientes $(s_i/n_i)$ por los de los tiempos previos. La supervivencia acumulada para cada tiempo corresponde a este producto.

## 11.3. PASOS PARA REALIZAR CURVAS DE SUPERVIVENCIA DE KAPLAN-MEIER

1. *Ordenar los datos* de menor a mayor según tiempo de supervivencia (o de observación), tal como se muestra en la tabla 11.2.
2. *Hacer una tabla de supervivencia.* Completar las tres primeras columnas de la tabla 11.3:
   a. La primera columna $(t_i)$ corresponde a los tiempos de observación (en el ejemplo, medidos en años). Se inicia un nuevo tiempo *solo cuando alguien fallece.*
   b. La segunda columna $(n_i)$ corresponde al número de individuos que están en riesgo de fallecer al comenzar ese período. Son los que *inician* vivos el período. Se incluye al individuo o individuos que morirán precisamente en ese tiempo.
   c. La tercera *columna* $(d_i)$ corresponde a los que mueren en el período de tiempo dado.
   d. Para entender mejor esta tabla, se pueden representar gráficamente los datos como en la figura 11.3: cada punto negro es una defunción; un punto blanco es un dato censurado. Debajo aparece la escala del tiempo en años.

***Tabla 11.3 Cálculo de la supervivencia acumulada y del error estándar transformado***

| AÑOS | MUERTE | $n_i$ | $s_i$ | $S_t$ | EET |
|---|---|---|---|---|---|
| 1 | 1 | 6 | 5 | 0,833 | $EEt = \sqrt{\frac{1}{(\ln[0,833])^2} \times \left[\frac{1}{6\times5}\right]} = 1,00$ |
| 2 | 0 | | | | |
| 3 | 1 | 4 | 3 | 0,625 | $EEt = \sqrt{\frac{1}{(\ln[0,6250])^2} \times \left[\frac{1}{6\times5} + \frac{1}{4\times3}\right]} = 0,73$ |
| 4 | 1 | 3 | 2 | 0,417 | $EEt = \sqrt{\frac{1}{(\ln[0,4167])^2} \times \left[\frac{1}{6\times5} + \frac{1}{4\times3} + \frac{1}{3\times2}\right]} = 0,61$ |
| 5 | 1 | 2 | 1 | 0,208 | $EEt = \sqrt{\frac{1}{(\ln[0,2083])^2} \times \left[\frac{1}{6\times5} + \frac{1}{4\times3} + \frac{1}{3\times2} + \frac{1}{2}\right]} = 0,56$ |
| 6 | 0 | | | | |

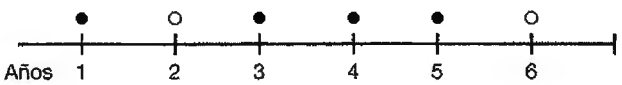



**Figura 11.3** Representación gráfica de los datos para construir una tabla de supervivencia.

3. *Calcular para cada tiempo el cociente entre supervivientes y sujetos en riesgo de fallecer.* Se trata de calcular la cuarta columna de la tabla 11.3 según la fórmula $s_i/n_i$. Así se obtendrá la supervivencia específica para cada tiempo que se considera.
4. *Multiplicar en cada tiempo los cocientes ($s_i/n_i$) por los de los tiempos previos.* La supervivencia acumulada para cada tiempo será precisamente este producto:

$$S_t = \prod \frac{s_i}{n_i}$$

Los valores de supervivencia acumulada (Kaplan-Meier) serán:

a. Desde el inicio hasta <1 año: $S_0$ = **1,00** (el 100% están vivos).
b. De 1 a <3 años: $S_1$ = 5/6 = **0,833.**
c. A los 3 años (hasta <4): $S_3$ = 5/6 × 3/4 = **0,625.**
d. A los 4 años (hasta <5): $S_4$ = 5/6 × 3/4 × 2/3 = **0,417.**
e. A los 5 años: $S_5$ = 5/6 × 3/4 × 2/3 × 1/2 = **0,208.**

## 11.4. REPRESENTACIÓN GRÁFICA DEL ESTIMADOR DE KAPLAN-MEIER

Cualquier análisis de supervivencia se suele acompañar de su representación gráfica para expresar visualmente cómo va disminuyendo la probabilidad de sobrevivir a medida que pasa el tiempo. Siempre se sitúa el tiempo en el eje de abscisas («x») y el porcentaje de los que sobreviven en el de ordenadas («y»). Esta representación se denomina curva de Kaplan-Meier y es muy utilizada en las publicaciones médicas (fig. 11.4).

Se debe empezar con una supervivencia de 1, hasta que se produce el primer fallecimiento. Entonces, la gráfica baja con el salto correspondiente a la reducción de supervivencia a partir de ese momento y así sucesivamente. Cuando el más largo de los tiempos corresponde a un sujeto que seguía vivo al término del período de observación, se deja una línea horizontal al final. Cuando el paciente que ha tenido el tiempo de observación más prolongado haya

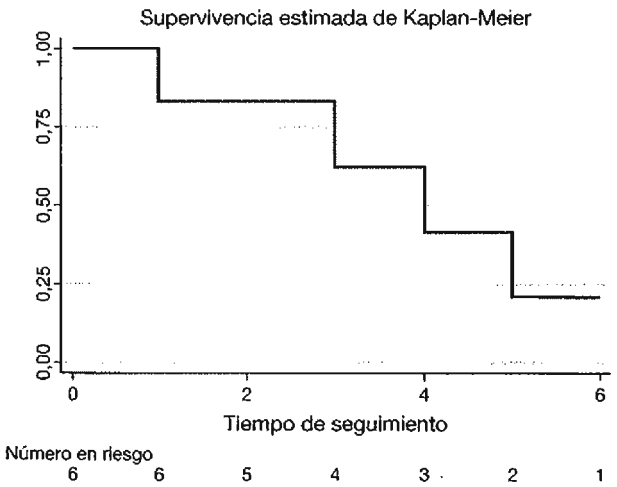



**Figura 11.4** Curva de Kaplan-Meier.

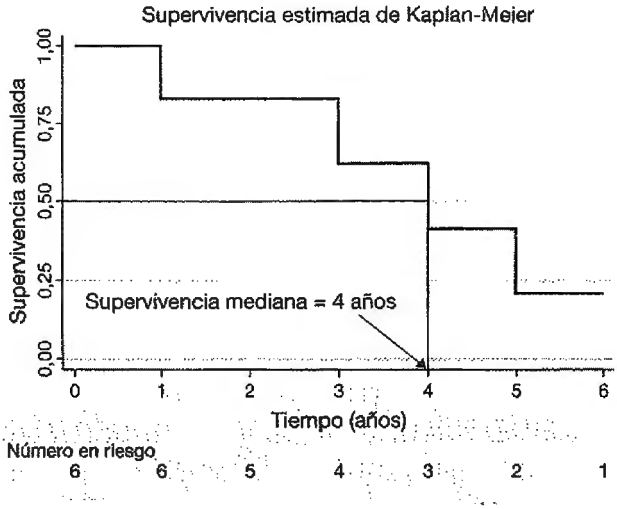



**Figura 11.5** Estimación de la mediana de supervivencia.

fallecido al final de este tiempo, la gráfica acabará verticalmente para cortar el eje de abscisas (supervivencia final = 0).

Se observa que, entre 0 y 1 año, la supervivencia es 1 (no ha fallecido nadie); justamente al llegar al año, la supervivencia disminuye a 0,833 y se mantiene ahí hasta los 3 años, en que desciende a 0,625; se mantiene en ese valor hasta que experimenta otro salto a los 4 años (disminuye a 0,417), y el último salto se produce a los 5 años (0,208). A partir de los 5 años no puede decirse nada sobre la supervivencia, porque ningún sujeto ha fallecido tras 5 años de observación. Un asunto importante es que hay que completar la gráfica, indicando debajo del eje de abscisas los sujetos que están a riesgo ($n_j$) en cada momento.

Para estimar gráficamente la *supervivencia mediana,* se traza una perpendicular desde el valor $S_t = 0,5$ del eje de ordenadas (fig. 11.5). En el punto en que esta recta corta a la curva de Kaplan-Meier se situará la supervivencia mediana. En el ejemplo será de 4 años. Es el primer momento durante el seguimiento en el que la supervivencia global acumulada sea ≤50%.

Es interesante tener en cuenta este aspecto, ya que un error frecuente de interpretación proviene de pensar que la supervivencia mediana correspondería a la mediana *de la variable tiempo de supervivencia.* Esto *no* es así. *La mediana de la variable tiempo de supervivencia no es la supervivencia mediana.* En el ejemplo, los tiempos de supervivencia son los correspondientes a la columna *años* de las tablas 11.1 y 11.2. Los datos ordenados de esta variable serían 1, 2, 3, 4, 5 y 6. Su mediana sería 3,5, pero esa no es la supervivencia mediana. Si se eliminasen los sujetos censurados, los datos que quedarían (1, 3, 4 y 5) también tendrían una mediana de 3,5, pero tampoco sería la supervivencia mediana. La supervivencia mediana es el tiempo en el que están vivos el 50% de los pacientes. En el ejemplo, la supervivencia mediana es 4 años y se calcula por el método gráfico que hemos visto.

En algunos casos, como en el ejemplo de la tabla 11.4, *no existe mediana,* ya que afortunadamente incluso al final del seguimiento sobreviven más del 50% de los pacientes.

Dado que los saltos se producen solo cuando se observa alguna muerte, ¿cómo intervienen los censurados en la supervivencia acumulada? Cada dato censurado influye *disminuyendo el denominador* de los cocientes $s/n_j$ siguientes. Aunque un dato censurado no provoque un salto, su influencia es notable, pues hace que el siguiente salto tenga mayor magnitud.

*Tabla 11.4 Ejemplo de datos de supervivencia sin (afortunadamente) supervivencia mediana*

| TIEMPO | ESTADO | $s_i$ | $n_i$ | $s_i/n_i$ | SUPERVIVENCIA ACUMULADA |
|---|---|---|---|---|---|
| 1 | 1 | | | | |
| 1 | 1 | 9 | 11 | 0,818 | 0,818 |
| 1 | 0 | | | | |
| 2 | 0 | | | | |
| 5 | 1 | 6 | 7 | 0,857 | (0,818)(0,857) = 0,701 |
| 5 | 0 | | | | |
| 6 | 1 | 4 | 5 | 0,8 | (0,8)(0,701) = 0,561 |
| 6 | 0 | | | | |
| 6 | 0 | | | | |
| 12 | 0 | | | | |
| 12 | 0 | | | | |

## 11.5. INTERVALOS DE CONFIANZA PARA LA ESTIMACIÓN DE SUPERVIVENCIA ACUMULADA

La supervivencia acumulada (Kaplan-Meier) se ha calculado en una muestra. Para estimar la supervivencia poblacional pueden construirse intervalos de confianza a partir de la supervivencia acumulada en la muestra y de su error estándar (EE). Sin embargo, no se debe usar directamente el error estándar que produce SPSS o STATA para sumarlo y restarlo $z$ veces a la supervivencia estimada, pues no siempre es válida la aproximación a la normal con ese error estándar. Lo más adecuado es obtener los intervalos de confianza a partir de un error estándar transformado ($EE_t$).

$$EE_t = \sqrt{\frac{1}{(\ln[S_t])^2} \times \sum \frac{n_i - s_i}{n_i s_i}}$$

donde ln significa logaritmo natural (neperiano) y $S_t$ es la supervivencia acumulada en el tiempo $t$. Las cantidades $n_i$ y $s_i$ son, respectivamente, el número de sujetos en riesgo y el número de supervivientes en cada tiempo.

Una vez obtenido el error estándar transformado, se determinan los límites de confianza para la supervivencia acumulada según la siguiente expresión:

$$IC_{1-\alpha} = S_t^{EXP(z_{\alpha/2}\, EEt)}$$

en la que $z_{\alpha/2}$ es el valor de la distribución normal para el error alfa respectivo. En concreto, $z_{\alpha/2} = 1{,}96$ para un intervalo de confianza al 95%. EXP supone elevar a la cantidad correspondiente el número e, base de los logaritmos naturales.

En el primer ejemplo, los errores estándar serían los que muestra la quinta columna de la tabla 11.3. En la tabla 11.5 se recogen los intervalos de confianza, calculados a partir de estos errores estándar. Se observa que solo hay una ligerísima diferencia con los calculados por STATA.

*Tabla 11.5 Construcción de intervalos de confianza a partir del EEt*

| AÑOS | $S_t$ | EEt | IC 95% ($S_t$) |
|---|---|---|---|
| 1 | 0,833 | 1,00 | $0{,}833^{EXP(\pm 1{,}96 \times 1)}$ = 0,274 a 0,975 |
| 2 | | | |
| 3 | 0,625 | 0,73 | $0{,}625^{EXP(\pm 1{,}96 \times 0{,}73)}$ = 0,142 a 0,893 |
| 4 | 0,417 | 0,61 | $0{,}417^{EXP(\pm 1{,}96 \times 0{,}61)}$ = 0,056 a 0,767 |
| 5 | 0,2083 | 0,56 | $0{,}208^{EXP(\pm 1{,}96 \times 0{,}56)}$ = 0,009 a 0,595 |
| 6 | | | |

Los cálculos se complican a medida que transcurre más tiempo de observación. Por eso es preferible usar STATA. Si no se dispone de STATA, puede transformarse el error estándar (EE) de SPSS para lograr el error estándar transformado (EEt), mediante la siguiente expresión:

$$EEt = \sqrt{\frac{1}{(\ln[S_t])^2} \times \left(\frac{S_t}{EE}\right)^2}$$

Se puede dejar programado en Excel[3]. Si se ha introducido el valor de la supervivencia en la casilla A2 y su error estándar convencional (el que aparece, por ejemplo, en SPSS en la casilla B2), deberá indicarse:

C2=(((B2/A2)^2)*(1/(LN(A2))^2))^0,5 devolverá: error estándar transformado (EEt).

D2=(A2)^EXP(1,96*C2) devolverá: límite inferior de confianza al 95%.

E2=A2^EXP(–1,96*C2) devolverá: límite superior de confianza al 95%.

## 11.6. ANÁLISIS DE SUPERVIVENCIA CON STATA

Para realizar análisis de supervivencia con STATA, lo primero que se ha de hacer es indicar al programa la variable que indica el ***tiempo*** de seguimiento y la variable que recoge si el participante ha desarrollado o no el evento de interés al final de su tiempo de seguimiento. Así, si denominamos ***tiempo*** a la primera de estas variables y ***estado*** a la segunda, codificada como 0 para los censurados y 1 para los que han presentado el evento de interés, se deberá ordenar:

**Statistics → Survival analysis → Setup and utilities → Declare data to be survival-time data**

y, en el menú que se despliega, indicar que la *Time variable* es la variable *tiempo*, que el evento de interés *(Failure event)* está definido por la variable *(Failure variable) estado* y que el valor que indica el evento de interés *(Failure values)* es 1. También se puede usar directamente la orden **stset** y ejecutarla:

```
stset tiempo, failure(estado==1)
```

STATA facilitará un resumen de la información referente a los datos de supervivencia:

```
     failure event:  estado == 1
obs. time interval:  (0, tiempo]
 exit on or before:  failure

------------------------------------------------------------------------------
         6  total obs.
         0  exclusions
------------------------------------------------------------------------------
         6  obs. remaining, representing
         4  failures in single record/single failure data
        21  total analysis time at risk, at risk from t =         0
                          earliest observed entry t =         0
                               last observed exit t =         6
```

Como puede observarse, 6 sujetos aportan información al análisis de supervivencia *(6 obs. remaining)*, 4 de ellos presentan el evento de interés *(failures in single record/single failure data)* y el sujeto con mayor tiempo de seguimiento ha sido seguido durante 6 años *(last observed exit t)*.

3 Puede encontrarse un programa en Excel ya preparado en la siguiente dirección de internet: **www.unav.es/preventiva → docencia → bioestadística.**

Para obtener el gráfico de Kaplan-Meier por menús, se deberá elegir:

**Statistics → Survival analysis → Graphs → Kaplan-Meier survivor function**

En la pestaña *At-risk table* conviene señalar la opción *Show at-risk table beneath graph* e indicar en las opciones que se desea para cada año desde el año 0 hasta el año 6, indicando, para ello, en la línea *Evaluation points: 0(1)6.*

También se puede emplear la orden:

```
sts graph, risktable(0(1)6)
```

Así se obtendrá la figura 11.6.

Para producir la tabla de supervivencia, que además ofrece directamente los intervalos de confianza al 95%, se puede aplicar la siguiente ruta:

**Statistics → Survival analysis → Summary statistics, tests, and tables → Life tables for survival data**

indicando en el menú que aparece que la variable tiempo *(Time variable)* es ***tiempo*** y la variable que indica el evento de interés *(Failure variable)* es ***estado***. Se puede ejecutar también la orden:

```
ltable tiempo estado, survival
```

Así se obtendrá el resultado:

```
. ltable tiempo estado, survival
```

| Interval | | Beg. Total | Deaths | Lost | Survival | Std. Error | [95% Conf. | Int.] |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 6 | 1 | 0 | 0.8333 | 0.1521 | 0.2731 | 0.9747 |
| 2 | 3 | 5 | 0 | 1 | 0.8333 | 0.1521 | 0.2731 | 0.9747 |
| 3 | 4 | 4 | 1 | 0 | 0.6250 | 0.2135 | 0.1419 | 0.8931 |
| 4 | 5 | 3 | 1 | 0 | 0.4167 | 0.2218 | 0.0560 | 0.7665 |
| 5 | 6 | 2 | 1 | 0 | 0.2083 | 0.1844 | 0.0087 | 0.5951 |
| 6 | 7 | 1 | 0 | 1 | 0.2083 | 0.1844 | 0.0087 | 0.5951 |

STATA muestra, para cada intervalo, el número de sujetos en riesgo, el número de sujetos que desarrollan el evento de interés, los sujetos que se censuran en un determinado período *(Lost)* y la supervivencia global acumulada para cada período, junto con su error estándar y su intervalo de confianza al 95%.

## 11.7. ANÁLISIS DE SUPERVIVENCIA CON OTROS PROGRAMAS

Para describir la supervivencia con SPSS mediante el método de Kaplan-Meier, se debe aplicar la siguiente secuencia de instrucciones (se insiste en que no siempre se podría elegir el nombre «AÑOS» para una variable, porque algunas versiones de SPSS no aceptan la ñ):

**Analizar → Supervivencia → Kaplan-Meier... →**
**Tiempo: TIEMPO DE SEGUIMIENTO → Estado: MUERTE →**
**Definir evento... → Valor único: 1 (Continuar) →**
**Opciones... → Estadísticos: Tabla(s) de supervivencia, Media y mediana de supervivencia (señalados por defecto). Gráficos → Supervivencia (Continuar) (Aceptar)**

Se obtendrá un resultado numérico y una gráfica. El resultado será:

| Survival Analysis for AÑOS | | | | tiempo de seguimiento | |
|---|---|---|---|---|---|
| Time | Status | Cumulative Survival | Standard Error | Cumulative Events | Number Remaining |
| 1,0 | Muerte | ,8333 | ,1521 | 1 | 5 |
| 2,0 | Censurado | | | 1 | 4 |
| 3,0 | Muerte | ,6250 | ,2135 | 2 | 3 |
| 4,0 | Muerte | ,4167 | ,2218 | 3 | 2 |
| 5,0 | Muerte | ,2083 | ,1844 | 4 | 1 |
| 6,0 | Censurado | | | 4 | 0 |

Number of Cases: 6 Censored: 2 (33,33%) Events: 4

| | Survival Time | Standard Error | 95% Confidence Interval |
|---|---|---|---|
| Mean: | 3,9 | ,7 | (2,5; 5,3) |
| (Limited to | 6,0) | | |
| Median: | 4,0 | 1,1 | (1,9; 6,1) |

*La primera línea indica cuál es el objetivo de la tabla. En nuestro ejemplo: «Survival analysis for AÑOS tiempo de seguimiento». Es decir, señala que el programa ha realizado un análisis de supervivencia y que la variable AÑOS (etiqueta: tiempo de seguimiento) es la que recoge el tiempo transcurrido hasta la muerte de cada paciente (o tiempo total de observación, si el paciente no ha muerto).*

Después aparece una tabla, en la que se indica el tiempo *(Time)* durante el cual se ha observado a cada sujeto, si se ha producido o no el evento de interés *(Status)*, la supervivencia global acumulada *(Cumulative survival)* y el error estándar de la supervivencia acumulada *(Standard error)*. Este error estándar corresponde a la expresión $EE = S_t\sqrt{\Sigma \frac{n_i - s_i}{n_i \times s_i}}$, pero, como ya se ha explicado con anterioridad, salvo que se trate de muestras muy grandes, no puede usarse directamente para obtener una aproximación por la distribución normal (IC 95% = $S_t \pm 1{,}96 \times EE$). A continuación se muestran los desenlaces acumulados *(Cumulative events)* hasta ese tiempo y el número de individuos que aún no han experimentado el evento de interés, es decir, los supervivientes *(Number remaining)*.

La información que aparece debajo de la tabla corresponde al número de sujetos que comienzan el estudio *(Number of cases)*, los sujetos con información truncada o censurada *(Censored)* y el número de eventos de interés *(Events)* que se han producido durante el seguimiento.

Por último, aparece una segunda tabla que proporciona los valores del tiempo medio de supervivencia y la mediana, con sus correspondientes errores estándar y el intervalo de confianza al 95%, siendo *survival time* el tiempo de supervivencia, *standard error* el error estándar y *95% confidence interval* el intervalo de confianza para la media *(Mean)* y la mediana *(Median)*. La anotación *limited to* indica que solo considera el período de tiempo de seguimiento, aunque, como es lógico, al existir pacientes censurados habrá algunos individuos que fallezcan después de los 6 años.

Después se mostrará la representación gráfica de la supervivencia acumulada ya vista. Conviene tener en cuenta que la gráfica por defecto no aparece si no se pide expresamente a SPSS desde «Opciones». También es importante añadir siempre al pie de la gráfica el número de sujetos en riesgo ($n_i$) que había para cada tiempo.

## 11.8. CURVAS DE INCIDENCIA DE NELSON-AALEN

El estimador de Nelson-Aalen es un estimador no paramétrico de la tasa instantánea *(hazard) acumulada* de presentar el evento de interés que se está considerando (**8**). En un determinado tiempo *t*, el *hazard* se definiría como el cociente entre las personas que presentan el desenlace de interés (p. ej., fallecimiento) y el número de personas en riesgo de fallecer en ese momento dado *(d/n)*. A su vez, la función del *hazard* acumulado hasta un determinado momento sería simplemente la suma de todos los *hazards* observados en todos los momentos en los que se haya producido un evento de interés hasta el tiempo *t*. Por ello, la ecuación se podría formular como:

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Para obtener una representación gráfica de este estimador con STATA, simplemente se deberá añadir la opción **na** en la orden **sts graph.**

## 11.9. COMPARACIÓN DE CURVAS DE SUPERVIVENCIA: TEST DEL *LOG-RANK*

Para comparar *2 o más* curvas de supervivencia se usan diversas pruebas estadísticas de contraste de hipótesis. La *hipótesis nula* supone que los grupos comparados presentan *igual supervivencia globalmente,* es decir, no hay diferencias persistentes.

La prueba más empleada para comparar curvas de supervivencia es el *test del log-rank.* Este test tiene en cuenta las diferencias de supervivencia entre grupos en todos los puntos del tiempo que dura el seguimiento. En medicina, *desafortunadamente,* se hizo práctica frecuente (que, por suerte, va desapareciendo) comparar la supervivencia tomando como referencia *un solo punto* común en el tiempo. Por ejemplo, para comparar dos tratamientos o dos series se tiende a usar como medida de resultado la supervivencia de cada grupo de pacientes a los 5 años. Esta decisión es básicamente *incorrecta.* Dos situaciones muy distintas pueden dar lugar a los mismos resultados de supervivencia a 5 años, como, por ejemplo, si la supervivencia del grupo A fuese mejor durante el seguimiento pero al final ambas curvas se igualasen a los 5 años. En esta situación, al comparar supervivencias en un solo punto de tiempo (5 años), se concluiría erróneamente que ambos grupos son iguales y se desperdiciaría mucha información. El test del *log-rank* tiene en cuenta todos los puntos en el tiempo para comparar la supervivencia de los dos grupos. Cuando se comparan varios grupos, este test detecta *heterogeneidad* (al menos un grupo es distinto de otro). En este sentido, es análogo al ANOVA.

El test del *log-rank* se ha llamado también test generalizado de Savage (**6,8**) y es esencialmente idéntico al test de Mantel y Haenszel, que se usa en el análisis epidemiológico estratificado para contrastar la hipótesis nula de que la *odds ratio* vale 1 (**8**). Además del test del *log-rank,* existen otros métodos de comparación de curvas de supervivencia:

- Test generalizado de Wilcoxon (también llamado de Breslow).
- Test de Tarone-Ware.
- Test de Peto-Peto-Prentice.
- Test de Fleming-Harrington.

Estos test están incorporados en los distintos programas de *software,* pero en la actualidad se tiende a reemplazarlos por la regresión de Cox (v. apartado 14.6). Todos ellos se basan en comparar las muertes observadas en cada grupo con las esperadas si la mortalidad fuese igual en todos los grupos ($H_0$). Pero difieren entre sí porque en cada uno se realiza una ponderación distinta y se otorga distinto peso a las diferencias según ocurran antes o después a lo largo del seguimiento.

Como siempre que se dispone de más de un test estadístico para responder a una pregunta dada, es necesario manejar criterios sobre los resultados que se presentarán. El más frecuente y aceptado para comparar curvas de supervivencia es el test del *log-rank*. Una aproximación sensata consiste en pedir los diversos test al ordenador y, si aportan resultados concordantes (valores *p* muy parecidos), lo más apropiado será presentar solo el test del *log-rank*. En cambio, si hay diferencias entre los resultados de los test, por ejemplo, si el test del *log-rank* y el de Wilcoxon arrojan resultados diferentes, se deberían presentar los resultados de ambos (**8**). De esta forma, el lector se hará una idea más clara de hasta qué punto las curvas de supervivencia pueden considerarse diferentes. Los valores de los test de Tarone-Ware, Peto-Peto-Prentice y Fleming-Harrington son, en general, intermedios entre el del *log-rank* y el de Wilcoxon.

Cuando el *evento es poco frecuente* o las curvas son *divergentes* (no se cruzan), el *log-rank* es el método más indicado y, en general, existirá concordancia entre los diversos test. Cuando las curvas se cruzan (al principio hay mejor supervivencia en un grupo y luego en el otro), estará indicado también realizar al menos otro test de comparación de curvas de supervivencia, preferentemente el de Wilcoxon o Breslow, ya que puede existir más disparidad de resultados.

En estos test se calcula una ji cuadrado que tiene como grados de libertad el número de grupos comparados menos 1. Para cada tiempo en que se produce un evento se calcula una diferencia entre los eventos observados en un grupo (solo en uno y siempre el mismo) y los esperados en ese grupo si la probabilidad de morir fuese igual en todos los grupos. Para este fin se crea una tabla de contingencia para cada tiempo en que alguien fallece. También se calcula en cada tiempo una varianza basada en la distribución hipergeométrica, que en el test del *log-rank* corresponde a un cociente cuyo numerador es el producto de los marginales y el denominador el producto del gran total al cuadrado ($T^2$) de la tabla por T-1. Al final se suman todas las diferencias entre valores observados y esperados (O – E) y todas las varianzas. La suma de las diferencias (O – E) se eleva al cuadrado y se usa como numerador. La suma de las varianzas se incluye en el denominador, y la ji cuadrado se calcula como:

$$\chi^2 = \frac{\left[\sum (O_i - E_i)\right]^2}{\sum \text{var}_i}$$

El cálculo manual del *log-rank* suele resultar muy tedioso y es preferible siempre realizarlo con ordenador.

Imagínese que los seis participantes que se han sometido a estudio en el ejemplo visto hasta ahora habían recibido un tipo de tratamiento, que se dispone de información de otros seis participantes a los que se había administrado otro tratamiento y se desea comparar las dos supervivencias. Si se representasen gráficamente ambas curvas de supervivencia con la orden:

```
sts graph, risktable(0(1)6) by(tratamiento)
```

en STATA se obtendría la figura 11.6.

Para comparar ambas curvas, se pueden obtener estos test con STATA a partir de los menús según:

**Statistics → Survival analysis → Summary statistics, tests, and tables → Test equality of survivor functions**

indicando la variable que define los grupos *(Variables), tratamiento,* y el test que se desea obtener; sin embargo, antes se debe haber ejecutado el **stset.**

A continuación se muestra cómo obtener en STATA cada uno de los test mencionados mediante órdenes, así como su correspondiente resultado:

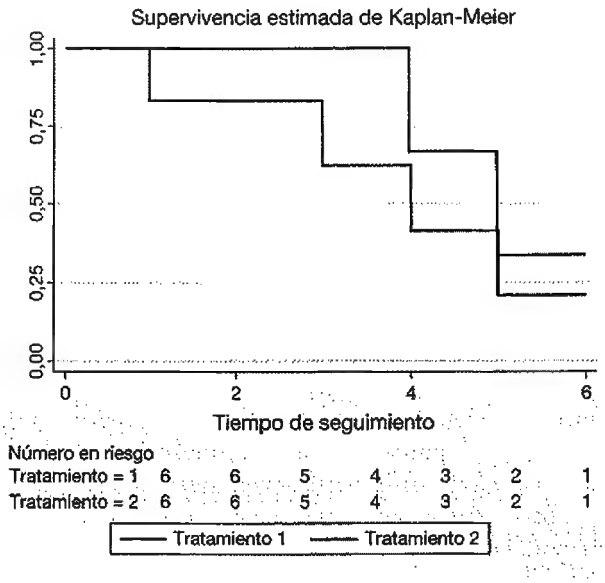



**Figura 11.6** Representación gráfica de la supervivencia global acumulada de dos grupos de pacientes.

```
. sts test tratamiento, logrank

        failure _d:  estado == 1
  analysis time _t:  tiempo
```

Log-rank test for equality of survivor functions

| tratamiento | Events observed | Events expected |
|---|---|---|
| 1 | 4 | 3.00 |
| 2 | 2 | 3.00 |
| Total | 6 | 6.00 |

```
                chi2(1) =       0.81
                Pr>chi2 =     0.3679
```

```
. sts test tratamiento, wilcoxon

        failure _d:  estado == 1
  analysis time _t:  tiempo
```

Wilcoxon (Breslow) test for equality of survivor function

| tratamiento | Events observed | Events expected | Sum of ranks |
|---|---|---|---|
| 1 | 4 | 3.00 | 10 |
| 2 | 2 | 3.00 | -10 |
| Total | 6 | 6.00 | 0 |

```
                chi2(1) =       1.39
                Pr>chi2 =     0.2377
```

```
. sts test tratamiento, tware

        failure _d:  estado == 1
  analysis time _t:  tiempo
```

Tarone-Ware test for equality of survivor functions

| tratamiento | Events observed | Events expected | Sum of ranks |
|---|---|---|---|
| 1 | 4 | 3.00 | 3.1462644 |
| 2 | 2 | 3.00 | -3.1462644 |
| Total | 6 | 6.00 | 0 |

chi2(1) = 1.13
Pr>chi2 = 0.2870

```
. sts test tratamiento, peto

        failure _d:  estado == 1
  analysis time _t:  tiempo
```

Peto-Peto test for equality of survivor functions

| tratamiento | Events observed | Events expected | Sum of ranks |
|---|---|---|---|
| 1 | 4 | 3.00 | .87179487 |
| 2 | 2 | 3.00 | -.87179487 |
| Total | 6 | 6.00 | 0 |

chi2(1) = 1.36
Pr>chi2 = 0.2440

```
. sts test tratamiento, fh (0 0)

        failure _d:  estado == 1
  analysis time _t:  tiempo
```

Fleming-Harrington test for equality of survivor function

| tratamiento | Events observed | Events expected | Sum of ranks |
|---|---|---|---|
| 1 | 4 | 3.00 | 1 |
| 2 | 2 | 3.00 | -1 |
| Total | 6 | 6.00 | 0 |

chi2(1) = 0.81
Pr>chi2 = 0.3679

***Interpretación:*** *si los dos grupos tuviesen la misma supervivencia, la probabilidad de encontrar unas diferencias iguales o mayores a las observadas sería superior al 23%, con independencia del test que se emplee. Por tanto, no podrá rechazarse la hipótesis nula que mantiene la igualdad en la supervivencia entre los participantes que recibieron los dos tratamientos.*

En SPSS, desde el menú de Kaplan-Meier basta incluir la variable que define los grupos en la ventana «Factor» y seleccionar el test en «Comparar factor».

## 11.10. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Operación | STATA | SPSS |
|---|---|---|
| Tabla de supervivencia | **ltable *tiempo estado*, survival** | **KM tiempo /STATUS=estado(1) /PRINT TABLE MEAN.** |
| Curva de Kaplan-Meier | **stset *tiempo*, failure(*estado*==1)**<br>**sts graph, risktable(0(1)6)** | **KM tiempo /STATUS=estado(1) /PRINT TABLE MEAN /PLOT SURVIVAL.** |

| Operación | STATA | SPSS |
|---|---|---|
| Curva de Nelson-Aalen | **stset *tiempo*, failure(*estado*==1)**<br>**sts graph, risktable(0(1)6) na** | |
| Curva de Kaplan-Meier para distintos grupos | **stset *tiempo*, failure(*estado*==1)**<br>**sts graph, risktable(0(1)6) by(*tratamiento*)** | **KM**<br>**tiempo BY tratamiento /**<br>**STATUS=estado(1)**<br>**/PRINT TABLE MEAN**<br>**/PLOT SURVIVAL.** |
| Comparación de curvas de supervivencia | **sts test *tratamiento*, logrank**<br>**sts test *tratamiento*, wilcoxon**<br>**sts test *tratamiento*, tware**<br>**sts test *tratamiento*, peto**<br>**sts test *tratamiento*, fh (0 0)** | **KM**<br>**tiempo BY tratamiento /**<br>**STATUS=estado(1)**<br>**/PRINT TABLE MEAN**<br>**/TEST LOGRANK**<br>**BRESLOW TARONE**<br>**/COMPARE OVERALL**<br>**POOLED.** |

## REFERENCIAS


1. Cox DR. Regression model and life tables. J Roy Statist Soc B 1972;34:187-220.
2. Cox DR, Oakes D. Analysis of survival data. London: Chapman & Hall; 1984.
3. Lagakos SW. Statistical analysis of survival data. En: Bailar JC III, Mosteller F, editors. Medical uses of statistics. 2nd ed. Boston: NEJM Books; 1992. p. 281-91.
4. Lee ET. Statistical methods for survival data analysis. New York: Wiley; 1992.
5. Collett D. Modelling survival data in medical research. London: Chapman & Hall; 1994.
6. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Berlin: Springer Verlag; 1997.
7. Bland JM, Altman DG. Time to event (survival) data. BMJ 1998;317(7156):468-9.
8. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd ed. Hoboken: John Wiley & Sons; 2008.
9. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ 1998;317 (7172):1572.
10. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel; 2008.
11. Coviello V, Boggess M. Cumulative incidence estimation in the presence of competing risks. Stata J 2004;4(2):103-12.
12. Pintilie M. An introduction to competing risks analysis. Rev Esp Cardiol 2011;64(7):599-605.

# INTRODUCCIÓN A LOS MODELOS MULTIVARIABLES. REGRESIÓN LINEAL MÚLTIPLE

12

*A. Sánchez-Villegas, C. López del Burgo, M. Á. Martínez-González*

## 12.1. INTRODUCCIÓN

El análisis multivariable se define como aquel análisis estadístico que incluye y analiza tres o más variables simultáneamente. Aunque los análisis multivariables más habituales son los *modelos de regresión* (múltiple, logística, Cox y Poisson), existen otros, como el análisis factorial o el análisis de conglomerados.

Los modelos de regresión intentan explicar un fenómeno de salud (o enfermedad) (variable Y, dependiente) teniendo en cuenta varias variables simultáneamente (variables X, independientes)[1].

La aplicación de un modelo de regresión responderá siempre a uno o varios de los siguientes objetivos de investigación:

1. Determinar los predictores de una determinada variable biosanitaria (variable Y) a partir de una lista más o menos amplia de posibles variables explicativas (variables X). Por ejemplo, de un listado de 20 posibles polimorfismos genéticos, ¿cuántos y cuáles de estos polimorfismos son capaces de predecir en más o menos grado la probabilidad de ser obeso? ¿Cuál será la probabilidad de ser obeso si se posee el polimorfismo A? ¿Y si se tienen los polimorfismos A y B? ¿Y si se poseen los polimorfismos A, B y C?
2. Construir un índice pronóstico (ecuación) para predecir una determinada condición (variable Y) a partir de los valores recogidos en otras variables (variables X). Por ejemplo, predecir la probabilidad de que un paciente presente enfermedad coronaria en los siguientes 10 años conociendo su sexo, su edad, el nivel de colesterol, la tensión arterial y el hábito tabáquico (ecuación de Framingham [**1**]).
3. Determinar el efecto de una variable $X_1$ sobre otra variable Y teniendo en cuenta otras características ($X_2$, $X_3$... $X_p$; factores de confusión) que pudieran distorsionar la verdadera asociación entre estas variables (**2**). Por ejemplo, determinar el efecto del consumo de comida rápida sobre el riesgo de desarrollar depresión, considerando la cantidad de actividad física realizada por el individuo y su hábito tabáquico[2] (**3**).
4. Detectar y describir fenómenos de interacción entre variables (modificación del efecto) sobre un determinado resultado. Es decir, si la presencia de una variable $X_2$ es capaz de modificar el efecto ejercido por la variable $X_1$ sobre la variable dependiente Y (**4**). Por ejemplo, se desea determinar si el efecto del consumo de una dieta rica en grasas saturadas sobre el cambio de peso es diferente según exista o no presencia de un determinado polimorfismo genético (**5**).

1 En ocasiones se conoce también como análisis multivariante. No obstante, en sentido estricto, se habla de multivariable cuando existe una sola variable dependiente (respuesta), aunque haya muchas variables predictoras o independientes. El término «multivariante» se reserva para cuando también hay muchas variables dependientes o de respuesta.

2 Se ha observado que la práctica de ejercicio reduce el riesgo de depresión, mientras que el consumo de tabaco lo aumenta. Como es posible que exista una agrupación de estilos de vida poco saludables en el mismo individuo, puede ocurrir que quien consuma mucha comida rápida sea también fumador y no haga deporte, por lo que el verdadero efecto de la comida rápida sobre la depresión puede estar confundido.

*Tabla 12.1 Ejemplos típicos de análisis multivariable en ciencias de la salud*

| | EJEMPLO 1 | EJEMPLO 2 | EJEMPLO 3 |
|---|---|---|---|
| Se desea explicar... | Cambio de peso | Aparición de depresión en la edad adulta | Mortalidad durante una media de seguimiento de 7 años |
| Variable dependiente («efecto») | Cambio de peso (kg) | Incidencia de depresión en la edad adulta (sí/no) | Mortalidad (sí/no) y tiempo (años) |
| Supuesta(s) «causa(s)» que se trata de valorar | Intervención dietética | Imagen corporal a los 5 años | Adhesión a dieta mediterránea |
| Variables independientes... | Centro sanitario<br>Edad<br>Sexo | Índice de masa corporal en la edad adulta (IMC) | Sexo<br>Edad<br>Años de universidad<br>Índice de masa corporal<br>Tabaco<br>Actividad física<br>Prevalencia de enfermedades |
| Se empleará... | **Regresión MÚLTIPLE** | **Regresión LOGÍSTICA** | **Regresión de COX** ***(proportional hazards model)*** |
| ¿Por qué? | La variable dependiente es **cuantitativa** | La variable dependiente es **dicotómica** | La variable dependiente es del tipo **«tiempo hasta un evento»** |
| Referencia bibliográfica | (6) | (7) | (8) |

Los principales modelos de regresión se caracterizan por incluir como variables independientes (X), variables de tipo cuantitativo o cualitativas dicotómicas (cuando se disponga de variables independientes cualitativas de más de dos categorías, deberán crearse variables indicadoras o *dummy;* v. apartado 12.15). Sin embargo, estos modelos difieren principalmente en el tipo de variable dependiente que incluyen. En la tabla 12.1 se presentan tres ejemplos de modelos de regresión según las características de la variable dependiente.

En el primer ejemplo de la tabla 12.1 se valoró si el cambio de diferentes factores de riesgo cardiovascular (incluido el cambio de peso) podía explicarse por una intervención dietética (6). Aunque hay que tener en cuenta otros factores distintos de la dieta que influyen en el peso, en este ejemplo se deberán controlar pocos factores más, ya que se trataba de un ensayo de gran tamaño correctamente aleatorizado (asignación de dietas al azar). La aleatorización tiende a producir grupos comparables en cuanto a características conocidas y desconocidas que pudieran distorsionar la comparación entre dichos grupos (7). El cambio de peso es una variable cuantitativa o numérica. Por lo tanto, lo indicado es una regresión que tiene una sola variable dependiente cuantitativa, pero más de una variable independiente, y se llama regresión lineal múltiple o, más escuetamente, *regresión múltiple.* La regresión múltiple es una extensión de la regresión lineal simple:

- Regresión lineal simple: $y = a + bx$
- Regresión múltiple: $y = a + b_1x_1 + b_2x_2 + \ldots\ldots b_px_p$

En el segundo ejemplo se valoró en una muestra de enfermeras (*Nurses' Health Study* II) si la obesidad en la edad infantil (cuantificada a través de la elección de una entre nueve siluetas corporales que definían a las participantes a los 5 años) se relacionaba con la aparición de depresión en la edad adulta (8). La variable dependiente es ahora de tipo dicotómico (*depre* = 1 si el sujeto desarrollaba esta patología en la edad adulta y *depre* = 0 cuando no lo hacía). Además, se tiene en cuenta el grado de obesidad de la participante en la edad adulta para eliminar el efecto que la obesidad en esta fase de la vida pudiera tener sobre la depresión, y se considerará solo el efecto de la obesidad en la edad infantil. Como la variable dependiente es cualitativa dicotómica, se usará la regresión *logística,* que viene a ser una extensión multivariable de la ji cuadrado.

*Tabla 12.2 Aspecto parcial de las bases de datos para diferentes modelos de regresión*

| REGRESIÓN MÚLTIPLE | | | | REGRESIÓN LOGÍSTICA | | | REGRESIÓN DE COX | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | $X_1$ | $X_2$ | $X_3$ | Y | $X_1$ | $X_2$ | Y | | $X_1$ | $X_2$ | $X_3$ |
| −1,28 | 2 | 65 | 0 | No | 4 | 28,1 | No | 34 | 3 | 1 | 38 |
| −5,22 | 1 | 71 | 0 | Sí | 7 | 22,3 | Sí | 14 | 1 | 2 | 45 |
| 0,56 | 3 | 69 | 1 | No | 2 | 26,2 | No | 51 | 5 | 1 | 51 |
| cambio peso | grupo tratam. | edad | sexo | diagn. depr. | imagen 5 años | IMC | fallec. | meses seguim. | quintil adh. | sexo | edad |

El tercer ejemplo (9) valoró si la adhesión a un patrón de dieta mediterránea podría reducir el riesgo de mortalidad en un seguimiento prospectivo durante una media de seguimiento de 7 años (estudio SUN, Seguimiento Universidad de Navarra). Si solo se hubiese tenido en cuenta si la persona moría o no en este período de tiempo (1 = sí y 0 = no), la situación sería idéntica a la del ejemplo 2. En cambio, ahora interesa también *el tiempo* que transcurre hasta que fallece.

Se dispone ahora de dos variables por participante:

1. Si muere o no durante el período de seguimiento:
   a. 1 = el participante fallece.
   b. 0 = no fallece.
2. Cuánto tiempo ha estado sometido a seguimiento (hasta morir o hasta acabar el estudio).

La primera variable es cualitativa dicotómica; la segunda es cuantitativa. Se deben combinar ambas, según técnicas análogas a las del análisis de supervivencia. Además, podrían existir otros factores, relacionados con el estilo de vida del participante y que, además, podrían influir en su riesgo de mortalidad (p. ej., el consumo de tabaco o la práctica de actividad física). En este caso, el contexto es un análisis multivariable. Cuando se desea realizar un análisis multivariable en esta situación, se aplicará la regresión de Cox o *proportional hazards model*. La regresión de Cox es una extensión multivariable de los métodos de *Kaplan-Meier*.

Las bases de datos presentarían el aspecto parcial mostrado en la tabla 12.2.

La figura 12.1 muestra el aspecto parcial de los datos en STATA para los diferentes modelos de regresión. El ejemplo 1 corresponde a un modelo de regresión múltiple. En el ejemplo 2 (regresión logística) se suele codificar con un valor de 1 a quienes son casos (diagnóstico de depresión en el ejemplo) y con un valor de 0 a los no diagnosticados. En el ejemplo 3 (regresión de Cox) hacen falta dos variables para construir la respuesta («efecto» o variable dependiente), ya que es preciso combinar el dato de si se ha producido o no el evento *(fall)* con el tiempo que ha tardado en producirse dicho evento. En quienes no se produce el evento, se asignará el tiempo total durante el cual han sido observados. El aspecto de la base de datos sería similar en el programa SPSS.

## 12.2. PRIMERA APROXIMACIÓN AL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

El modelo de regresión múltiple es una extensión a varias variables de un modelo de regresión simple (fig. 12.2). La ecuación de la regresión lineal *simple* es:

$$y = a + bx$$

donde Y es la variable dependiente (cuantitativa) y X la variable independiente. Esta ecuación se puede generalizar para el caso en que haya más de una variable independiente. Supóngase que existen tres variables independientes: $X_1$, $X_2$, $X_3$. Entonces puede construirse la ecuación:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3$$

**Ejemplo 1**

Variable dependiente: cambio de peso

Grupo 1 = dieta A
Grupo 2 = dieta B
Grupo 3 = control

| | camb_peso | grupo | edad | sexo |
|---|---|---|---|---|
| 1 | -1.28 | 2 | 65 | 0 |
| 2 | -5.22 | 1 | 71 | 0 |
| 3 | .56 | 3 | 69 | 1 |

**Ejemplo 2**

Variable dependiente: depresión

| | depre | imagen5 | IMC |
|---|---|---|---|
| 1 | 0 | 4 | 28.1 |
| 2 | 1 | 7 | 22.3 |
| 3 | 0 | 2 | 26.2 |

**Ejemplo 3**

Variable dependiente: fallecimiento + tiempo que tarda en ocurrir

Quintiles adhesión a dieta mediterránea

Sexo 1 = varón
Sexo 2 = mujer

| | fall | seguim_mes | qDM | sexo | edad |
|---|---|---|---|---|---|
| 1 | 0 | 34 | 3 | 1 | 38 |
| 2 | 1 | 14 | 1 | 2 | 45 |
| 3 | 0 | 51 | 5 | 1 | 51 |

**Figura 12.1** Aspecto parcial en STATA de las bases de datos para llevar a cabo modelos de regresión.

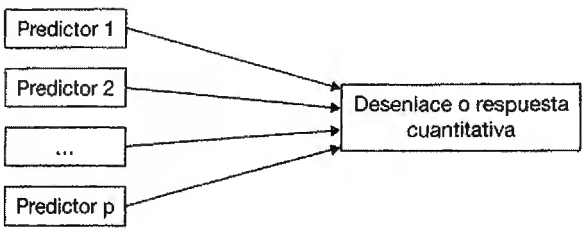



**Figura 12.2** Aplicación de la regresión múltiple.

Cada variable independiente $X_i$ tiene un coeficiente de regresión o pendiente propia $b_i$. Este coeficiente se interpretará como el cambio en la variable dependiente Y, por unidad de cambio en cada variable independiente ($X_1$, $X_2$ o $X_3$) *a igualdad de nivel de las otras variables independientes.* Es imposible interpretar una regresión si no se conocen las unidades de medida de cada variable. Esto se aplica tanto a la regresión simple como a la múltiple.

La regresión lineal múltiple puede llevarse a cabo a través del programa STATA, utilizando el menú o a través de instrucciones. Con el menú:

**Statistics → Linear models and related → Linear regression**

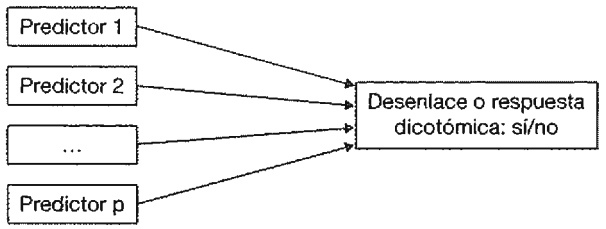



**Figura 12.3** Aplicación de la regresión logística.

Utilizando instrucciones:

```
regress    y x1 x2 x3        xp
```

siendo *y* la variable dependiente y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

## 12.3. PRIMERA APROXIMACIÓN AL MODELO DE REGRESIÓN LOGÍSTICA

Se utilizará la regresión logística cuando se disponga de *una variable dependiente dicotómica* (**10-12**). Esta situación es muy frecuente, ya que, a menudo, en la investigación biomédica o epidemiológica se desea identificar los predictores de la aparición de un determinado fenómeno, de que ocurra o no un suceso (p. ej., estar sano o enfermo, aprobar el MIR o no aprobarlo, etc.). Todas las variables que son candidatas a predecir la ocurrencia de ese fenómeno se utilizarán como variables independientes en un modelo de regresión logística, como muestra la figura 12.3.

La ecuación de la regresión logística es:

$$\ln\left(\frac{p}{1-p}\right) = a + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

Puede apreciarse su semejanza con la regresión múltiple. En este modelo de regresión siguen obteniéndose tantos coeficientes de regresión *b* como variables independientes se incluyan en el modelo matemático.

La diferencia con el modelo de regresión múltiple es que ahora se ha sustituido la variable dependiente Y por otra expresión. En la regresión logística, la variable dependiente no tiene un sentido numérico en sí misma, sino que es el logaritmo neperiano (ln) de la probabilidad *(p)* de que ocurra un suceso, dividido por la probabilidad de que no ocurra *(1 – p)*. El cociente *p/1 – p* se conoce en inglés como *odds*, que se ha traducido por «ventaja».

$$Odds = \frac{p}{1-p}$$

Una *odds* se calcula dividiendo el número de individuos que tienen una característica por el número de quienes no la tienen. Si en un estudio hay 50 pacientes reclutados en un centro de salud y 25 que no proceden de un centro de salud (son de un hospital), la *odds* de proceder del centro de salud es 2. Esto significa que hay el doble de pacientes que vienen del centro de salud que del hospital.

$$Odds_{\text{Centro de Salud}} = \frac{\text{n.º pacientes del centro de salud}}{\text{n.º pacientes que no son del centro de salud}} = \frac{50}{25} = 2$$

Por tanto, para calcular una *odds* basta con dividir el número de individuos con la característica de interés por el número de individuos que carecen de ella.

### 12.3.1. La *odds ratio*

¿Qué es una *odds ratio* (OR)? Una OR es una medida de asociación entre dos variables (X e Y) y consiste, tal y como sugiere su nombre, en un cociente o razón entre dos *odds*. Además, esta medida de asociación (o efecto) es la que se obtiene cuando se aplica un modelo de regresión logística.

Imagine el modelo más sencillo de regresión logística. Se intenta predecir un determinado suceso Y a partir de una sola variable independiente dicotómica X.

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds) = a + bx$$

A través de un modelo de regresión logística podría determinarse, por ejemplo, cuál es la *odds* (probabilidad/1 – probabilidad) de desarrollar cáncer de pulmón (variable Y = *cpulmón*; 0 = no; 1 = sí) según se fume o no se fume (variable X = *fumar*; 0 = no fumador; 1 = fumador).

Como la variable independiente es dicotómica, solo puede tomar dos valores (0 y 1), por lo que únicamente existirán dos funciones logísticas. Según se fume o no se fume, el aspecto de esta función logística variará:

Si el sujeto fuma:

$$\ln\left(\frac{p_{\text{cáncer}}}{1-p_{\text{cáncer}}}\right) = \ln(odds_{\text{cáncer}}) = a + b*1 = a + b \qquad \textit{Odds} \text{ de cáncer} = e^{a+b}$$

Si el sujeto no fuma:

$$\ln\left(\frac{p_{\text{cáncer}}}{1-p_{\text{cáncer}}}\right) = \ln(odds_{\text{cáncer}}) = a + b*0 = a \qquad \textit{Odds} \text{ de cáncer} = e^{a}$$

¿Podría determinarse cuál es la asociación entre el tabaco y el cáncer de pulmón? Una forma de hacerlo es comparar las *odds* obtenidas para cada supuesto (ser y no ser fumador).

Si se restaran las dos ecuaciones obtenidas:

$$\ln(p_{\text{cáncer|fumar}}/1-p_{\text{cáncer|fumar}}) - \ln(p_{\text{cáncer|no fumar}}/1-p_{\text{cáncer|no fumar}}) = a + b - a = b$$

$$\ln\left(\frac{p_{\text{cáncer|fumar}}/1-p_{\text{cáncer|fumar}}}{p_{\text{cáncer|no fumar}}/1-p_{\text{cáncer|no fumar}}}\right) = \ln\left(\frac{odds_{\text{cáncer|fumar}}}{odds_{\text{cáncer|no fumar}}}\right) = b$$

Como se ha indicado, el cociente entre dos *odds* se denomina OR, por lo que la ecuación anterior podría reescribirse como:

$$\ln\left(\frac{odds_{\text{cáncer|fumar}}}{odds_{\text{cáncer|no fumar}}}\right) = \ln(\text{OR}) = b \qquad \text{OR} = e^{b}$$

De hecho, la OR sirve para comparar dos *odds* (según se dé o no la variable X, fumar en este ejemplo): si realmente el tabaco no se asociara con el cáncer de pulmón, las dos *odds* serían iguales, *b* valdría 0 (compruébese la similitud con la regresión múltiple) y el valor de la OR sería 1. Si el tabaco se asociara de forma directa con el cáncer, *b* sería superior a 0 y la OR > 1. Si se asociara de forma inversa, *b* sería menor que 0 y la OR < 1.

El modelo de regresión logística puede obtenerse a través del menú de STATA con:

**Statistics → Binary outcomes → Logistic regression**

Se obtendrá así el valor de los coeficientes de regresión asociados a cada variable independiente.

**Statistics → Binary outcomes → Logistic regression (reporting odds ratios)**

Producirá el valor de las OR asociadas a cada predictor X.

Si en vez de usar el menú se acudiera a las instrucciones de STATA, se escribiría:

```
logit    y x1 x2 x3      xp        #para obtener b
logistic y x1 x2 x3      xp        #para obtener OR
```

siendo $y$ la variable dependiente y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

## 12.4. PRIMERA APROXIMACIÓN AL MODELO DE REGRESIÓN DE COX

La regresión de Cox, también llamada modelo de riesgos proporcionales *(proportional hazards model)*, es una técnica muy difundida (**12-17**). Su uso está indicado cuando la variable dependiente está relacionada con la supervivencia de un grupo de sujetos o, en general, con el tiempo que transcurre hasta que se produce en ellos un suceso o evento. Como ocurre con otras técnicas de análisis de supervivencia (Kaplan-Meier, *log-rank*), el evento de interés no tiene por qué ser la muerte. También puede ser otro tipo de suceso, como, por ejemplo, el fallo de una prótesis, la incidencia de una enfermedad o la ocurrencia de una complicación en quien padece ya una patología de base. Lo importante es que se trate de un suceso que ocurra una sola vez como máximo. Para sucesos o complicaciones que se repiten más veces en algunos pacientes durante el seguimiento, la regresión de Cox sería dudosamente válida y probablemente habría que recurrir a la regresión de Poisson (**12**; v. apartado 12.6).

La regresión de Cox se usa para valorar simultáneamente el efecto independiente de una serie de variables explicativas o factores pronósticos sobre la supervivencia (es decir, sobre la tasa de mortalidad) o sobre la tasa de ocurrencia de otro fenómeno que vaya apareciendo tras un período de tiempo variable en cada sujeto (fig. 12.4). Esta regresión es la extensión multivariable del análisis de supervivencia para evaluar de manera general variables dependientes del tipo «tiempo hasta un suceso o evento» y usa modelos de regresión, próximos al modelo de regresión logística. El modelo de regresión de Cox también permite predecir las probabilidades de supervivencia (o, en general, de permanencia libre del evento) para un determinado sujeto a partir del patrón de valores que presenten sus variables pronósticas.

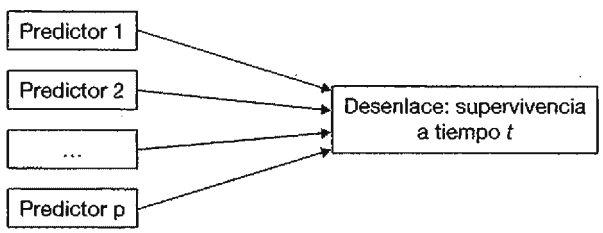



**Figura 12.4** Aplicación de la regresión de Cox.

Por tanto, para calcular una *odds* basta con dividir el número de individuos con la característica de interés por el número de individuos que carecen de ella.

### 12.3.1. La *odds ratio*

¿Qué es una *odds ratio* (OR)? Una OR es una medida de asociación entre dos variables (X e Y) y consiste, tal y como sugiere su nombre, en un cociente o razón entre dos *odds*. Además, esta medida de asociación (o efecto) es la que se obtiene cuando se aplica un modelo de regresión logística.

Imagine el modelo más sencillo de regresión logística. Se intenta predecir un determinado suceso Y a partir de una sola variable independiente dicotómica X.

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds) = a + bx$$

A través de un modelo de regresión logística podría determinarse, por ejemplo, cuál es la *odds* (probabilidad/1 – probabilidad) de desarrollar cáncer de pulmón (variable Y = *cpulmón*; 0 = no; 1 = sí) según se fume o no se fume (variable X = *fumar*; 0 = no fumador; 1 = fumador).

Como la variable independiente es dicotómica, solo puede tomar dos valores (0 y 1), por lo que únicamente existirán dos funciones logísticas. Según se fume o no se fume, el aspecto de esta función logística variará:

Si el sujeto fuma:

$$\ln\left(\frac{p_{\text{cáncer}}}{1-p_{\text{cáncer}}}\right) = \ln(odds_{\text{cáncer}}) = a + b*1 = a + b \qquad \textit{Odds} \text{ de cáncer} = e^{a+b}$$

Si el sujeto no fuma:

$$\ln\left(\frac{p_{\text{cáncer}}}{1-p_{\text{cáncer}}}\right) = \ln(odds_{\text{cáncer}}) = a + b*0 = a \qquad \textit{Odds} \text{ de cáncer} = e^{a}$$

¿Podría determinarse cuál es la asociación entre el tabaco y el cáncer de pulmón? Una forma de hacerlo es comparar las *odds* obtenidas para cada supuesto (ser y no ser fumador).

Si se restaran las dos ecuaciones obtenidas:

$$\ln(p_{\text{cáncer}|\text{fumar}}/1-p_{\text{cáncer}|\text{fumar}}) - \ln(p_{\text{cáncer}|\text{no fumar}}/1-p_{\text{cáncer}|\text{no fumar}}) = a + b - a = b$$

$$\ln\left(\frac{p_{\text{cáncer}|\text{fumar}}/1-p_{\text{cáncer}|\text{fumar}}}{p_{\text{cáncer}|\text{no fumar}}/1-p_{\text{cáncer}|\text{no fumar}}}\right) = \ln\left(\frac{odds_{\text{cáncer}|\text{fumar}}}{odds_{\text{cáncer}|\text{no fumar}}}\right) = b$$

Como se ha indicado, el cociente entre dos *odds* se denomina OR, por lo que la ecuación anterior podría reescribirse como:

$$\ln\left(\frac{odds_{\text{cáncer}|\text{fumar}}}{odds_{\text{cáncer}|\text{no fumar}}}\right) = \ln(\text{OR}) = b \qquad \text{OR} = e^{b}$$

De hecho, la OR sirve para comparar dos *odds* (según se dé o no la variable X, fumar en este ejemplo): si realmente el tabaco no se asociara con el cáncer de pulmón, las dos *odds* serían iguales, *b* valdría 0 (compruébese la similitud con la regresión múltiple) y el valor de la OR sería 1. Si el tabaco se asociara de forma directa con el cáncer, *b* sería superior a 0 y la OR > 1. Si se asociara de forma inversa, *b* sería menor que 0 y la OR < 1.

El modelo de regresión logística puede obtenerse a través del menú de STATA con:

**Statistics → Binary outcomes → Logistic regression**

Se obtendrá así el valor de los coeficientes de regresión asociados a cada variable independiente.

**Statistics → Binary outcomes → Logistic regression (reporting odds ratios)**

Producirá el valor de las OR asociadas a cada predictor X.

Si en vez de usar el menú se acudiera a las instrucciones de STATA, se escribiría:

```
logit    y x1 x2 x3      xp        #para obtener b
logistic y x1 x2 x3      xp        #para obtener OR
```

siendo *y* la variable dependiente y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

## 12.4. PRIMERA APROXIMACIÓN AL MODELO DE REGRESIÓN DE COX

La regresión de Cox, también llamada modelo de riesgos proporcionales *(proportional hazards model)*, es una técnica muy difundida (**12-17**). Su uso está indicado cuando la variable dependiente está relacionada con la supervivencia de un grupo de sujetos o, en general, con el tiempo que transcurre hasta que se produce en ellos un suceso o evento. Como ocurre con otras técnicas de análisis de supervivencia (Kaplan-Meier, *log-rank*), el evento de interés no tiene por qué ser la muerte. También puede ser otro tipo de suceso, como, por ejemplo, el fallo de una prótesis, la incidencia de una enfermedad o la ocurrencia de una complicación en quien padece ya una patología de base. Lo importante es que se trate de un suceso que ocurra una sola vez como máximo. Para sucesos o complicaciones que se repiten más veces en algunos pacientes durante el seguimiento, la regresión de Cox sería dudosamente válida y probablemente habría que recurrir a la regresión de Poisson (**12**; v. apartado 12.6).

La regresión de Cox se usa para valorar simultáneamente el efecto independiente de una serie de variables explicativas o factores pronósticos sobre la supervivencia (es decir, sobre la tasa de mortalidad) o sobre la tasa de ocurrencia de otro fenómeno que vaya apareciendo tras un período de tiempo variable en cada sujeto (fig. 12.4). Esta regresión es la extensión multivariable del análisis de supervivencia para evaluar de manera general variables dependientes del tipo «tiempo hasta un suceso o evento» y usa modelos de regresión, próximos al modelo de regresión logística. El modelo de regresión de Cox también permite predecir las probabilidades de supervivencia (o, en general, de permanencia libre del evento) para un determinado sujeto a partir del patrón de valores que presenten sus variables pronósticas.

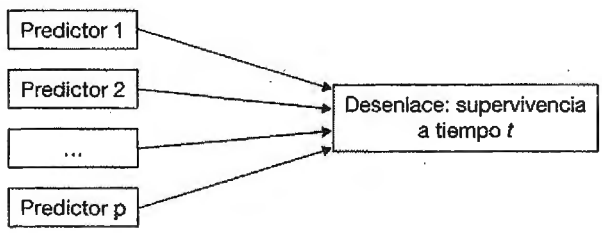



**Figura 12.4** Aplicación de la regresión de Cox.

Debe tenerse en cuenta que la regresión de Cox asume algunos de los mismos supuestos que el método de Kaplan-Meier: el suceso debe ser irreversible, ha de ocurrir una sola vez y la censura no debe ser informativa.

La ecuación de la regresión de Cox es:

$$\ln(\lambda_t) = a + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

Puede verse que, salvo el cambio de la variable dependiente, lo demás es bastante parecido al análisis de regresión logística. En la regresión logística, la variable de respuesta o dependiente era el ln *(odds)*, mientras que en la de Cox la respuesta depende del tiempo y la variable dependiente es el logaritmo del *hazard* ($\lambda$) o logaritmo de la tasa instantánea del evento.

El término *hazard* corresponde a una tasa instantánea, que conceptualmente solo requiere una duración de tiempo infinitesimal (instantánea) para que ocurra el suceso. La tasa se diferencia del riesgo en que tiene en cuenta el tiempo (fallecimientos por unidad de tiempo), mientras que el riesgo es una proporción y solo considera el número de sujetos inicialmente en riesgo de fallecer. La tasa instantánea o *hazard* de morir en el instante «t» se obtendría, según el modelo de Cox antes visto al tomar antilogaritmos:

$$\lambda_t = e^{a+b_1x_1+b_2x_2+\ldots+b_px_p} = e^a \times e^{b_1x_1+b_2x_2+\ldots+b_px_p}$$

Se denomina $\lambda_{0t}$ a la exponencial de «a», que sería la ordenada en el origen.

$$e^a = \lambda_{0t}$$

Se obtiene la siguiente expresión:

$$\lambda_t = \lambda_{0t}e^{b_1x_1+b_2x_2+\ldots+b_px_p}$$

La primera cantidad de la parte derecha de la ecuación, $\lambda_{0t}$, es análoga a la ordenada en el origen ya vista en otros modelos de regresión, y es la tasa *(hazard)* basal cuando todas las variables independientes $X_i$ valen 0.

$$\text{Si } x_1 = 0, x_2 = 0, \ldots x_i = 0 \rightarrow \lambda_t = \lambda_{0t}$$

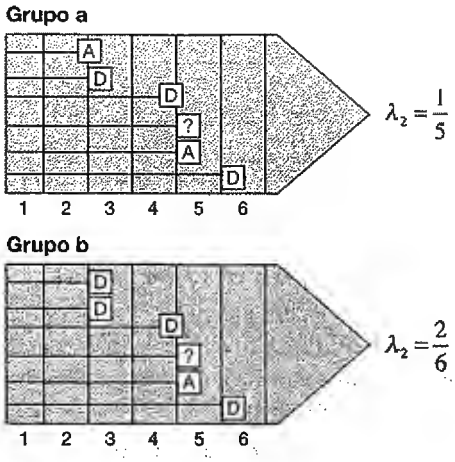
Esta tasa no es una constante, sino que depende del tiempo, de ahí el subíndice *t*. La supervivencia en el tiempo *t* ($S_t$) no es una cantidad numérica única, sino que varía con el tiempo. Téngase en cuenta que lo mismo sucede en el modelo de Cox con las tasas instantáneas o *hazards* en las que se basa. Tanto $\lambda_t$ como $\lambda_{0t}$ variarán a lo largo del tiempo de observación o seguimiento durante el cual se prolongue el estudio.

### 12.4.1. *Hazard ratio*

¿Qué es una *hazard ratio* (HR)? Una HR es una medida de asociación entre dos variables (X e Y) y consiste, tal y como sugiere su nombre, en un cociente o razón entre dos *hazards*. Además, esta medida de asociación (o efecto) es la que se obtiene cuando se aplica un modelo de regresión de Cox.

El planteamiento es muy parecido al utilizado en la regresión logística, aunque ahora la comparación no es entre *odds*, sino entre *hazards*.

Se estudiará con un ejemplo recogido en la figura 12.5. En esta figura se representan dos grupos de pacientes (*a* y *b*). En cada grupo hay seis pacientes, objeto de seguimiento hasta un máximo de 5 años, si bien el tiempo de seguimiento varía de uno a otro paciente; el seguimiento de cada paciente se representa por una línea horizontal: una D significa el momento en que ocurre la muerte de un paciente y una A señala el final del seguimiento para un paciente que se encontraba vivo al término del estudio. Se usa el signo de interrogación para aquellos pacientes que se perdieron para los que la última noticia que se tiene de ellos es que seguían vivos.




| Grupo a | | Grupo b | |
|---|---|---|---|
| *Años* | *Estado* | *Años* | *Estado* |
| 1,5 | Vivo | 2 | Muerto |
| 2 | Muerto | 2 | Muerto |
| 3,5 | Muerto | 3,5 | Muerto |

**Figura 12.5** Cálculo del *hazard* a 2 años ($\lambda_2$) en dos grupos (*a* y *b*) de seis pacientes.

En el grupo *a* se ha producido una muerte a los 2 años y había cinco pacientes en riesgo de morir en ese momento (el primero solo había permanecido 1,5 años en el estudio y, por eso, a los 2 años ya no estaba «en riesgo»). El *hazard* se calcula simplemente dividiendo los sucesos ocurridos en ese instante (tiempo = 2 años) entre el total de sujetos en riesgo.

$$hazard_t = \lambda_t = \frac{\text{sucesos ocurridos en el instante t}}{\text{sujetos en riesgo en el instante t}} = \frac{d_t}{n_t}$$

$$\text{Grupo } a \rightarrow hazard_{2\text{años}} = \lambda_{2a} = \frac{d_2}{n_2} = \frac{1}{5} = 0{,}2$$

$$\text{Grupo } b \rightarrow hazard_{2\text{ años}} = \lambda_{2b} = \frac{d_2}{n_2} = \frac{2}{6} = 0{,}33$$

El *hazard* a los 2 años será 0,2 en el grupo *a* y 0,33 en el grupo *b*.

Para comparar ambos grupos (*a* y *b*) se dividirá un *hazard* entre otro (tal y como se hacía con las *odds* en el modelo de regresión logística). Si se toma como referencia el grupo *a*, la HR para el grupo *b* con respecto al *a* será:

$$\textit{Hazard ratio} = \text{HR} = \frac{hazard_b}{hazard_a} = \frac{0{,}33}{0{,}2} = 1{,}67$$

El grupo *b* tiene una mortalidad que es 1,67 veces mayor que la del *a;* es decir, la mortalidad es un 67% superior en *b*. Esta comparación solo se refiere a los 2 años. Si se comparasen a los 3,5 años, el HR sería 1. En una comparación a los 5 años también será 1. No se pueden hacer más comparaciones, pues en esta base de datos solo hay muertes a los 2, 3,5 y 5 años. Obviamente, si no se observan muertes no se pueden comparar sus riesgos de mortalidad. Lo que hace el modelo de Cox es promediar de manera ponderada las HR de los diversos momentos en los que acontece alguna muerte. Viene a ser como hacer muchas regresiones logísticas, una para cada momento en que se observa algún fallecimiento.

El modelo de regresión de Cox puede obtenerse a través del menú de STATA, empleando:

**Statistics → Survival analysis → Regression models → Cox proportional hazards model**

o con las instrucciones:

```
stset tiempo_seguimiento, failure(evento==1)
stcox y x1 x2 x3   xp
```

siendo $y$ la variable dependiente y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

## 12.5. ASPECTOS COMUNES Y DIFERENCIALES DE LOS MODELOS LINEAL, LOGÍSTICO Y DE COX

1. Los tres modelos son funciones matemáticas que pueden incluir más de una variable independiente (predictora, variable X) y una sola variable dependiente Y. Las variables predictoras solo pueden ser introducidas en el modelo como variables cuantitativas o cualitativas dicotómicas. Sin embargo, los tres modelos difieren en las características de la variable dependiente Y.

   En el modelo de regresión lineal múltiple, esta variable es cuantitativa, como sería el índice de masa corporal (IMC) (kg/m$^2$). De esta forma puede observarse cómo influye, por ejemplo, la práctica o no de actividad física (variable X dicotómica) en este índice de adiposidad (variable Y cuantitativa). El efecto, por tanto, se mide en escala aditiva.

   En la regresión logística, la variable dependiente es cualitativa dicotómica (es decir, no se valoran cambios en el IMC según los valores que tome X, sino, por ejemplo, sobre la probabilidad de ser o no ser obeso o de tener o no un IMC $\geq$ 30 kg/m$^2$). De igual forma, podría valorarse el efecto de practicar o no actividad física (variable X dicotómica), si bien en este caso sobre la probabilidad (mejor dicho, sobre la *odds*) de presentar obesidad (variable Y cualitativa dicotómica). Si, además de recoger si el sujeto engordará o no tras realizar o no práctica deportiva, se recaba información acerca de en qué momento del tiempo lo hará (velocidad con que engordará), en ese caso se aplicará un modelo de regresión de Cox. Los modelos de regresión logística y de Cox se mueven, por tanto, en escala mutiplicativa.

2. Los tres modelos pueden construirse con diferentes finalidades:
   a. Predecir los valores de la variable Y a partir de los valores que toman diferentes variables X ($X_1 \ldots X_p$).
   b. Determinar la asociación entre dos variables ($X_1$ e Y) independientemente del valor que tomen otras variables X ($X_2 \ldots X_p$) (controlar la confusión).
   c. Servir para detectar y describir la posible interacción entre variables X ($X_1 \ldots X_p$) sobre la variable Y.
   d. Además, los tres utilizan los mismos procedimientos para valorar la confusión (se controla el factor de confusión añadiéndolo al modelo) y la interacción (se crean términos multiplicativos que son incorporados al modelo).

3. En los tres modelos se calculan tantos coeficientes de regresión $b$ como variables independientes se introduzcan en el modelo, aunque su interpretación solo es directa en la regresión múltiple. En las regresiones logística y de Cox, el valor práctico radica en el exponencial del coeficiente ($e^b$), que se traduce como OR y HR, respectivamente.

Además, en la regresión lineal, los coeficientes son estimados a través del método de mínimos cuadrados. En cambio, en la regresión logística y en la de Cox no sirve el método de los mínimos cuadrados. En estos dos últimos casos, los parámetros son estimados mediante el método de máxima verosimilitud *(maximum likelihood)*.

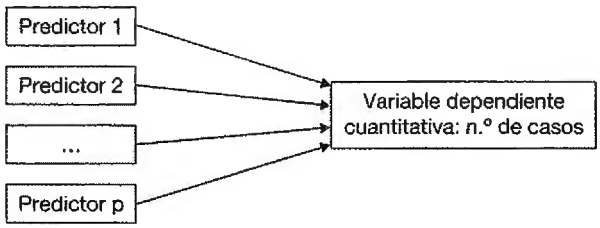



**Figura 12.6** Aplicación de la regresión de Poisson.

## 12.6. REGRESIÓN DE POISSON

La regresión de Poisson es un modelo matemático que se utiliza cuando la variable dependiente Y es una variable cuantitativa discreta que no puede tomar valores negativos (p. ej., número de casos de un determinado evento). Se usa para valorar simultáneamente el efecto independiente de una serie de variables explicativas o factores pronósticos sobre la velocidad de ocurrencia de un determinado fenómeno (número de casos producidos en un período de tiempo dado y entre un número determinado de sujetos susceptibles de sufrirlo) (fig. 12.6).

La ecuación de la regresión de Poisson se expresa como:

$$\ln(DI) = a + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

donde ln(DI) es el logaritmo neperiano de la densidad de incidencia, $b_1$-$b_p$ son los coeficientes de regresión y $x_1$-$x_p$ son las variables independientes o predictoras.

¿Qué es la densidad de incidencia (DI)? La tasa de incidencia o densidad de incidencia es una medida de frecuencia de enfermedad que expresa la velocidad con la que ocurre un determinado evento. Por ejemplo: cinco eventos por segundo, 20 eventos por 1.000 personas-año[3], etc. (9).

Su fórmula general es:

$$DI = \frac{\text{n.º eventos}}{\text{personas*tiempo}}$$

Imagine un ejemplo sencillo, en el que quiere valorarse si existen diferencias en las tasas de enfermedad (velocidad con la que se enferma) de acuerdo con el consumo o no de tabaco (0 = no fumador, 1 = fumador).

$$\ln(DI) = a + b * tabaco$$

Se podrían crear dos modelos, uno para no fumadores y otro para fumadores.

Para fumadores: $\ln(DI) = a + b$

Para no fumadores: $\ln(DI) = a$

Si se restaran las dos ecuaciones, se obtendría:

$$\ln(DI)_{\text{fumador}} - \ln(DI)_{\text{no fumador}} = a + b - a = b$$

$$\ln\left(\frac{DI_{\text{fumador}}}{DI_{\text{no fumador}}}\right) = \ln(RDI) = b$$

3 Veinte eventos por 1.000 personas-año se traduciría como la existencia, después de 1 año de seguimiento, de 20 casos de enfermedad de un total de 1.000 personas; o también, después de 2 años de seguimiento, la existencia de 20 casos de enfermedad de un total de 500 personas; o bien, después de 4 años de seguimiento, la ocurrencia de 20 casos de un total de 250 personas. Es decir, considera no solo el número de casos que se producen sobre el total de personas que pueden sufrir dicho evento, sino el tiempo en el que son susceptibles de sufrirlo.

*Tabla 12.3 Formato que suelen tener los datos cuando se aplica una regresión de Poisson*

| | SEXO | TABACO | DEPORTE | CASOS | PERS_MES | DI |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 8 | 100 | 0,08 |
| 2 | 1 | 1 | 0 | 17 | 100 | 0,17 |
| 3 | 1 | 0 | 1 | 6 | 100 | 0,06 |
| 4 | 1 | 1 | 1 | 12 | 100 | 0,12 |
| 5 | 0 | 0 | 0 | 7 | 100 | 0,07 |
| 6 | 0 | 1 | 0 | 11 | 100 | 0,11 |
| 7 | 0 | 0 | 1 | 3 | 100 | 0,03 |
| 8 | 0 | 1 | 1 | 6 | 100 | 0,06 |

*DI, densidad de incidencia.*

donde RDI es la razón de densidad de incidencia. La razón de densidad de incidencia es una medida de asociación entre una exposición X (tabaco en este caso) y un desenlace Y (casos de enfermedad). Si RDI > 1, la exposición es dañina, lo cual significaría que se enferma más deprisa en el grupo expuesto al tabaco (en el mismo período de tiempo y sobre el mismo número de personas, se observan más casos de enfermedad) que en el grupo no expuesto. Si RDI < 1, la exposición resultaría protectora. Si RDI fuera igual a 1, el tabaco no se asociaría con la tasa de enfermedad.

Se cumple que:

$$\text{Razón de densidad de incidencia} = e^{b}$$

Véase en un ejemplo concreto. Se observaron 96 casos de una enfermedad en 800 personas-mes. Su distribución por sexo, tabaco y deporte se muestra la tabla 12.3. Las variables predictoras o independientes (X) son, en este caso, el *sexo* (0 = mujer, 1 = varón), el *tabaco* (0 = no fumador, 1 = fumador) y el *deporte* (0 = no deporte, 1 = deporte). En esa tabla, cada fila no corresponde a una persona, sino a un grupo de personas definidas por su sexo, su hábito tabáquico y su práctica de deporte. La variable dependiente es el total del número de casos de enfermedad que se han producido en cada grupo *(casos)*. Se ha simplificado el análisis otorgando un seguimiento de 100 personas-mes *(pers-mes)* a cada posible combinación de características. Así, existirán 100 pers-mes varones, no fumadores y no deportistas (fila 1), 100 pers-mes varones, fumadores y no deportistas (fila 2), etc. Se define 100 pers-mes como 100 personas seguidas durante 1 mes, 50 personas seguidas durante 2 meses, 25 personas seguidas durante 4 meses o incluso como una persona seguida durante 100 meses. Con los datos de la tabla, puede calcularse la DI, es decir, el número de casos de enfermedad que se observan *(casos)* entre las personas en riesgo de enfermar y sus tiempos en riesgo *(pers_mes)*. Se obtendrán tantas DI como posibles combinaciones de variables X se dispongan (8, en este ejemplo).

De la misma forma, pueden hallarse las RDI para la asociación entre cada variable X y la variable dependiente Y. Habrá así una RDI para el sexo (se calculará quiénes enferman más rápido, hombres o mujeres), otra para el tabaco (se calculará si enferman antes los fumadores o los no fumadores) y una última para el deporte.

$$RDI_{sexo} = \left(\frac{DI_{hombres}}{DI_{mujeres}}\right) = \frac{(8+17+6+12)/400}{(7+11+3+6)/400} = \frac{0,1075}{0,0675} = 1,593$$

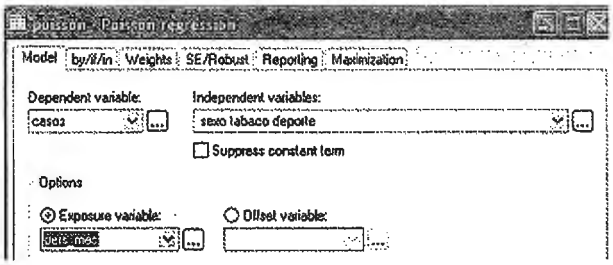
$$RDI_{tabaco} = \left(\frac{DI_{fumador}}{DI_{no\ fumador}}\right) = \frac{(17+12+11+6)/400}{(8+6+7+3)/400} = 1,917$$

$$RDI_{deporte} = \left(\frac{DI_{deporte}}{DI_{no\ deporte}}\right) = \frac{27/400}{43/400} = 0,628$$

Según estos resultados, enferman más rápido los varones que las mujeres, los fumadores que los no fumadores y quienes no practican deporte comparados con los que sí lo practican.

La regresión de Poisson puede llevarse a cabo a través del programa STATA utilizando el menú o a través de instrucciones. Con el menú:

**Statistics → Count outcomes → Poisson regression**




Utilizando instrucciones:

**poisson *y* $x_1$ $x_2$ $x_3$ $x_p$, exposure (*personas-tiempo*)**

siendo *y* la variable dependiente y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

En el ejemplo:

Con esta instrucción STATA devolverá un listado en el que solo se expresarán los coeficientes de regresión (Coef.). Un coeficiente para cada variable independiente X. Recuérdese que la RDI asociada al sexo, por ejemplo, se calcularía como $e^{0,465}$

```
. poisson casos sexo tabaco deporte, exposure (pers_mes)

Iteration 0:   log likelihood = -16.034759
Iteration 1:   log likelihood = -16.034759

Poisson regression                              Number of obs     =          8
                                                LR chi2(3)        =      14.41
                                                Prob > chi2       =     0.0024
Log likelihood = -16.034759                     Pseudo R2         =     0.3101

------------------------------------------------------------------------------
       casos |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |   .4653632    .245546     1.90   0.058    -.0158981    .9466246
     tabaco |   .6505876   .2518051     2.58   0.010     .1570587    1.144116
    deporte |  -.4653632    .245546    -1.90   0.058    -.9466246    .0158981
      _cons |  -2.86707    .2709024   -10.58   0.000    -3.398029   -2.336111
   pers_mes |  (exposure)
------------------------------------------------------------------------------
```

Para obtener RDI deberá especificarse la instrucción **irr**
poisson *y* $x_1$ $x_2$ $x_p$, **exposure**(*personas-tiempo*) **irr**
*irr: incidence rate ratio*

```
. poisson casos sexo tabaco deporte, exposure (pers_mes) irr

Iteration 0:   log likelihood = -16.034759
Iteration 1:   log likelihood = -16.034759

Poisson regression                              Number of obs     =          8
                                                LR chi2(3)        =      14.41
                                                Prob > chi2       =     0.0024
Log likelihood = -16.034759                     Pseudo R2         =     0.3101

------------------------------------------------------------------------------
       casos |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sexo |   1.592593   .3910548     1.90   0.058     .9842276    2.576997
     tabaco |   1.916667   .4826264     2.58   0.010     1.170064    3.139666
    deporte |    .627907   .1541801    -1.90   0.058     .3880486    1.016025
   pers_mes |  (exposure)
------------------------------------------------------------------------------
```

***Interpretación:*** *en el primer listado, STATA produce las estimaciones de los coeficientes $b_i$. Puede comprobarse que, utilizando estos coeficientes como exponentes del número e, se obtienen las razones de tasas o de densidad de incidencia. Por ejemplo, Exp (0,6506) = 1,917 (RDI de fumadores comparados con no fumadores).*

Como en otros modelos de regresión, cada coeficiente viene seguido de su respectivo error estándar (Std. Err.). Dividiendo el coeficiente por su error estándar se obtiene un valor z que sigue una distribución normal (empieza a ser significativo al 5% a dos colas cuando $z > 1,96$). La siguiente columna ($P > |z|$) corresponde al valor p de significación estadística a dos colas. Por último, se presentan los intervalos de confianza al 95% para los coeficientes.

El segundo listado aporta directamente las RDI asociadas a cada variable independiente (*sexo, tabaco* y *deporte*), así como sus errores estándar, su nivel de significación estadística (valor *p*) y su intervalo de confianza al 95%. Debe señalarse que estas RDI están ajustadas. Es decir, cuando se valora el efecto del tabaco sobre la velocidad de enfermar, este efecto es independiente del sexo o de la práctica o no de ejercicio físico por parte del participante, ya que el modelo incluye las tres variables X en sus estimaciones. En este caso, las estimaciones ajustadas coincidirían con las obtenidas «a mano» directamente de la tabla 12.3. Ninguna de las variables X actúa como factor de confusión (v. apartado 12.16).

El modelo de Poisson, como todo modelo de regresión, sirve para hacer predicciones de riesgos (tasas en este caso) absolutos. Para predecir la tasa (DI) de enfermedad de un varón, fumador y no deportista, se utilizaría la ecuación:

$$\ln(DI_{varón,fumador,nodeportista}) = -2,867 + 0,4654 * 1 + 0,6506 * 1 - 0,4653 * 0 = -1,328$$
$$DI_{varón,fumador,nodeportista} = \text{Exp}(-1,328) = 0,17$$

Puede observarse que el valor de DI observado (17/100 pers_mes) coincide con lo predicho por el modelo.

## 12.7. OTROS MÉTODOS MULTIVARIANTES

### 12.7.1. MANOVA

El MANOVA o análisis multivariante de la varianza no solo puede tener en cuenta muchas variables independientes, sino que usa, además, varias variables dependientes que, de algún modo, miden la misma respuesta desde distintos puntos de vista. Por ejemplo, si se desea ver predictores de adiposidad, en vez de usar como respuesta o variable dependiente solamente el IMC se puede decidir usar tres variables de respuesta: IMC, índice cintura-cadera y grosor de pliegues cutáneos. El MANOVA permite manejar a la vez, en un solo análisis, varias variables dependientes. Lógicamente, también valorará una serie de predictores o variables independientes (en el ejemplo, edad, actividad física, hábitos alimentarios, variantes genéticas, etc.).

El análisis de MANOVA puede ser realizado con el programa STATA, bien a través del menú del programa o bien con instrucciones. Con el menú:

**Statistics → Multivariate analysis → MANOVA, multivariate regression, and related → MANOVA**

A través del uso de instrucciones:

**manova** $y_1\ y_2\ y_3 \quad y_p = x_1\ x_2\ x_3 \quad x_p$

siendo $y_1$-$y_p$ las variables dependientes (cuantitativas) y $x_1$-$x_p$ las variables independientes (cuantitativas o cualitativas dicotómicas).

### 12.7.2. Análisis factorial

En este análisis no hay una variable dependiente y muchas independientes que se usen para explicarla o predecirla, sino que pretende extraer de una base de datos con muchas variables un pequeño grupo de factores que consigan proporcionar de manera resumida gran parte de la información contenida en todas las variables iniciales. Es, por tanto, una técnica de reducción de variables. Existen diferentes aproximaciones para llevar a cabo este tipo de análisis. Una de las más comunes es el análisis factorial de componentes principales.

Desde el menú de STATA:

**Statistics → Multivariate analysis → Factor and Principal components analysis → Factor analysis**

Con instrucciones:

```
factor v1 v2 v3  vp, pcf
```

siendo $v_1$-$v_p$ las variables implicadas en el análisis.

### 12.7.3. Análisis de conglomerados o de clúster

Al igual que el análisis factorial, el análisis por conglomerados o de clúster es una técnica descriptiva que busca sintetizar los datos, pero, en vez de resumir el número de variables (columnas), tiende a formar grupos homogéneos de sujetos (vendría a ser como reducir las filas). Este análisis facilita la clasificación de sujetos, pues coloca en el mismo grupo a quienes tienen valores parecidos de un conjunto de variables. Un clúster o conglomerado es, por tanto, un grupo de sujetos próximos entre sí en el espacio multidimensional definido por las variables consideradas para su clasificación.

Desde el menú de STATA:

**Statistics → Multivariate analysis → Cluster analysis → Cluster data → Single linkage**

Con instrucciones:

```
cluster singlelinkage      v1 v2 v3      vp
```

siendo $v_1$-$v_p$ las variables implicadas en el análisis.

## 12.8. HIPÓTESIS NULAS EN UNA REGRESIÓN MÚLTIPLE

El modelo de regresión múltiple es una generalización a varias variables de un modelo de regresión simple. La regresión lineal múltiple se empleará cuando se desee estudiar cómo influyen varios factores (o variables independientes) en una variable de respuesta (la variable dependiente) que es cuantitativa, como, por ejemplo, la talla o el peso.

La ecuación de la regresión lineal *simple* es:

$$y = a + bx$$

donde Y es la variable dependiente y X es la independiente. Esta ecuación se puede generalizar para el caso en que haya más de una variable independiente. Supóngase que existen tres variables independientes: $X_1$, $X_2$ y $X_3$. Puede construirse la ecuación:

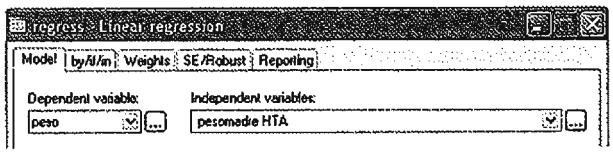
$$y = a + b_1x_1 + b_2x_2 + b_3x_3$$

Cada variable independiente $X_i$ tiene un coeficiente de regresión o pendiente propia $b_i$. Este coeficiente se interpretará como el cambio en la variable dependiente Y, por unidad de cambio en cada variable independiente ($X_1$, $X_2$ o $X_3$) *a igualdad de nivel de las otras variables independientes.*

Por ejemplo, supóngase que el peso (g) de una muestra de niños recién nacidos se utiliza como variable dependiente Y intentando predecirla a partir de dos variables independientes, $X_1$ y $X_2$, que corresponden, respectivamente, al peso de la madre en kg (*pesomadre*: $X_1$) y a la presencia de hipertensión arterial durante el embarazo (*HTA*: $X_2$; 0 = no; 1 = sí).

Este análisis puede ser realizado con STATA a través del menú:

**Statistics → Linear models and related → Linear regression**




O a través de la instrucción:

```
regress
```

```
. regress peso pesomadre HTA
```

| Source | SS | df | MS | Number of obs = | 189 |
|---|---|---|---|---|---|
| Model | 7265986.74 | 2 | 3632993.37 | F( 2, 186) = | 7.29 |
| Residual | 92651065.9 | 186 | 498124.01 | Prob > F = | 0.0009 |
| | | | | R-squared = | 0.0727 |
| | | | | Adj R-squared = | 0.0627 |
| Total | 99917052.6 | 188 | 531473.684 | Root MSE = | 705.78 |

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 12.36788 | 3.852452 | 3.21 | 0.002 | 4.767763 | 19.968 |
| HTA | -599.4701 | 216.6363 | -2.77 | 0.006 | -1026.85 | -172.0899 |
| _cons | 2864.785 | 285.2792 | 10.04 | 0.000 | 2301.986 | 3427.584 |


| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 12.36788 | 3.852452 | 3.21 | 0.002 | 4.767763 | 19.968 |
| HTA | -599.4701 | 216.6363 | -2.77 | 0.006 | -1026.85 | -172.0899 |
| _cons | 2864.785 | 285.2792 | 10.04 | 0.000 | 2301.986 | 3427.584 |


***Interpretación:*** *se obtiene la siguiente ecuación:*

$$y = 2.864,8 + 12,4x_1 - 599,5x_2$$

*Sustituyendo $X_i$ por sus nombres, se obtendrá:*

$$\text{Peso (g)} = 2.864,8 + 12,4 * \text{peso madre} - 599,5 * \text{HTA}$$

*La interpretación será que, por cada kg de peso adicional que presenta la madre, aumenta el peso del recién nacido en 12,4 g, independientemente de si la madre tuvo o no HTA durante el embarazo. Además, la diferencia de peso entre hijos de mujeres con y sin presencia de esta patología es de 599,5 g (pesarán más los hijos de madres sin HTA), a igualdad de peso de la madre, es decir, sea cual sea su peso.*

En una regresión múltiple se efectúan pruebas de hipótesis para cada variable, dividiendo su coeficiente *($b_i$)* por el error típico o error estándar (test de Wald) respectivo. Este estadístico sigue una *t de Student* con tantos grados de libertad como los de los residuales. La hipótesis nula para cada una de las pruebas de hipótesis es que la respectiva variable no mejora la predicción, *condicionalmente a que otras variables ya hayan sido tenidas en cuenta.* En el ejemplo puede concluirse que tanto el peso de la madre como la presencia de HTA durante el embarazo mejoran la predicción del peso del recién nacido una vez que se tiene en cuenta cada una de ellas.

***Tabla 12.4 La significación estadística en el modelo de regresión múltiple se debe interpretar como* condicional *a las otras variables***

| | COEFICIENTE DE REGRESIÓN | ERROR ESTÁNDAR | T | P |
|---|---|---|---|---|
| Constante (a) | −360,913 | 76,744 | −4,703 | 0,002 |
| Temperatura rectal | 4,424 | 6,646 | 0,666 | 0,527 |
| Temperatura axilar | 7,587 | 7,405 | 1,025 | 0,340 |

***Tabla 12.5 El mismo ejemplo de la Tabla 12.4, pero con dos modelos con las dos variables independientes por separado***

| | COEFICIENTE DE REGRESIÓN | ERROR ESTÁNDAR | T | P |
|---|---|---|---|---|
| **Modelo 1. Regresión lineal simple, variable independiente: temperatura rectal** | | | | |
| Constante (a) | −327,021 | 69,464 | −4,708 | 0,002 |
| Temp. rectal | 10,962 | 1,866 | 5,876 | <0,001 |
| **Modelo 2. Regresión lineal simple, variable independiente: temperatura axilar** | | | | |
| Constante (a) | −369,154 | 73,055 | −5,053 | 0,001 |
| Temp. axilar | 12,319 | 1,999 | 6,164 | <0,001 |

## 12.9. INTERPRETACIÓN CONDICIONAL DE LOS VALORES *p*

El siguiente ejemplo ilustra muy bien el significado de la interpretación condicional de los valores *p*. Imagine un listado de coeficientes, errores estándar, *t* y *p*, como el de la tabla 12.4 (**18**). Se trata de predecir la frecuencia cardíaca (pulsaciones/min) en unos niños, con o sin febrícula, a partir de la temperatura corporal. La temperatura se mide en dos localizaciones: rectal y axilar.

En una interpretación ingenua de este listado se diría que ni la temperatura rectal ni la axilar sirven para predecir la frecuencia cardíaca, pues sus respectivos valores *p* de significación estadística ($p = 0{,}527$ y $p = 0{,}340$) indican falta de evidencia para apoyar una asociación. No podría rechazarse la hipótesis nula, pero sería equivocado descartar que la temperatura fuese capaz de predecir la frecuencia cardíaca.

De hecho, el conocimiento médico apoya que la frecuencia cardíaca sí se puede predecir a partir de la temperatura. Si se ajustara una regresión lineal simple con estos mismos datos, los resultados serían los de la tabla 12.5, en la que se demuestra que tanto la temperatura rectal como la axilar son predictores significativos de la frecuencia cardíaca ($p < 0{,}001$ para ambas).

Lo que sucede en este ejemplo es que los valores *p* de la tabla 12.4 son *condicionales*. Responderían a la siguiente pregunta: una vez introducida en el modelo la temperatura axilar, ¿mejora la predicción si además se tiene en cuenta la temperatura rectal? La respuesta es No ($p = 0{,}527$). La segunda pregunta que puede formularse es: una vez tenida en cuenta la temperatura rectal, ¿mejora la predicción si se añade la temperatura axilar? La respuesta también es No ($p = 0{,}340$). No hay evidencia para rechazar la hipótesis nula en ninguno de los dos casos, pues los valores *p* no son significativos. El error que se comete con frecuencia al interpretar un listado como el de la tabla 12.4 es pensar equivocadamente que no hay relación alguna entre temperatura corporal y frecuencia cardíaca. No es verdad, sí existe relación, pero lo que sucede en el primer modelo (v. tabla 12.4) es que las dos variables están contando la misma historia.

Las hipótesis nulas que se contrastan en una regresión múltiple establecen que cada una de las variables no mejora la predicción del modelo *si el resto de variables ya se ha tenido en cuenta*. Esta idea no debe olvidarse nunca, porque todas las interpretaciones de los valores *p* en una regresión múltiple deben basarse en el principio de que son condicionales a que las otras variables ya estén en el modelo.

## 12.10. INTERVALOS DE CONFIANZA EN LA REGRESIÓN MÚLTIPLE

La significación estadística de un coeficiente de regresión puede establecerse, además, a partir de la observación de su intervalo de confianza. La fórmula del intervalo de confianza del coeficiente de regresión (*b*) es:

$$\text{IC } 95\%(b) = b \pm 1{,}96(\text{EE}_b)$$

siendo *b* el valor del coeficiente de regresión y $\text{EE}_b$ el valor de su error estándar.

Existirán diferencias estadísticamente significativas ($p < 0,05$) cuando el intervalo de confianza no incluya el valor nulo, 0 en este caso.

En el ejemplo del peso de los recién nacidos:

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 12.36788 | 3.852452 | 3.21 | 0.002 | 4.767763 | 19.968 |
| HTA | -599.4701 | 216.6363 | -2.77 | 0.006 | -1026.85 | -172.0899 |
| _cons | 2864.785 | 285.2792 | 10.04 | 0.000 | 2301.986 | 3427.584 |

***Interpretación:*** *por término medio, los hijos nacidos de madres que han sufrido HTA durante el embarazo pesarán 599,5 g menos, independientemente del peso que presente la madre. Sin embargo, este valor, con un nivel de confianza del 95%, oscilará entre 1.026,9 y 172,1 g. Es decir, el resultado es estadísticamente significativo. En la población, sea cual sea el peso de la madre, siempre pesarán menos los hijos de madres con HTA (obsérvese que la conclusión a la que se llega observando el intervalo de confianza coincide con la obtenida tras la observación del valor p, 0,006 en este ejemplo).*

## 12.11. COEFICIENTE DE DETERMINACIÓN $R^2$ Y SU VERSIÓN AJUSTADA

El coeficiente de determinación al cuadrado ($R^2$) es el coeficiente de determinación en una regresión y se interpreta como la proporción de la variabilidad de Y que puede ser explicada a partir de X. Cuando se añaden más variables $X_i$ al modelo, el coeficiente de determinación $R^2$ se incrementa, es decir, se incrementa el porcentaje de variabilidad de la variable Y que puede ser explicada por los predictores (de hecho, si se dispusiera de la información de infinitas variables predictoras X, se explicaría el 100% de la variabilidad de la variable dependiente). Pero ¿hasta qué punto ese incremento es real o artificial? La respuesta la da el $R^2$ ajustado.

El $R^2$ ajustado es el coeficiente de determinación, pero corregido al tener en cuenta que puede haber un número de predictores variable en el modelo. La observación del coeficiente de determinación ajustado será, por tanto, el estimador que establezca si la inclusión de nuevas variables en el modelo mejora o no la capacidad de predicción del mismo.

Se calcula con la siguiente fórmula:

$$R^2_{\text{ajustado}} = R^2 - \frac{p\left(1-R^2\right)}{N-1-p}$$

donde $p$ es el número de predictores independientes y $N$ el tamaño de la muestra.

Imagine que se pretende predecir la tensión arterial sistólica de un grupo de 33 sujetos a partir de su edad y sexo. Se encuentran los resultados recogidos en la tabla 12.6.

Como se observa, para cada modelo se obtienen dos coeficientes $R^2$. El primero equivale al coeficiente de correlación múltiple elevado al cuadrado (modelo 1: $0,658^2 = 0,432$; modelo 2: $0,665^2 = 0,442$) y el segundo al coeficiente de determinación al cuadrado, pero corregido (modelo 1: $R^2$ ajustado = 0,412; modelo 2: $R^2$ ajustado = 0,401).

***Tabla 12.6 Coeficientes de determinación para el ejemplo de la predicción de la tensión arterial a partir de la edad y el sexo***

| | R | R CUADRADO | R CUADRADO AJUSTADA* |
|---|---|---|---|
| Modelo 1 (variable independiente $X_1$: *edad*) | 0,658 | 0,432 | 0,412 |
| Modelo 2 (variables independientes $X_1$: *edad* y $X_2$: *sexo*) | 0,665 | 0,442 | 0,401 |

$$R^2_{\text{ajustado}} = R^2 - \frac{p\left(1-R^2\right)}{N-1-p}$$

Para el modelo que solo incluye edad:

$$R^2_{\text{ajustado}} = R^2 - \frac{p(1-R^2)}{N-1-p} = 0,432 - \frac{1(1-0,432)}{33-1-1} = 0,41$$

Para el modelo con edad y sexo:

$$R^2_{\text{ajustado}} = R^2 - \frac{p(1-R^2)}{N-1-p} = 0,442 - \frac{2(1-0,442)}{33-1-2} = 0,40$$

Se aprecia cómo, una vez que se corrige por la nueva variable introducida en el modelo (en el ejemplo se introduce el sexo creándose el modelo 2), el $R^2$ ajustado no aumenta, sino que disminuye. Eso sugiere que esta segunda variable, el sexo, no mejora la predicción de la tensión arterial obtenida utilizando los datos de edad de los participantes.

## 12.12. CONDICIONES DE APLICACIÓN DEL MODELO DE REGRESIÓN MÚLTIPLE. ANÁLISIS DE RESIDUALES Y VERIFICACIÓN DE SUPUESTOS

El procedimiento utilizado para calcular una regresión lineal simple es el ajuste por mínimos cuadrados. El objetivo es encontrar la ecuación que mejor se ajuste a los puntos observados. En una regresión múltiple, el procedimiento de estimación es semejante al utilizado en la regresión lineal simple; se estima la superficie que mejor se ajusta a la nube de puntos observados. El método, denominado ajuste por mínimos cuadrados, minimiza las distancias desde cada punto observado hasta el plano (residuales al cuadrado).

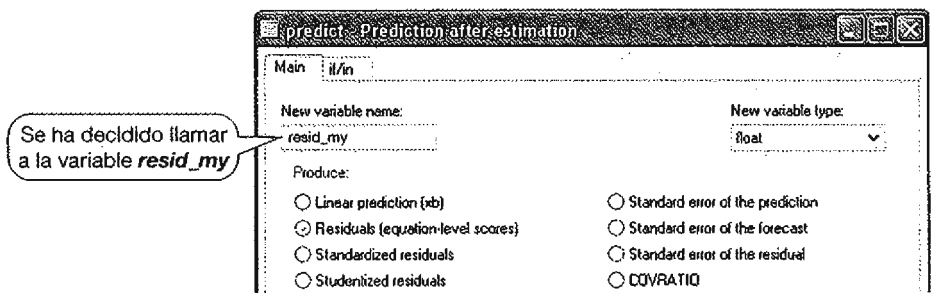
Al igual que en la regresión lineal simple, el modelo se basa en unos supuestos similares, que son los siguientes:

- Las variables están relacionadas linealmente.
- La distribución de la variable dependiente condicionada a cada posible combinación de valores de las independientes es una distribución normal multivariable.
- Las variables son independientes unas de otras.
- Existe homogeneidad de las varianzas (homoscedasticidad): las varianzas de la variable Y condicionadas a los valores de X son homogéneas.

Para comprobar estos supuestos, se deben guardar los residuales y valorar si se adaptan a la normalidad, igual que se hace en la regresión simple.

Los residuales del modelo pueden crearse a través del menú de STATA:

**Statistics → Postestimation → Predictions, residuals, etc.**




O a través de la instrucción donde el usuario ha decidido poner ***resid_my*** como nombre a los residuales:

```
predict resid_my, residuals  #(en regresión lineal,
puede emplearse también la instrucción score)
```

La comprobación de la normalidad puede realizarse a través de distintos test:

**Statistics → Summaries, tables, and tests → Distributional plots and tests**
**Shapiro-Wilk normality test**

```
swilk resid_my
```

**Statistics → Summaries, tables, and tests → Distributional plots and tests →**
**Shapiro-Francia normality test**

```
sfrancia resid_my
```

**Statistics → Summaries, tables, and tests → Distributional plots and tests →**
**Skewness and kurtosis normality test**

```
sktest resid_my
```

En el ejemplo del modelo que predecía el peso de los niños recién nacidos a través del peso de la madre y la presencia de HTA durante el embarazo se observa que el modelo es válido y se cumple el criterio de normalidad de los residuales con las tres aproximaciones propuestas.

Obsérvese que como solo existe una variable en la base de datos que empiece con la letra r, puede introducirse el nombre reducido de la misma, *r*, en vez del nombre completo ***resid_my***

```
swilk r
```

Shapiro-Wilk W test for normal data

| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| res | 189 | 0.99411 | 0.837 | -0.409 | 0.65877 |

```
sfrancia r
```

Shapiro-Francia W' test for normal data

| Variable | Obs | W' | V' | z | Prob>z |
|---|---|---|---|---|---|
| res | 189 | 0.99296 | 1.080 | 0.164 | 0.43499 |

```
sktest r
```

Skewness/Kurtosis tests for Normality

| | | | | joint | |
|---|---|---|---|---|---|
| Variable | Obs | Pr(Skewness) | Pr(Kurtosis) | adj chi2(2) | Prob>chi2 |
| res | 189 | 0.3292 | 0.7296 | 1.08 | 0.5818 |

Si el tamaño muestral es grande, habitualmente resultarán significativos los test de normalidad de los residuales, lo cual tiene escasa relevancia práctica (**19**). Resulta entonces más importante valorar la magnitud del alejamiento de la normalidad con métodos gráficos. Habitualmente, con tamaños muestrales grandes ($n > 500$), la regresión suele ser suficientemente robusta.

Los residuales pueden representarse a través de los gráficos *Q-Q* y *P-P* a través de las instrucciones:

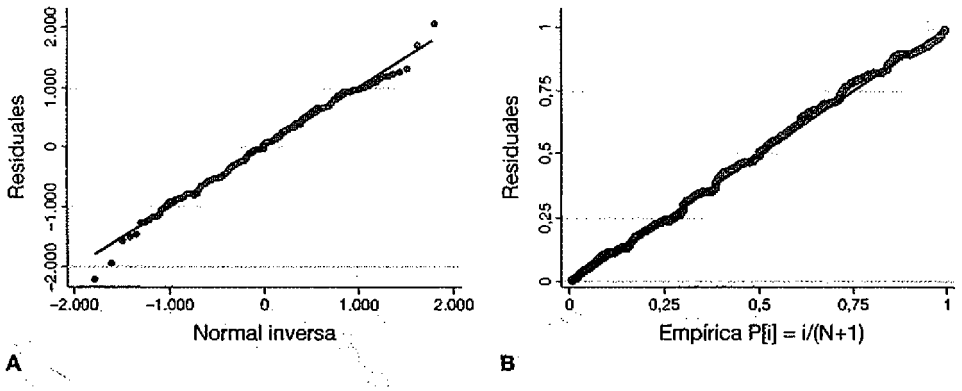



**Figura 12.7** Representación gráfica de los residuales. **A.** Gráfico *Q-Q* normal. **B.** Gráfico *P-P* normal.

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Normal quantile plot**

```
qnorm resid_my
```

**Statistics → Summaries, tables, and tests → Distributional plots and tests → Normal probability plot, standardized**

```
pnorm resid_my
```

En la figura 12.7 puede observarse la representación gráfica de los residuales para el ejemplo del peso de los recién nacidos.

***Interpretación:*** *a través del gráfico Q-Q puede observarse que los residuales de este modelo no se apartan de la normalidad. Lo mismo ocurre con el gráfico P-P, en el que los puntos se sitúan sobre la diagonal, por lo que puede concluirse que el modelo lineal es correcto.*

Además, STATA ofrece otras posibilidades, como la realización de un diagrama de dispersión entre la variable independiente X (eje de abscisas) y el residual del modelo (eje de ordenadas).

**Statistics → Linear models and related → Regression diagnostics → Residual-versus-predictor plot**

```
rvpplot x, yline(0)
```

(Se le ha solicitado a STATA que trace una línea recta sobre el valor $y = 0$ para una mejor visualización e interpretación de la gráfica.)

O bien entre el valor predicho por el modelo (eje de abscisas) y la residual del modelo (eje de ordenadas):

**Statistics → Linear models and related → Regression diagnostics → Residual-versus-fitted plot**

```
rvfplot, yline(0)
```

Ambas representaciones permiten comprobar si los residuales presentan una dispersión constante a lo largo de todos los sujetos de la muestra, es decir, si existe homogeneidad de varianzas. La figura 12.8 muestra estas representaciones para el ejemplo del peso de los recién nacidos.

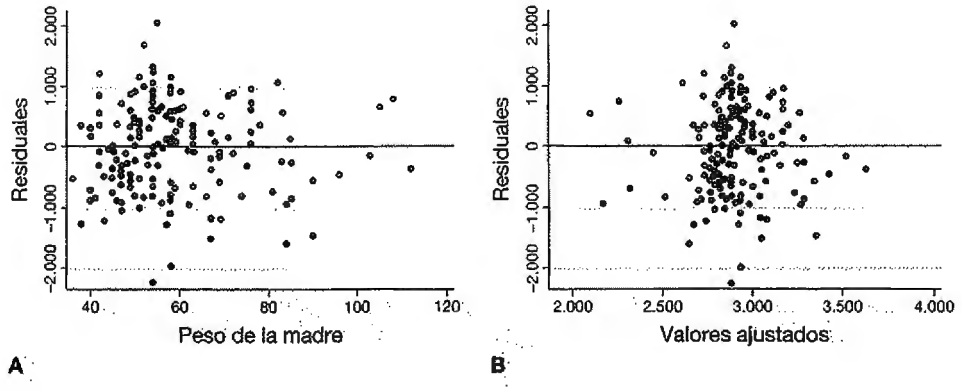



**Figura 12.8** Representación gráfica de los residuales para comprobar si existe homogeneidad de varianzas. **A.** Residual *vs.* valor X. **B.** Residual *vs.* valor predicho.

***Interpretación:*** *la primera gráfica representa los valores de peso de la madre ($X_p$) en el eje de abscisas frente al valor de los residuales del modelo final (según peso de la madre y presencia de HTA) (eje de ordenadas). La segunda gráfica representa los valores predichos (eje de abscisas) frente a los residuales del modelo. Sin embargo, ambas gráficas aportan una información similar. No existe heterogeneidad de varianzas, ya que los puntos de ambas gráficas se distribuyen homogéneamente a lo largo de la línea horizontal. Los puntos representados no siguen un patrón establecido.*

Cuando exista un alejamiento notorio de la normalidad en los residuales, se puede probar un término cuadrático para alguna de las variables independientes cuantitativas más importantes. Esto conduciría a modelos polinómicos y permitiría incluir relaciones curvilíneas. Existen amplias posibilidades de modelización no lineal en regresión (**20-23**) (v. apartado 12.18).

## 12.13. TOLERANCIA, FACTOR DE INFLACIÓN DE VARIANZA Y MULTICOLINEALIDAD

Cuando se produce una alta correlación (asociaciones lineales) entre el conjunto de variables independientes (predictoras) incluidas en el modelo, se habla de colinealidad. En este caso, las estimaciones obtenidas en el modelo son muy inestables, ya que resulta complicado separar el efecto aportado por cada una de las variables predictoras.

Existen diferentes índices para valorar la colinealidad en un modelo de regresión lineal múltiple. Un primer paso podría ser estudiar la matriz de correlaciones de las diferentes variables independientes X incluidas en el modelo. Si existen altas correlaciones entre ellas, puede sospecharse la presencia de multicolinealidad. Otras mediciones que suelen utilizarse son las medidas de tolerancia (T) y su recíproco, el factor de inflación de la varianza (FIV).

La T de una variable $X_i$ se define como la proporción de variabilidad de dicha variable (proporción de su varianza) que no se explica por el resto de las variables independientes X incluidas en la ecuación de regresión.

Para calcular la colinealidad de una variable predictora $X_i$, es necesario conocer en primer lugar el cuadrado del coeficiente de correlación múltiple entre dicha variable predictora $X_i$ y el resto de variables predictoras ($p - 1$).

$$R^2_{x_i|x_1,\ldots,x_p}$$

La T se calcula a través de la siguiente fórmula:

$$T = 1 - R^2_{x_i|x_1,\ldots,x_p}$$

Una variable con un valor de T muy bajo contribuye con poca información a un modelo y puede dar lugar a coeficientes de regresión muy inestables.

El FIV, denominado también factor de incremento de la varianza, se define como la proporción de variabilidad de la variable $X_i$, que es explicada por el resto de las variables predictoras del modelo. Corresponde al inverso de la tolerancia, por lo que su fórmula será:

$$FIV = \frac{1}{1 - R^2_{x_i|x_1,\ldots,x_p}}$$

Cuando el FIV crece, también lo hace la varianza del coeficiente de regresión, y el modelo se vuelve inestable. Los valores de FIV grandes son un indicador de la existencia de multicolinealidad.

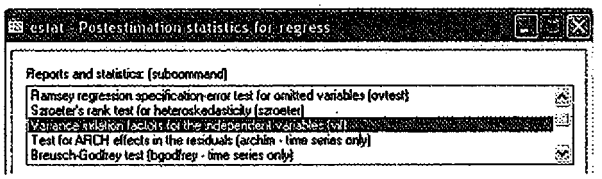
Algunos autores consideran que existe un problema grave de multicolinealidad cuando el FIV de alguna variable independiente $X_i$ es mayor de 10 (**24**). Este valor correspondería a un $R^2_i$ de 0,9 y a una $T_i < 0{,}1$.

Sin embargo, la mayoría de los investigadores confían en reglas informales para valorar la multicolinealidad (**25**). Estas son:

1. El FIV más grande es mayor que 10.
2. La media de todos los FIV es considerablemente mayor que 1.

En STATA pueden obtenerse las medidas de tolerancia y del factor de inflación de la varianza a través del menú:

**Statistics → Postestimation → Reports and statistics**




O de la instrucción:

```
estat vif
```

Véase con un ejemplo. Se desea valorar el efecto que tienen dos tipos de alimentos, la comida rápida y los lácteos desnatados, sobre el IMC. Para ello, se realiza un modelo de regresión múltiple donde se introducen como variables independientes: el consumo de comida rápida, expresado en cambios en 100 g de consumo *(crapida100);* el de lácteos desnatados, expresado en cambios en 100 g de consumo *(lacteos100);* la ingesta de grasas totales, expresado en cambios en 100 g de ingesta *(grasas100),* y, además, el sexo y la edad de los sujetos.

Se obtiene el siguiente modelo de regresión en STATA:

```
. regress imc crapida100 lacteos100 grasas100 edad sexo
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 16088.4242 | 5 | 3217.68484 | Number of obs | = 4248 |
| Residual | 38774.186 | 4242 | 9.1405436 | F( 5, 4242) | = 352.02 |
| | | | | Prob > F | = 0.0000 |
| Total | 54862.6102 | 4247 | 12.917968 | R-squared | = 0.2932 |
| | | | | Adj R-squared | = 0.2924 |
| | | | | Root MSE | = 3.0233 |

| imc | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| crapida100 | .4028426 | .1867108 | 2.16 | 0.031 | .0367918 | .7688934 |
| lacteos100 | -.1085813 | .0212874 | -5.10 | 0.000 | -.1503157 | -.0668468 |
| grasas100 | -.001143 | .0044374 | -0.26 | 0.797 | -.0098426 | .0075566 |
| edad | .0811918 | .0041555 | 19.54 | 0.000 | .0730449 | .0893388 |
| sexo | -2.820769 | .1014208 | -27.81 | 0.000 | -3.019607 | -2.621931 |
| | | | 82.88 | 0.000 | 24.50084 | 25.68799 |

***Interpretación:*** *el incremento en el consumo de 100 g de comida rápida se asocia con un incremento significativo en el IMC (0,40 kg/m²) (p = 0,031), a igualdad de consumo de lácteos desnatados, grasas totales, edad y sexo de los sujetos. Por otro lado, el consumo de lácteos desnatados se asocia con menores incrementos en el IMC de los participantes una vez considerado el resto de las variables del modelo. De hecho, por cada incremento en 100 g en el consumo de este tipo de alimentos, se produce una reducción media del IMC de 0,11 unidades (b = –0,1086). La ingesta de grasas totales no parece asociarse con el IMC en esta muestra.*

Sin embargo, podría pensarse que quizá el consumo de comida rápida podría asociarse de forma lineal y directa con el de grasas totales (es decir, al aumentar el consumo de comida rápida, lo hace también la ingesta de grasas totales, en una correlación positiva). De igual forma, cabría pensar que el consumo de lácteos desnatados podría presentar una correlación también con la ingesta total de grasas, pero, en este caso, de forma inversa (correlación negativa). Por ello, se decide valorar la T y el FIV del modelo. El resultado obtenido es el siguiente:

```
. estat vif
```

| Variable | VIF | 1/VIF |
|---|---|---|
| edad | 1.23 | 0.816211 |
| sexo | 1.14 | 0.879251 |
| crapida100 | 1.11 | 0.899829 |
| lacteos100 | 1.07 | 0.930659 |
| grasas100 | 1.04 | 0.961964 |
| Mean VIF | 1.12 | |

***Interpretación:*** *ninguna de las variables presenta un FIV superior a 10, es más, el valor más alto corresponde a la edad, con un valor de 1,23. Además, la media de FIV del modelo no difiere sustancialmente de 1 (es 1,12). Puede concluirse que no existe multicolinealidad.*

No obstante, sería interesante repetir el análisis excluyendo la variable *grasas100,* que no parece predecir el IMC una vez tenidas en cuenta el resto de las variables del modelo (condicionalmente a que otras variables ya hayan sido tenidas en cuenta).

## 12.14. VARIABLES QUE DEBEN REGISTRARSE EN UNA INVESTIGACIÓN

La relación entre un factor en estudio (exposición) y una respuesta (desenlace) puede estar afectada, de distintas maneras, por terceras variables. Si se estudia la relación entre las dos variables de interés sin tener en cuenta otros factores relacionados con ambas, puede desaparecer la relación o aparecer una relación inexistente, espuria, a causa del problema de la confusión. Además, otros factores podrían actuar modificando la intensidad y/o el sentido de la relación evaluada, debido al fenómeno de la interacción.

En investigación experimental se tiene una mayor garantía de control sobre todas estas variables externas a la hipótesis del estudio, ya que la aleatorización reparte dicha influencia de forma similar entre los diferentes grupos. No obstante, cuando no es posible la aleatorización, e incluso cuando lo ha sido, pero quedan dudas de su efectividad real, las técnicas de ajuste multivariante permitirán controlar el efecto de las variables de confusión que puedan influir en la relación entre la exposición y el desenlace. Evidentemente, este ajuste estadístico requiere conocimiento previo de las variables que pueden afectar a la relación que se va a evaluar, es decir, todos los factores de riesgo conocidos del desenlace. Esto implica la necesidad de una completa revisión teórica y bibliográfica previa al diseño de cualquier investigación, para introducir en el estudio todas aquellas variables que puedan afectar a la relación entre las dos variables de interés. El olvido de una de estas variables puede invalidar el estudio de manera irreversible.

Otro problema importante en los análisis multivariantes es la existencia de datos faltantes en determinadas variables. Esto lleva a la eliminación del caso completo, lo que puede conducir a

una reducción drástica del tamaño muestral efectivo. Así sucede en algunos trabajos en los que determinadas variables se miden solo en algunos sujetos, por ejemplo, porque *«la técnica es muy cara y no se puede medir en todos»*. A efectos prácticos hay que ser precavidos a la hora de incluir en un modelo una variable en la que faltan más de un 10% de los casos.

Para terminar, es importante tener en cuenta que en este proceso de ajuste estadístico no se deben incluir las variables intermedias (eslabones intermedios entre la exposición y el desenlace) ni las variables posteriores, aquellas que son consecuencia de la variable desenlace.

## 12.15. VARIABLES CATEGÓRICAS Y VARIABLES INDICADORAS *(DUMMY)*

Mientras la variable independiente Y es una variable cuantitativa, las independientes X pueden ser tanto cuantitativas como cualitativas dicotómicas. Sin embargo, cuando existan variables cualitativas de más de dos categorías, estas pueden ser introducidas en el modelo de regresión múltiple, aunque previamente es preciso «reconvertirlas». Es necesaria la construcción de las denominadas variables ficticias, indicadoras o variables «dummy». Si la variable categórica está formada por $k$ categorías, será posible analizarla introduciendo en el modelo simultáneamente $k - 1$ variables *dummy*. Estas variables ficticias son dicotómicas y toman los códigos 0 y 1. El valor 0 se asigna a aquella categoría que se toma como referencia, y habrá una variable *dummy* por cada una de las otras categorías, que solo valdrán 1 cuando el sujeto pertenezca a la respectiva categoría. Generalmente se asigna el valor 0 a aquella categoría en que se espera un nivel menor o basal o la referencia «natural» (p. ej., la categoría inferior cuando se hacen cuartiles o quintiles o la categoría de no expuestos a un determinado factor). Existen otros métodos, pero el más usual es el de codificación *dummy*.

Estas variables solo tienen sentido consideradas en conjunto. Por tanto, siempre serán tratadas como un bloque y no podrán analizarse separadamente.

Véase con un ejemplo concreto. Estruch et al. (**6**) compararon tres dietas en cuanto a su eficacia para reducir el peso: una dieta rica en aceite de oliva virgen (AOV), otra rica en frutos secos (FS) y una dieta control baja en grasas (control). La variable cualitativa de agrupación *(dieta)* tenía, por tanto, tres niveles o categorías. El grupo control se consideró la categoría de referencia y se crearon dos nuevas variables (*AOV* y *FS*). Estas nuevas variables, de tipo *dummy*, servían para comparar cada una de ellas frente al grupo control. La nueva variable *AOV* valía 1 cuando el participante era asignado al grupo de aceite de oliva virgen y 0 en caso contrario (control o FS). La nueva variable *FS* valía 1 cuando el participante era asignado al grupo de frutos secos y 0 en caso contrario (control o AOV) (tabla 12.7).

Esta recodificación puede ser llevada a cabo con STATA con la instrucción:

```
generate AOV=1 if grupo==1
generate FS=1 if grupo==2
mvencode AOV FS, mv(0)
```

(Esta instrucción permite transformar los valores faltantes *[missing values]* en las variables *AOV* y *FS* en valores 0.)

*Tabla 12.7 Dos variables* **dummy** *sustituyen a una variable con tres categorías*

| CODIFICACIÓN | NUEVAS VARIABLES (VARIABLES *DUMMY*) | |
|---|---|---|
| VARIABLE ORIGINAL (DIETA) | AOV | FS |
| 1 = Aceite de oliva | 1 | 0 |
| 2 = Frutos secos | 0 | 1 |
| 3 = Control | 0 | 0 |

En ocasiones, no es necesario crear las variables *dummy* a través de instrucciones dadas al programa estadístico. STATA permite la creación directa de variables indicadoras o *dummy* sin necesidad de manipular las variables originales. La creación es directa siempre y cuando la categoría elegida como referencia sea la primera. Bastaría con incluir el término «**i.**» delante de la variable cualitativa que debe ser transformada. En el ejemplo presentado:

```
regress camb_peso i.dieta
```

(Esta instrucción creará igualmente dos *dummy*. Sin embargo, la categoría de referencia será, en este caso, la dieta rica en AOV. Así, las *dummy* creadas compararán el cambio de peso entre la dieta rica en FS y la rica en AOV, y entre la dieta control y la rica en AOV. Si se desease otra comparación, se debería recodificar la variable original *dieta*.)

La instrucción regress permite llevar a cabo un análisis de regresión múltiple con el programa STATA. Se introduce a continuación la variable dependiente (***camb_peso*** en el ejemplo) y posteriormente las variables independientes (***AOV*** y ***FS***, variables *dummy* en el ejemplo)

```
. regress  camb_peso  AOV FS
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 1.25991775 | 2 | .629958874 |
| Residual | 5698.79764 | 702 | 8.11794535 |
| Total | 5700.05756 | 704 | 8.09667266 |

| | |
|---|---|
| Number of obs | = 705 |
| F( 2, 702) | = 0.08 |
| Prob > F | = 0.9253 |
| R-squared | = 0.0002 |
| Adj R-squared | = -0.0026 |
| Root MSE | = 2.8492 |

| camb_peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| AOV | .0962443 | .2638161 | 0.36 | 0.715 | -.4217188 | .6142075 |
| FS | .0845728 | .2664315 | 0.32 | 0.751 | -.4385251 | .6076707 |
| _cons | -.3308597 | .191658 | -1.73 | 0.085 | -.7071512 | .0454318 |

***Interpretación:*** *el listado de salida de STATA sirve para crear tres ecuaciones de predicción de cambio de peso, una para cada tipo de dieta. Así, se puede comparar el cambio de peso (kg) predicho por el modelo para el grupo de dieta rica en aceite de oliva virgen, lo anticipado para dieta rica en frutos secos y lo predicho para el grupo control (baja en grasa).*

*Modelo para dieta rica en* ***aceite de oliva*** *virgen (AOV = 1, FS = 0):*

$$camb\_peso = -0,33 + 0,096 * 1 + 0,085 * 0 = -0,33 + 0,096 = -0,234$$

*Modelo para dieta rica en* ***frutos secos*** *(AOV = 0, FS = 1):*

$$camb\_peso = -0,33 + 0,096 * 0 + 0,085 * 1 = -0,28 + 0,085 = -0,195$$

*Modelo para dieta baja en grasa (grupo* ***control****) (AOV = 0, FS = 0):*

$$camb\_peso = -0,33 + 0,096 * 0 + 0,085 * 0 = -0,33$$

*El análisis establece que, por término medio, los tres grupos han perdido peso. El cambio medio de peso ha sido de –234 g en el grupo de aceite de oliva, de –195 g en el grupo de frutos secos y de –330 g en el grupo control.*

*La interpretación de los dos coeficientes (0,096 y 0,085) es, por tanto, muy sencilla y directa. El primero (+0,096) es la diferencia en el cambio de peso entre el grupo de aceite y el grupo control, el segundo (+0,085) es la diferencia entre el grupo de frutos secos y el grupo control. Ninguna de estas diferencias resultó estadísticamente significativa.*

Estas comparaciones son equivalentes a un ANOVA con contrastes y se podrían haber hecho con dos contrastes *a priori* (coeficientes –1, 0 y +1 para el primer contraste y coeficientes 0, –1 y +1 para el segundo).

Contraste 1: comparación AOV-control.

Contraste 2: comparación FS-control.

El resultado sería exactamente idéntico al de la regresión. La ventaja de hacerlo por regresión es que basta con introducir también otras variables en el modelo (p. ej., edad, sexo, IMC inicial, etc.) para obtener estas mismas estimaciones ya ajustadas por esos posibles factores de confusión (v. apartado 12.16) (**12**).

## 12.16. FACTORES DE CONFUSIÓN EN LA REGRESIÓN MÚLTIPLE

El fenómeno de la confusión es un concepto clave en el análisis multivariable. Su comprensión es necesaria para entender el proceso de construcción de un modelo de regresión múltiple.

### 12.16.1. Definición de factor de confusión

Un factor de confusión es una variable que distorsiona la medida de asociación entre una exposición (variable independiente) y un desenlace (variable dependiente) (**2,26-29**). El resultado de la presencia de una variable de confusión puede ser la observación de un efecto donde en realidad no existe o una exageración de una asociación real (confusión positiva), o, por el contrario, la atenuación de un asociación real e incluso una inversión del sentido de la misma (confusión negativa). Un factor de confusión es el resultado de relaciones específicas existentes entre las variables de una base de datos. El factor de confusión está, con frecuencia, presente a pesar de los esfuerzos que se hacen para evitarlo.

Para que una variable sea factor de confusión de la asociación entre una exposición o factor de riesgo y un desenlace o respuesta, tiene que cumplir las siguientes condiciones:

1. Estar asociada con el desenlace, independientemente de su asociación con la exposición de interés (es decir, tanto en sujetos expuestos como en no expuestos), pero no debe ser un resultado del efecto. Un refinamiento de esta primera condición es que el factor de confusión debe ser un factor causal o un marcador de un factor que cause el efecto (**30**).
2. Estar asociada con la exposición (independientemente del efecto).
3. No ser un eslabón intermedio entre la exposición y el desenlace (**31**), como recoge la figura 12.9. Los eslabones intermedios son los mecanismos por los que la supuesta causa (variable independiente o exposición) ejercería el supuesto efecto (variable dependiente o desenlace). Eslabones intermedios serían, por ejemplo, los niveles de colesterol HDL entre una causa como la obesidad y su efecto correspondiente, el infarto de miocardio: la obesidad produciría una reducción de colesterol HDL, que, a su vez, aumentaría el riesgo de infarto de miocardio. La aplicación práctica es que la relación entre obesidad y riesgo de infarto de miocardio no se debería ajustar por los niveles de HDL (**32**).

El fenómeno de confusión se puede representar gráficamente mediante un *diagrama acíclico dirigido* (DAG), donde las flechas indican relaciones causa-efecto (**33**) (fig. 12.10).

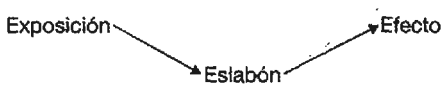



**Figura 12.9** No se debe ajustar por eslabones intermedios en la cadena causal.

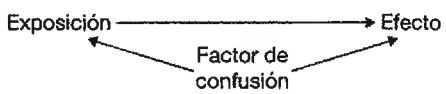



**Figura 12.10** Diagrama acíclico dirigido para representar un factor de confusión.

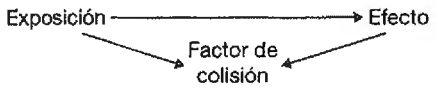



**Figura 12.11** No se debe ajustar por un *collider* (factor de colisión).

No sería un factor de confusión aquella variable asociada a la exposición y al efecto (directa o indirectamente) a la que apuntasen dos cabezas de flecha (fig. 12.11). A esta variable a la que le llegan dos puntas de flecha se le llama *collider* (34) o factor de colisión (2). Nunca se debe ajustar por un *collider,* pues se corre el riesgo de crear confusión donde no la había.

No basta con verificar el cumplimiento de estas condiciones para establecer que una variable es un factor de confusión, sino que se debe añadir un concepto clave: solo habrá confusión en la relación entre dos variables cuando ambas compartan una causa común (35). Esto requiere un gran conocimiento de las relaciones causales entre variables y solo se logra si quien analiza los datos tiene gran familiaridad con el «estado del arte» en las investigaciones que se hayan realizado sobre ese tema específico. No se pueden valorar factores de confusión si no se conocen muy bien los avances científicos sobre el tema que se estudia.

### 12.16.2. No se deben usar valores *p* ni procedimientos *stepwise* para valorar la confusión

Resulta especialmente desafortunado guiarse por criterios meramente estadísticos para identificar un factor de confusión. No solo no se recomiendan, sino que se desaconsejan totalmente los métodos automáticos de selección de variables *(stepwise, forward, backward)* presentes en la mayoría de los programas estadísticos para identificar factores de confusión. Estos métodos solo están basados en valores *p* (contrastes de hipótesis). No hay ningún valor *p* ni ningún contraste de hipótesis que indique que una variable actúa como factor de confusión. La confusión no depende de la significación estadística. La significación estadística, además, está muy condicionada por el tamaño muestral. Lo que hay que hacer para valorar la confusión es basarse en las relaciones causales entre variables que se esperan a partir del conocimiento previo. Cuando haya razones fundadas para pensar que las dos variables cuya relación se valora comparten una causa común, entonces esa causa común debe considerarse un factor de confusión, sean cuales sean los valores *p*.

### 12.16.3. Cambio en la magnitud del coeficiente como criterio de confusión

En muchas ocasiones se aplican dos modelos para comparar la estimación sin ajustar (cruda) con la estimación ajustada por el posible factor de confusión (tras introducir la variable potencialmente confusora en el modelo). Se piensa que, si hay un cambio suficientemente grande (del 10% o mayor) en la magnitud del coeficiente para la variable independiente entre los dos modelos (al pasar del crudo al ajustado), entonces el factor por el que se ajustó es un factor de confusión y el modelo válido sería el ajustado. Este criterio supone una mejoría respecto al planteamiento anterior de guiarse por la mera significación estadística (valor *p*). Ahora ya no se miran los valores *p*, sino las magnitudes del efecto, antes y después de ajustar. Es un procedimiento práctico que se ha recomendado y usado muy frecuentemente (de hecho, es el que más se utiliza), pero que tampoco es ideal si se aplica de manera «automática» y sin tener en cuenta otros criterios. Debe pensarse siempre en términos de relaciones causales y ajustar solo por aquellas variables que supongan realmente una posible causa común (y no estén contando la misma historia que otra variable que ya está en el modelo).

Lo ideal es conocer muy bien el estado de la ciencia sobre las posibles relaciones causales entre las variables que se manejan y recurrir a gráficos causales que expliciten las posibles relaciones causales entre variables (36).

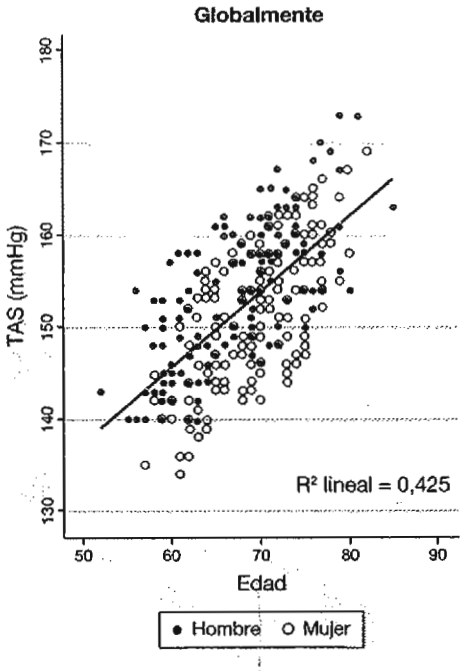



**Figura 12.12** Gráfico de dispersión de la variable edad y la variable tensión arterial sistólica (TAS) considerando conjuntamente ambos sexos.

La figura 12.12 presenta de forma gráfica la relación establecida entre la edad (eje *x*) (coeficiente *b* asociado a la edad) y la presión arterial (eje *y*) sin ajustar por ningún otro factor (p. ej., sin tener en cuenta el sexo de los sujetos). En la figura 12.13 se presenta un ejemplo en el que la variable sexo no actúa como factor de confusión. Al separar a los sujetos en dos grupos en función de esta tercera variable (sexo en este caso), se obtienen las dos rectas de regresión, una para hombres y otra para mujeres, de la asociación entre la edad y la presión arterial. Puede apreciarse visualmente que sus pendientes son prácticamente idénticas no solo entre sí, sino también iguales a la de la figura 12.12. El coeficiente de regresión (pendiente de la recta) obtenido en la figura 12.12 (toda la muestra conjuntamente) coincidiría con el coeficiente (ajustado por sexo) obtenido en cada subgrupo en la figura 12.13. Ajustar es análogo a estratificar. Si el sexo hubiese actuado como factor de confusión, las pendientes en la figura 12.13 (estratificada por sexo) serían distintas que en la figura 12.12 (global). Es importante insistir en que no se habla de pendientes distintas entre hombres y mujeres si hubiese confusión, sino de pendientes distintas cuando se analizan juntos y cuando se analizan por separado.

En términos generales, se habla de confusión cuando existen diferencias importantes entre las estimaciones brutas o crudas (toda la muestra conjuntamente) de una asociación y las estimaciones ajustadas por los factores de confusión (estratificadas). Estas diferencias se pueden valorar siguiendo varios criterios, aunque existe un cierto consenso en la importancia de valorar el efecto que tiene el ajuste sobre la magnitud de los cambios de las medidas de asociación. De este modo, un factor puede considerarse como de confusión cuando su ajuste es responsable de un cambio de al menos el 10% en la magnitud de la diferencia entre las estimaciones ajustadas y las brutas.

Por norma general, se emplea la siguiente fórmula:

$$\text{Magnitud} = \frac{\left|\text{Valor crudo} - \text{Valor ajustado}\right|}{\text{Valor ajustado}} \times 100$$

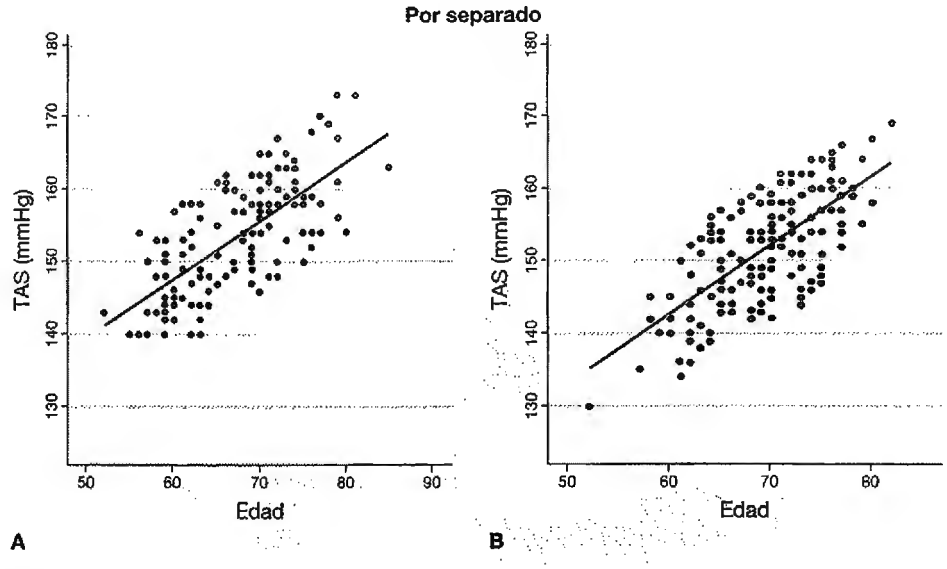



**Figura 12.13** Ejemplo de ausencia de confusión por la variable sexo. **A.** Hombre. **B.** Mujer.

Debe insistirse en que no se debe caer en una aplicación mecánica de este criterio, sino tener siempre presente cuál es el papel causal de unas y otras variables y valorar si la variable potencialmente de confusión es una causa común compartida por la exposición y el efecto o un correlato de ella (**35**).

La identificación de la confusión requiere, en cualquier caso, tener también en cuenta los coeficientes de regresión de la variable independiente $X_1$ estudiada en dos modelos de regresión, uno que no contenga al potencial factor de confusión y otro que sí lo incluya. El primer modelo de regresión incluirá únicamente la variable estudiada como exposición principal (variable independiente = $X_1$), pero no la variable de control (posible variable de confusión, $X_2$). Este primer modelo es la estimación bruta o cruda. Un segundo modelo incluirá también la variable independiente principal de exposición $X_1$, pero, además, el posible factor de confusión $X_2$. El coeficiente de este modelo para $X_1$ es la estimación ajustada. Si el coeficiente de regresión asociado a la variable independiente $X_1$ varía más de un 10% al ajustar, se podrá pensar que $X_2$ es un factor de confusión, pero hay que acudir también al conocimiento experto del tema y a las relaciones causales demostradas previamente entre las variables. Si hay argumentos biológicos para apoyar que el potencial factor de confusión pudiera ser un factor causal común de la exposición $X_1$ y el efecto Y, entonces solo sería válido el segundo modelo de regresión (el ajustado).

Véase un ejemplo. Suponga que se trata de predecir la presión arterial sistólica de un grupo de pacientes usando su edad y se valora la posible confusión introducida por la variable sexo (tabla 12.8).

El valor del coeficiente de regresión para la edad es 0,91 en la estimación cruda y 0,93 en la estimación ajustada por sexo. En este caso, el sexo NO actúa como factor de confusión en la relación entre edad y presión arterial sistólica, pues el coeficiente de la edad no varía en más de un 10% al ajustar (0,93 frente a 0,91).

*Tabla 12.8 Coeficientes de regresión en la estimación cruda y en la estimación ajustada*

| | COEFICIENTE DE REGRESIÓN | *P* SIGNIFICACIÓN |
|---|---|---|
| **Modelo 1 (estimación cruda)** | | |
| *Edad* (variable X) | 0,91 | <0,01 |
| **Modelo 2 (estimación ajustada por sexo)** | | |
| *Edad* (variable $X_1$) | 0,93 | <0,01 |
| *Sexo* (variable de control = $X_2$) | 8,44 | 0,239 |

*Variable dependiente: presión arterial sistólica.*

En teoría, el modelo final no necesitaría incluir la variable sexo. No obstante, hay tantos argumentos biológicos para tener en cuenta las diferencias entre hombres y mujeres en la tensión arterial, que siempre es mejor presentar las estimaciones ajustadas por sexo.

## 12.17. INTERACCIÓN (O MODIFICACIÓN DEL EFECTO) EN LA REGRESIÓN MÚLTIPLE

El concepto de modificación del efecto se debe a Miettinen (**36**) y es importante, porque, al igual que ocurre con el de confusión, la adecuada comprensión de este fenómeno determinará una interpretación correcta de los resultados de los estudios que se publican en la literatura científica. Es un concepto que debe distinguirse claramente de la confusión, ya que su identificación determinará una actitud radicalmente opuesta por parte del investigador: así como en presencia de confusión el objetivo es eliminar una distorsión de la medida de asociación objeto de la investigación, ante la presencia de interacción el objetivo es describir mejor este fenómeno, una riqueza existente en los datos (**4**).

Se habla de interacción biológica entre dos o más factores causales cuando existe una interdependencia entre sus mecanismos de acción para curar, prevenir o controlar una enfermedad de manera que la incidencia de enfermedad atribuible a este conjunto de factores causales es diferente a la esperable de sus efectos individuales (**37**).

No se puede hablar de interacción en términos absolutos, sino que depende de la escala de medición que se use. Se puede utilizar una escala aditiva o una multiplicativa.

- En escala aditiva, se habla de interacción cuando el efecto conjunto de dos variables es significativamente superior a la suma de sus efectos individuales (sinergia). También cabría la posibilidad de una interacción *negativa* (antagonismo). Habrá antagonismo cuando el efecto conjunto sea significativamente *inferior* a la suma de los efectos individuales.
- En escala multiplicativa, se habla de interacción si el efecto conjunto de dos variables es significativamente distinto a la multiplicación de cada uno de sus efectos individuales.

En regresión múltiple, la escala es *aditiva.*

Cuando se ha hablado antes de confusión, se insistía en que no se debía usar la significación estadística para valorarla, pues ningún valor *p* sirve para detectar la confusión. En cambio, la interacción se plantea como un efecto *significativamente* distinto del que sería esperable al combinar las acciones por separado. Es decir, *para la interacción sí hay un test estadístico y lo que importa más es su valor p.* La confusión no se evalúa por valores *p,* sino por la magnitud del efecto y el conocimiento biológico.

En una interacción en escala aditiva valorada con una regresión múltiple, si una variable es continua y la otra es dicotómica se verán dos rectas divergentes cuando se represente un diagrama de dispersión con ajuste de rectas de regresión. La pendiente será distinta en los dos grupos definidos por la variable dicotómica.

Para la identificación de la interacción (modificación de efecto) se siguen estos pasos:

1. Valorar la existencia de causalidad entre la exposición $X_1$ y el desenlace Y.
2. Excluir la existencia de confusión (o controlarla si es que existe).
3. Realizar un análisis, separado por estratos, estimando el efecto de la exposición $X_1$ dentro de cada subgrupo (estrato) de la variable que se piense que pueda ser un modificador de efecto ($X_2$). Este tipo de análisis se llama «análisis de subgrupos o análisis estratificado»[4].

Otra aproximación para medir la interacción entre dos variables $X_1$ y $X_2$ es a través de la creación de un nuevo modelo de regresión múltiple, en el cual debe añadirse el término de interacción (producto $X_1 \times X_2$). Si el coeficiente de regresión asociado a este término multiplicativo es estadísticamente significativo ($p < 0,05$), puede considerarse que existe una interacción entre las dos variables independientes $X_1$ y $X_2$ con respecto a la variable dependiente Y. La variable $X_2$ actuaría como variable modificadora del efecto en la asociación $X_1$ – Y.

Se volverá al ejemplo de la predicción del peso del niño recién nacido (g) (Y) en función del peso de la madre (kg) ($X_1$) y la presencia o no de HTA durante el embarazo ($X_2$). Van a desarrollarse tres modelos. La comparación del modelo 1 (solo con $X_1$: *pesomadre*) y del modelo 2 (con dos variables X: $X_1$ y $X_2$: *pesomadre* e *HTA*) permitirá valorar si la presencia de HTA en el embarazo introduce confusión en la asociación entre el peso de la madre y el del niño recién nacido.

**Modelo 1: crudo** ← Instrucción: **regress *peso pesomadre***

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 9.855486 | 3.809995 | 2.59 | 0.010 | 2.339389 | 17.37158 |
| _cons | 2372.986 | 227.0911 | 10.45 | 0.000 | 1924.996 | 2820.975 |

Instrucción: **regress *peso pesomadre HTA***

**Modelo 2: ajustado por HTA durante el embarazo**

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 12.36788 | 3.852452 | 3.21 | 0.002 | 4.767763 | 19.968 |
| HTA | -599.4701 | 216.6363 | -2.77 | 0.006 | -1026.85 | -172.0899 |
| _cons | 2864.785 | 285.2792 | 10.04 | 0.000 | 2301.986 | 3427.584 |

***Interpretación:*** *puede observarse que existe un cambio sustancial en el coeficiente de regresión asociado al peso de la madre del modelo 1 (crudo) al modelo 2 (ajustado por HTA). Puede concluirse que la HTA actúa como factor de confusión. El cambio es de aproximadamente un 20% en el coeficiente de regresión asociado al peso de la madre [(9,9 – 12,4)/12,4 = 0,20].*

El modelo 3 será aquel que incluya ambas variables X: *pesomadre* e *HTA*, y el término producto *pesomadre*HTA*.

Puede crearse el término producto en STATA a través de la instrucción:

```
generate pesomadreHTA= pesomadre* HTA
```

4 Es peligroso intentar valorar un efecto *dentro* de múltiples subgrupos, ya que los test múltiples de la misma hipótesis repetidos una y otra vez en cada subgrupo producen una inflación del error tipo I y habrá resultados «estadísticamente significativos» que se deban exclusivamente al azar. El peor escenario es cuando globalmente no se encuentra asociación significativa, pero los investigadores argumentan que en algún subgrupo peculiar sí han encontrado tal asociación. Es peor todavía cuando su argumento se basa en un valor $p$ que está entre 0,01 y 0,05. Se debe evitar este tipo de manipulaciones de los datos (**38**), y en todo caso hay que basarse en pruebas de interacción que valoren específicamente si el efecto difiere significativamente de un grupo a otro (**39**).

Modelo 3. Valoración de la interacción introduciendo el término producto *pesomadreHTA* en el modelo:

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 1.498528 | 13.08705 | 0.11 | 0.909 | -24.32052 | 27.31758 |
| HTA | -1248.854 | 777.9997 | -1.61 | 0.110 | -2783.746 | 286.0382 |
| pesomadreHTA | 9.482951 | 10.91119 | 0.87 | 0.386 | -12.0434 | 31.0093 |
| _cons | 3593.421 | 885.6449 | 4.06 | 0.000 | 1846.159 | 5340.684 |

***Interpretación:*** *en este ejemplo, la variable HTA no actúa como modificador del efecto en la asociación entre el peso de la madre y el peso del recién nacido. Es decir, no existe interacción, pues el valor p (p = 0,386) del término de interacción (pesomadreHTA) es > 0,05.*

Imagine este otro ejemplo. Se quiere establecer la relación entre el peso de la madre y el del recién nacido, pero esta vez se tiene en cuenta el sexo del recién nacido (*sexo*; 0 = niño; 1 = niña).

Se presentarán tres modelos: 1) crudo; 2) ajustado por sexo del recién nacido, y 3) que incluya el término de interacción *pesomadre*sexo*.

**Modelo 1**

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 9.855486 | 3.809995 | 2.59 | 0.010 | 2.339389 | 17.37158 |
| _cons | 2372.986 | 227.0911 | 10.45 | 0.000 | 1924.996 | 2820.975 |

**Modelo 2**

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 10.25549 | 3.896285 | 2.63 | 0.009 | 2.568902 | 17.94208 |
| sexo | 55.40391 | 107.9612 | 0.51 | 0.608 | -157.5819 | 268.3897 |
| _cons | 2318.124 | 251.4021 | 9.22 | 0.000 | 1822.158 | 2814.09 |

Una vez tenido en cuenta el peso de la madre, el sexo del recién nacido no parece estar asociado con su peso ($p = 0,608$). Además, no es un factor de confusión porque no cambia de forma sustancial el coef. asociado al peso de la madre (crudo = 9,6; ajustado = 10,3). Pero el sexo del recién nacido sí juega un papel esencial en la asociación entre el peso de la madre y el peso del recién nacido. Obsérvese el modelo 3

**Modelo 3**

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | -5.268235 | 6.58567 | -0.80 | 0.425 | -18.26091 | 7.724436 |
| sexo | -1332.542 | 491.1062 | -2.71 | 0.007 | -2301.43 | -363.653 |
| pesomadres~o | 23.40289 | 8.086056 | 2.89 | 0.004 | 7.450155 | 39.35563 |
| _cons | 3267.754 | 410.4259 | 7.96 | 0.000 | 2458.037 | 4077.471 |

***Interpretación:*** *el coeficiente de regresión asociado a la variable pesomadresexo (pesomadre*sexo) es estadísticamente significativo (p = 0,004), por lo que se puede concluir que el sexo del recién nacido modifica la asociación entre el peso de la madre y el de dicho recién nacido. El efecto del peso de la madre sobre el del hijo varía según el sexo del recién nacido (fig. 12.14). Habría que segmentar la base de datos y realizar un análisis estratificado, calculando una pendiente para las niñas y otra diferente para los niños.*

*Las dos ecuaciones, con las dos pendientes distintas, pueden deducirse directamente del modelo 3, teniendo en cuenta que:*

$$y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$$

$$Peso = 3.267,8 - 5,3 * pesomadre - 1.332,5 * sexo + 23,4 * pesomadre * sexo$$

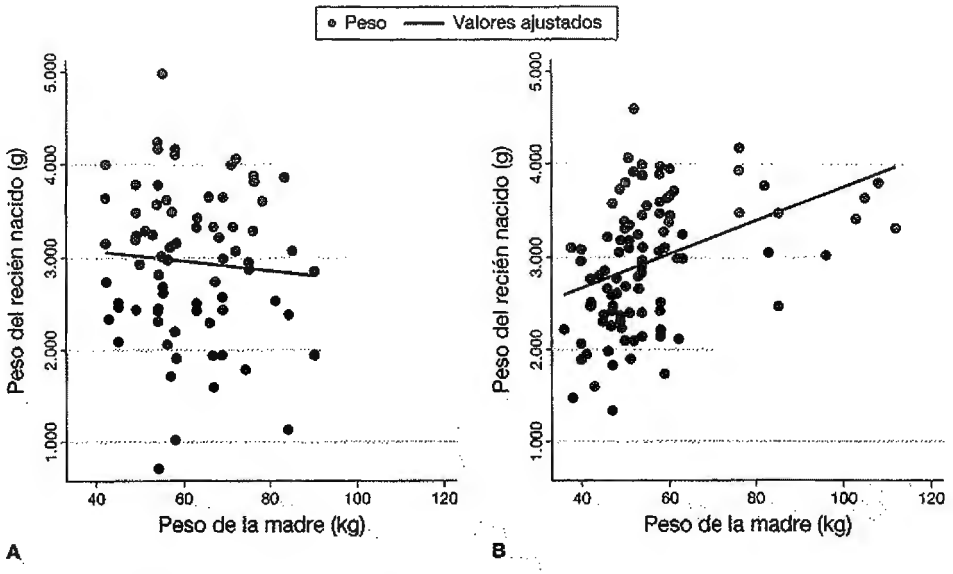



**Figura 12.14** Modificación del efecto del peso de la madre sobre el peso del recién nacido en función de su sexo. **A.** Niño. **B.** Niña.

*En niños:*

$$Peso\ niño\ (g) = 3.267,8 - 5,3 * pesomadre - 1.332,5 * 0 + 23,4 * pesomadre * 0 = 3267,8 - 5,3 * pesomadre$$

*En niñas:*

$$Peso\ niña\ (g) = 3.267,8 - 5,3 * pesomadre - 1.332,5 * 1 + 23,4 * pesomadre * 1$$

*Si se extrae el factor común:*

$$Peso\ niña\ (g) = (3.267,8 - 1.332,5) + (23,4 - 5,3) * pesomadre = 1.935,3 + 18,1 * pesomadre$$

*La pendiente de la recta para la variable peso de la madre es claramente diferente según se trate de un niño o de una niña recién nacida. En los niños, el coeficiente de regresión asociado al peso de la madre vale –5,3, mientras que en el caso de las niñas este valor es de +18,1. No solo la magnitud de efecto es diferente para niños y niñas; también es distinto el sentido, con lo que se trata de una interacción negativa. Un incremento en el peso de la madre se asocia con un descenso leve del peso de sus hijos, pero el aumento en el peso de la madre se relaciona con un incremento de mayor magnitud en el peso de sus descendientes mujeres. Este último cálculo permite valorar la magnitud de la interacción y no solo el valor p del término de interacción (37). La diferencia del efecto del peso de la madre sobre el del recién nacido según este sea niño o niña es de gran magnitud; por eso compensa realizar análisis separados en cada subgrupo y sería un error analizarlos juntos.*

Aunque en el modelo 3 ya quedaba claro que las pendientes eran distintas para recién nacidos niños o niñas (se habla de una interacción o modificación del efecto cualitativamente distinta), desde el punto de vista práctico no basta con tener los resultados del modelo 3, ya que, además de los coeficientes, se necesitan sus errores estándar y su significación. Así pues, debe hacerse el análisis segmentado. En ese análisis, tras segmentar la base de datos por sexo del recién nacido, solo haría falta introducir la variable peso de la madre en el modelo. Se obtendrán dos rectas de regresión, una para niños y otra para niñas:

La instrucción que utiliza STATA para segmentar el archivo por sexo es **bysort *sexo:*** Obsérvese que el nuevo modelo de regresión solicitado no incluye la variable independiente *sexo*

```
. bysort sexo: regress peso pesomadre
```

```
-> sexo = niño
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 318774.292 | 1 | 318774.292 | Number of obs = | 81 |
| Residual | 50473291.9 | 79 | 638902.43 | F( 1, 79) = | 0.50 |
| | | | | Prob > F = | 0.4820 |
| | | | | R-squared = | 0.0063 |
| | | | | Adj R-squared = | -0.0063 |
| Total | 50792066.2 | 80 | 634900.828 | Root MSE = | 799.31 |

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | -5.268235 | 7.458316 | -0.71 | 0.482 | -20.11364 | 9.577172 |
| _cons | 3267.754 | 464.8101 | 7.03 | 0.000 | 2342.573 | 4192.936 |

```
-> sexo = niña
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 7441963.65 | 1 | 7441963.65 | Number of obs = | 108 |
| Residual | 41682926.2 | 106 | 393235.153 | F( 1, 106) = | 18.92 |
| | | | | Prob > F = | 0.0000 |
| | | | | R-squared = | 0.1515 |
| | | | | Adj R-squared = | 0.1435 |
| Total | 49124889.9 | 107 | 459111.12 | Root MSE = | 627.08 |

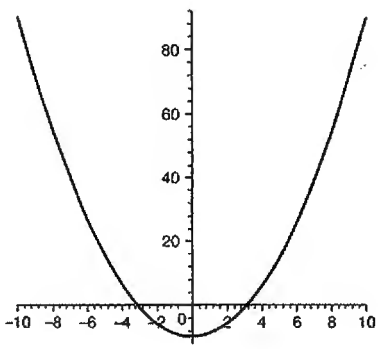
| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| pesomadre | 18.13466 | 4.168614 | 4.35 | 0.000 | 9.869972 | 26.39934 |
| _cons | 1935.213 | 239.6205 | 8.08 | 0.000 | 1460.142 | 2410.284 |

***Interpretación:*** *se aprecia que el efecto del peso de la madre sobre el del recién nacido es solo significativo en las mujeres que dieron a luz a niñas. Tal vez no se encuentre una asociación significativa en la relación entre el peso de la madre y el peso de los niños, porque esa asociación puede que no sea lineal. En ese caso, debería probarse un modelo con un término cuadrático.*

## 12.18. RELACIONES NO LINEALES, MODELOS POLINÓMICOS

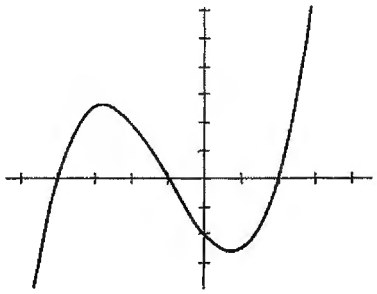
En ocasiones, la relación entre una variable independiente cuantitativa (X) y la variable desenlace (Y) no es lineal (forma de recta), sino que adopta otras distribuciones. A continuación, se muestran algunas de las más comunes:

La función cuadrática:



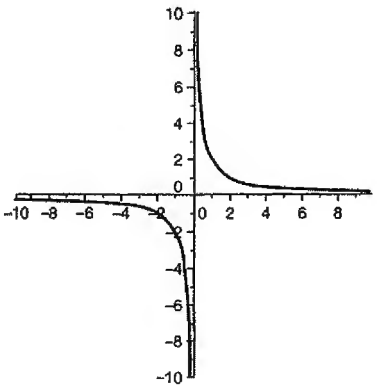

$$y = a + b_1 x_1 + b_2 x_1^2$$

La función cúbica:



$$y = a + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3$$

La función racional (hipérbola):




$$y = a + b_1 \frac{1}{x_1}$$

En estos casos, la variable cuantitativa debería transformarse en una cualitativa ordinal e introducirse de dicha forma en el modelo de regresión múltiple. Otra forma de proceder sería la creación de nuevas variables como $X^2$, $X^3$ o 1/X (según proceda), a través de transformaciones matemáticas sencillas que se introducirán en los modelos matemáticos como tales.

La determinación del tipo de relación entre una variable dependiente Y y una independiente X puede ser establecida de forma aproximada mediante representaciones gráficas (gráficos de dispersión) o bien con modelos matemáticos y comprobando la significación estadística de cada uno de los coeficientes de regresión *(b)* asociados a cada variable del modelo (p. ej., $X^2$, $X^3$, 1/X).

Véase con un ejemplo sencillo. Imagine que se quiere asociar la edad de un grupo de sujetos (variable X, independiente) con su peso (variable dependiente Y). Se considera que la relación edad-peso sigue una función lineal. De esta forma, una mayor edad se asociará siempre con un mayor peso:

$$y = a + b_1 x_1 \qquad peso = a + b * edad$$

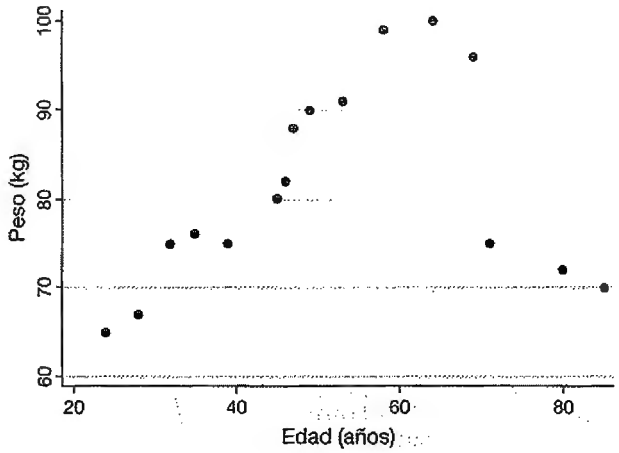



**Figura 12.15** Gráfico de dispersión de las variables X e Y para una asociación no lineal.

Se representará de forma gráfica la distribución de las variables (X e Y) a través de un gráfico de dispersión (fig. 12.15).

El apartamiento de la linealidad parece obvio al observar el gráfico de dispersión. Los datos parecen distribuirse como un modelo cuadrático, con altos pesos en edades intermedias de la vida y pesos más bajos en la juventud y en la vejez. Al realizar un modelo de regresión lineal, se obtiene el siguiente listado en STATA para comprobarlo:

```
. regress peso edad
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 153.779207 | 1 | 153.779207 |
| Residual | 1773.65829 | 14 | 126.689878 |
| Total | 1927.4375 | 15 | 128.495833 |

| | |
|---|---|
| Number of obs | = 16 |
| F( 1, 14) | = 1.21 |
| Prob > F | = 0.2892 |
| R-squared | = 0.0798 |
| Adj R-squared | = 0.0141 |
| Root MSE | = 11.256 |

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edad | .1747119 | .1585786 | 1.10 | 0.289 | -.1654054 | .5148291 |
| _cons | 72.30392 | 8.647351 | 8.36 | 0.000 | 53.75719 | 90.85064 |

No existe una asociación lineal significativa entre la edad de los sujetos y su peso ($p = 0{,}289$)

El resultado obtenido establece que la edad y el peso no se asocian de forma lineal. Esto no significa que no estén asociados, sino que el modelo matemático es diferente a la línea recta.

La representación gráfica parece sugerir una asociación cuadrática, por lo que será necesario crear, en primer lugar, el término $X^2$, es decir, la edad al cuadrado:

$$y = a + b_1x_1 + b_2x_1^2 \qquad peso = a + b_1 edad + b_2 edad^2$$

La variable edad² *(edadcuad)* se crea a través de esta instrucción

```
. gen edadcuad=edad*edad

. regress  peso edad edadcuad
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 1392.94478 | 2 | 696.472392 | Number of obs = | 16 |
| Residual | 534.492716 | 13 | 41.1148243 | F( 2, 13) = | 16.94 |
| | | | | Prob > F = | 0.0002 |
| Total | 1927.4375 | 15 | 128.495833 | R-squared = | 0.7227 |
| | | | | Adj R-squared = | 0.6800 |
| | | | | Root MSE = | 6.4123 |

| peso | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| edad | 3.21729 | .5615269 | 5.73 | 0.000 | 2.004185 | 4.430395 |
| edadcuad | -.0280656 | .0051122 | -5.49 | 0.000 | -.0391099 | -.0170213 |
| _cons | -1.124129 | 14.25343 | -0.08 | 0.938 | -31.91679 | 29.66853 |

Cuando las variables *edad* y *edadcuad* son introducidas simultáneamente en el modelo, el resultado es significativo. Ambos coeficientes explicarían la variación en la variable *peso*

El modelo final, cuadrático, sería:

$$peso = -1{,}12 + 3{,}21 * edad - 0{,}03 * edad^2$$

## 12.19. CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN MÚLTIPLE

En investigación es muy frecuente que haya muchas variables candidatas para su inclusión en un modelo final. El objetivo de cualquier método empleado es seleccionar aquellas variables que den lugar al «mejor» modelo en el contexto científico del problema. No es sencillo, ya que puede haber más de un modelo que sea perfectamente válido y no hay reglas absolutas que se puedan establecer universalmente para construirlos. El criterio para la inclusión de una determinada variable en un modelo puede variar según cada problema y cada disciplina científica.

La aproximación tradicional a la construcción de un modelo implica buscar el modelo más *parsimonioso* (es decir, con el menor número de variables) que explique bien los datos (**10,12,40,41**). El motivo para minimizar el número de variables implicadas es que el modelo resultante probable será numéricamente más estable y más fácilmente generalizable. Al aumentar el número de variables incluidas en un modelo, se incrementarán los errores típicos y el modelo se hará más dependiente de los datos observados. Recientemente ha surgido un movimiento entre los epidemiólogos tendente a incluir en un modelo todas las variables científicamente relevantes, con independencia de su contribución al modelo. Este enfoque se basa en el hecho de que es posible que variables individuales no provoquen una confusión importante, pero, tomadas en su conjunto, puede observarse una confusión considerable en los datos. El principal problema de este enfoque es que el modelo puede acabar «sobreparametrizado» y producir estimaciones inestables. Este sobreajuste se caracteriza, en general, por coeficientes estimados irrealmente grandes y/o errores típicos enormemente aumentados (**42**). Esta consecuencia es problemática, sobre todo en los casos en que el número de variables sea grande en relación con el número de sujetos. Se recomienda que en una regresión lineal múltiple no se introduzca más de una variable independiente por cada 20 sujetos de la base de datos (**12,43**). Así, una base de datos con 400 sujetos admitiría como máximo 20 variables

independientes en el modelo. Esto no significa que sea necesariamente inválido introducir 22 variables.

Los pasos que se explican a continuación pueden ayudar a la selección de variables en un modelo de regresión múltiple, pero nunca deben aplicarse de forma mecánica. El conocimiento biológico del problema, los hallazgos previos en otras investigaciones y la experiencia en el manejo de grandes bases de datos son más importantes para construir un buen modelo que ceñirse automáticamente a unos pasos «tipo receta».

### 12.19.1. Realización de gráficos de dispersión

La realización de gráficos de dispersión entre cada variable independiente X y la variable dependiente Y puede aportar una primera aproximación sobre la relación que se establece entre las variables. Esta relación puede ser lineal, cuadrática, cúbica u otras.

Recuérdese que, en STATA, los gráficos de dispersión pueden obtenerse a través del menú:

**Graphics → Twoway graph (scatter, line, etc.)**

o de instrucciones:

```
twoway (scatter y x)twoway (scatter y x)
(lfit y x)
```

(Con la instrucción **lfit** se ajusta el modelo lineal, STATA traza la línea de predicción de *y* a partir de los datos de *x*.)

### 12.19.2. Hacer un atento examen de la matriz de correlaciones

Esta matriz se obtiene solicitando las correlaciones (bivariables) de todas las posibles parejas de variables independientes entre sí, así como con la variable dependiente. Se obtiene un primer mapa de ruta que orienta sobre los resultados que se encontrarán. Cuando dos variables independientes tengan entre sí un coeficiente de correlación de gran magnitud (>0,90 en valor absoluto), hay que decidir cuál de ellas es la mejor candidata, o bien usar la media de ambas, pero nunca introducir las dos, pues se producirán casi siempre problemas de colinealidad.

Los coeficientes de correlación entre variables (matriz de correlación) pueden calcularse a través de STATA, utilizando el menú:

**Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Correlations and covariances**

o mediante la instrucción:

```
correlate  v1 v2 v3 vp
```

siendo $v_1$-$v_p$ las variables cuantitativas implicadas en el análisis.

### 12.19.3. Realizar regresiones univariantes de cada variable independiente con la dependiente

Se pueden establecer modelos de regresión simple entre cada variable independiente $X_i$ y la variable dependiente o respuesta Y. Como variables independientes se incluyen tanto el efecto principal como las posibles variables de confusión.

Si el tamaño de muestra es elevado (>200), conviene agrupar los valores de las variables independientes cuantitativas $X_i$ en categorías excluyentes e introducir estas como variables independientes en un modelo de regresión simple. Muchas veces resulta preferible establecer

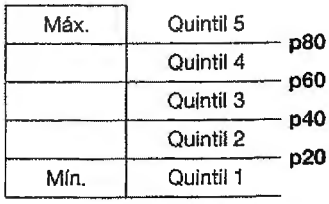



**Figura 12.16** Ejemplo de la distribución de una variable cuantitativa en quintiles (categorización en cinco grupos iguales).

***Tabla 12.9 Ejemplo de relaciones monotónicas y no monotónicas entre una variable X cualitativa ordinal y una variable Y cuantitativa***

| | COEFICIENTES PARA LOS CUARTILES 2-4 | |
|---|---|---|
| VARIABLES *DUMMY* PARA CUARTILES 2 A 4 (1 = REFERENCIA) | RELACIÓN MONOTÓNICA | RELACIÓN NO MONOTÓNICA |
| DUMQ2 | –0,3 | –1,6 |
| DUMQ3 | –1,2 | +0,7 |
| DUMQ4 | –2,0 | –1,1 |

categorías que contengan el mismo número de sujetos en cada una de ellas. La de la clasificación de los sujetos en cuantiles (quintiles, cuartiles, terciles)[5] (fig. 12.16), lo que permite apreciar si la relación entre la variable dependiente y la independiente es monotónica o no. Se dice que la relación es monotónica cuando Y aumenta sistemáticamente cuando X crece, o bien Y decrece sistemáticamente al aumentar X (tabla 12.9).

Cuando la relación es monotónica, puede usarse la variable cuantitativa como tal (lineal), lo que hace que no se consuma más que un grado de libertad. Si la relación no es monotónica, debe introducirse la variable categorizada (cuartiles o, preferiblemente, quintiles), lo que hará que consuma tres o cuatro grados de libertad, respectivamente, en vez de uno. Si el crecimiento del coeficiente a lo largo de los cuartiles tuviese un aspecto exponencial, podría usarse la variable como cuantitativa y probarse a añadir un término cuadrático además del lineal para valorar si así se ajusta mejor el modelo.

### 12.19.4. Uso de regresiones no paramétricas (LOESS)

Una alternativa a las regresiones lineales es la realización de regresiones no paramétricas con alisamiento, por ser ponderadas localmente (LOESS, *Locally Estimated Scatterplot Smoothing*) (44). Este tipo de regresión ofrece la ventaja de no proponer ninguna forma previa para la relación entre la variable dependiente y las variables predictoras. Por tanto, permite establecer asociaciones entre dos variables (una independiente y otra dependiente) no específicamente lineales. El método se basa en calcular una recta de regresión lineal, estimada por el método de los mínimos cuadrados, pero solamente a partir de los individuos más cercanos a un determinado valor de la variable independiente X. El porcentaje de puntos (observaciones) utilizados en el ajuste para ponderar localmente la regresión se denomina entorno. El entorno cambia según el tipo de variables. STATA considera un 80% de los sujetos de la muestra por defecto (**bwidth**(***0.8***)). Esta proporción puede ser modificada por el investigador.

5 Cuando se habla de cuartiles, la muestra es dividida en cuatro partes iguales (puntos de corte: p25, p50 y p75). Si se usan quintiles, la variable se divide en cinco categorías y cada categoría contendrá el 20% de la muestra (puntos de corte: p20, p40, p60 y p80). Al usar terciles, se divide la muestra en tres categorías iguales (puntos de corte: p33, p66). Para calcular los percentiles de una variable, las observaciones recogidas de dicha variable deben ordenarse de menor a mayor y se busca el punto de corte que deja un determinado porcentaje por debajo de él. Por ejemplo, el percentil 50 (p50) será aquel valor que deje por debajo al 50% de los datos para esa variable. Todos los valores anteriores reciben el nombre genérico de cuantiles.

Cuando la nube de puntos para valorar la forma de la relación entre X e Y no es visualmente interpretable por tratarse de una muestra muy grande, puede recurrirse al procedimiento LOESS para apreciar dicha forma.

STATA denomina a este procedimiento *Lowess smoothing*. Puede llevarse a cabo a través del menú:

**Statistics → Nonparametric analysis → Lowess smoothing**

o de la instrucción:

```
lowess y x, bwidth  #(proporción
 de muestra considerada)
```

En el ejemplo de la relación entre el peso del recién nacido y el de la madre, considerando el 75% de los puntos, se observa la representación gráfica recogida en la figura 12.17.

```
lowess peso pesomadre, bw (0.75)
```

***Interpretación:*** *parece existir un incremento del peso de los recién nacidos asociado al aumento de peso de sus madres siempre y cuando estas pesen de 35 a aproximadamente 55 kg. Sin embargo, el incremento del peso de la madre no se traduce en un aumento del peso del recién nacido si la madre pesa de 55 a 85 kg. A partir de los 85 kg de peso, nuevamente, se produce una asociación directa entre el peso de la madre y el del recién nacido.*

### 12.19.5. Selección de variables candidatas para el modelo multivariante

Se ha hablado muchas veces de que, en este proceso, deberían incluirse todas las variables independientes que en el análisis de regresión univariante se aproximen, aunque sea muy remotamente, a la significación estadística. Se ha propuesto el criterio de incluir todas aquellas que den lugar a una $p < 0{,}25$ en el análisis univariante, aunque uno de los mayores riesgos al construir modelos multivariables es el de aplicar mecánicamente este criterio, sin mayor juicio. El hecho es que puede haber variables de importancia biológica conocida (como la edad, el sexo, etc.) que no tengan ese valor $p$ ni se aproximen a la significación, y que se deban introducir en el modelo, aunque lo más probable es que, si no tienen un valor $p$ de 0,25 o menor, no crearán confusión. El problema del enfoque consistente en usar la $p < 0{,}25$ es que ignora la posibilidad de que un grupo de variables

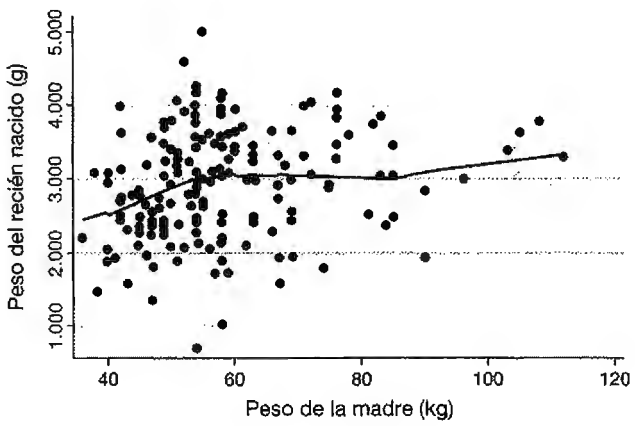



**Figura 12.17** Regresión no paramétrica para el ejemplo de la asociación entre el peso de la madre y el peso del recién nacido.

puede ser un predictor importante, aunque cada una de ellas por separado se asocie muy débilmente a la variable dependiente. Si se prevé que podría suceder así, es conveniente aumentar el valor de *p* univariante para seleccionar variables candidatas.

### 12.19.6. Valorar la contribución de cada variable incluida en el modelo: $R^2$ ajustado

Para valorar si cada variable debe permanecer en el modelo, se examina su significación y, sobre todo, el efecto que tiene su eliminación en el cambio de magnitud de los coeficientes del resto de las variables (posibles factores de confusión). Se deben probar una a una, eliminándolas secuencialmente del modelo y valorando si cambia sustancialmente el coeficiente de la variable independiente principal.

Sin embargo, existen otros criterios que han de tenerse en cuenta para valorar si una variable predictora debe permanecer en el modelo. Uno de ellos es el cambio que se produce en el $R^2$ ajustado tras su inclusión. Si el incremento en el $R^2$ ajustado es de magnitud importante, la variable mejora la capacidad predictora del modelo y debería permanecer en el mismo.

### 12.19.7. Valorar la colinealidad

Por otra parte, en el caso de variables independientes X cuantitativas, debe valorarse también la posible colinealidad (asociación lineal entre variables X que pueden producir coeficientes de regresión distorsionados) a través de la observación tanto de las medidas de tolerancia como del factor de inflación de la varianza. Recuérdese que puede acudirse al menú o a la instrucción de STATA:

**Statistics → Postestimation → Reports and statistics**

```
estat vif
```

### 12.19.8. Valorar relaciones dosis-respuesta que se aparten del modelo lineal

Siempre que sea posible, debe evitarse la inclusión de términos cuadráticos en el modelo de regresión. No obstante, a veces un modelo parabólico (función cuadrática $y = a + b_1x + b_2x^2$) se adapta mejor a los datos que el modelo $y = a + bx$. Esto se valorará comprobando la significación estadística de un término cuadrático añadido al modelo. Si el modelo mejora significativamente, se añadirá el término cuadrático.

### 12.19.9. Inclusión de términos de interacción entre variables

El modelo final resultante con lo realizado hasta ahora es el que se conoce como *modelo de efectos principales.* Sobre él deben valorarse las interacciones, una a una. Para ello se crean términos de producto y se van introduciendo, añadiéndolos de uno en uno al modelo principal. Cuando se comprueba que uno de ellos no es significativo, se elimina y se prueba el siguiente. Para este procedimiento iterativo, que es lento y tedioso, puede usarse, como método de *screening* o cribado, la opción automática *stepwise,* para que sea el programa el que seleccione aquellos términos multiplicativos que sean significativos.

Los métodos *stepwise* son estrategias de selección de variables *por pasos,* que dejan todo el proceso de especificación del modelo en manos del ordenador. Consisten en seleccionar, en cada *paso,* aquella variable que mejor cumple un criterio prefijado basado en la significación estadística de inclusión (para añadirla al modelo) o el criterio de exclusión (para eliminarla). Si el procedimiento parte de un número elevado de variables y estas se van eliminando, el procedimiento se denomina *backward selection* («hacia atrás»). Si, por el contrario, se parte de una sola variable X y se van agregando al modelo nuevas variables independientes, el procedimiento recibe el nombre de *forward selection* («hacia delante»).

En STATA, estos procedimientos por pasos se realizan a través del menú:

**Statistics → Other → Stepwise estimation**

o de las instrucciones:

```
stepwise,pr(.10): regress y x
```

Obsérvese que se solicita a STATA un procedimiento por pasos para regresión múltiple (**regress**).

(La instrucción **pr** (#) especifica el nivel de significación estadística fijado para excluir una variable. En este ejemplo, 0,10. Es el procedimiento hacia atrás *[backward].*)

```
stepwise, pe(.05): regress y x
```

(La instrucción **pe**(#) fija el valor *p* para la inclusión de la variable. En este ejemplo, 0,05. Es el procedimiento hacia delante *[forward].*)

Existen otras instrucciones en STATA que permiten realizar procedimientos por pasos (hacia atrás o hacia delante), incluyendo y excluyendo variables simultáneamente:

```
pr(#) pe(#)
```

```
pr(#) pe(#) forward
```

Por último, una de las principales características que deberían cumplir los modelos de regresión múltiple que contuvieran términos de interacción es que fueran modelos jerárquicos. El modelo jerárquico se define como un modelo tal que, si se elimina un término cualquiera, todos los términos de mayor orden en los que intervenga también deben ser eliminados. Inversamente, si se incluye un término cualquiera, todos sus términos de menor orden también deberán estar presentes en el modelo. Esto implica que si, por ejemplo, un modelo contiene la interacción $X_1$*$X_2$, también deberá contener la variable de exposición $X_1$ y la de control $X_2$. El uso de procedimientos por pasos podría, por tanto, dar lugar a modelos matemáticos no jerárquicos con la exclusión de términos de menor orden del modelo final. Sin embargo, existe una instrucción en STATA que soluciona este problema, permitiendo obtener modelos únicamente jerárquicos:

```
pr(#) hierarchical
```

```
pe(#) hierarchical
```

Los criterios para la inclusión de interacciones son estadísticos (deben ser significativos) y también tienen que ser prácticos, es decir, deben tener sentido desde el punto de vista biológico. Es muy importante representar gráficamente las interacciones para valorar adecuadamente su interpretación en términos de la vida real.

Conviene ser precavido al interpretar las pruebas de interacción. No se debe aceptar como importante una interacción simplemente porque el valor *p* para dicha interacción sea <0,05. Tiene que estar muy clara la significación y, además, ha de haber una magnitud en la diferencia del efecto entre los subgrupos que interaccionan que sea clínicamente relevante.

No obstante, al valorar varios factores a la vez, debe probarse siempre el test de interacción en el modelo, porque la interpretación variará notablemente dependiendo de si existe interacción cualitativa (cambio de dirección del efecto según niveles del modificador) o, al menos, una interacción cuantitativa que sea fuerte. Lo que sí está muy claro es que, cuando la interacción no sea significativa, no tiene sentido dejar el término de producto en el modelo.

### 12.19.10. Comprobar los residuales del modelo

Al igual que en la regresión simple, los valores residuales del modelo deben seguir una distribución normal para poder considerar el modelo como válido. Si no se cumplen los criterios de normalidad de los residuales, deberán llevarse a cabo transformaciones matemáticas de la variable dependiente Y para normalizar los residuales.

## 12.20. ELECCIÓN DEL MEJOR MODELO

Uno de los inconvenientes de la regresión múltiple es la dificultad para escoger el mejor modelo, ya que, a veces, hay varios candidatos adecuados, sobre todo si el número de variables en estudio es elevado, ya que el número de posibles modelos crece exponencialmente con el número de variables (en concreto $2^p - 1$, siendo $p$ el número de variables). El criterio de selección dependerá del objetivo del modelo. Básicamente, un modelo de regresión múltiple se construye con uno de estos tres objetivos:

1. **Control de la confusión.** Construcción de un modelo que mida la relación entre una variable (exposición) y su respuesta en presencia de otras variables que puedan influir. Se busca estimar con el menor sesgo posible los coeficientes de regresión relativos a la variable exposición. La alternativa a la causalidad es casi siempre la confusión, por la existencia de una causa común de la exposición y del efecto. Se debe ajustar por tales variables.
2. **Predicción.** Elaboración de un modelo que permita predecir la respuesta de nuevos sujetos con el menor error posible. En modelos con fines predictivos, el criterio de selección es maximizar el coeficiente de determinación *ajustado* ($R^2$ corregido).
3. **Descripción.** Preparación de un modelo que describa lo mejor posible la muestra estudiada, con el fin de eliminar las variables redundantes. En esta situación se buscan modelos que maximicen el coeficiente de determinación, $R^2$, aunque eliminando aquellas variables que solo producen incrementos mínimos en $R^2$.

## 12.21. LOS MÉTODOS AUTOMÁTICOS POR PASOS *(STEPWISE)* CASI NUNCA ESTÁN INDICADOS

Conviene no usar estos métodos o, en todo caso, emplearlos con precaución como primer paso de *screening* o cribado entre muchas variables o para realizar una selección rápida y automática de los términos de interacción (v. apartado 12.19.9). Las cautelas para no usar estos procedimientos se deben a que pueden llevar a modelos no jerárquicos (excluir de un modelo uno de los términos que intervienen en una interacción) o a excluir alguna de las *dummy* de una variable categórica policotómica, lo que supone una especificación incorrecta del modelo. El primer problema es solucionado por STATA eligiendo instrucciones que permitan obtener únicamente modelos jerárquicos (**pe(#) hierarchical** y **pr(#) hierarchical**).

Como se ha visto, para controlar la confusión el criterio de inclusión o exclusión de una variable de control (ajuste), no debería ser nunca su significación estadística, sino el efecto que su presencia o ausencia tiene sobre los coeficientes del resto de las variables del modelo, sobre todo sobre la variable que mide la exposición.

## 12.22. REGRESIÓN LINEAL MÚLTIPLE EN OTROS PROGRAMAS DISTINTOS DE STATA

### 12.22.1. SPSS

Se utilizará un ejemplo concreto. Supóngase que se desea determinar la contribución de diferentes variables independientes: *sexo* (1 = varón; 2 = mujer), *edad* (años, continua), *tabaco* (0 = no fumador; 1 = fumador actual; 2 = exfumador), consumo de alcohol (*calcohol* = cuartiles de consumo) y nivel de actividad física (*actfisica* (MET-h/semana)[6], continua) (variables X) sobre la variable dependiente índice de masa corporal (*imc* ($kg/m^2$), continua).

SPSS permite realizar regresiones lineales a través de tres procedimientos diferentes, la *regresión lineal*, los *modelos lineales generalizados* y el *modelo lineal general*. En este apartado solo se explicarán el primer y segundo procedimiento, por ser el más usado y el más completo, respectivamente.

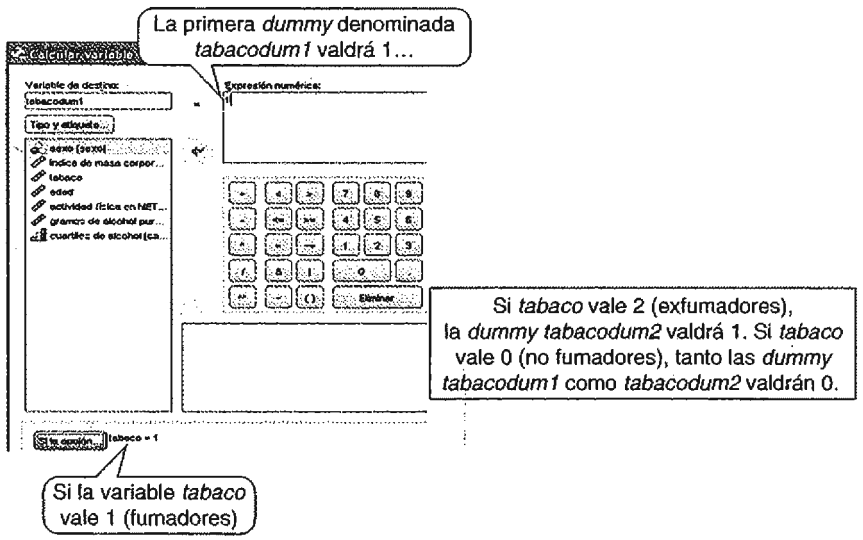
6 MET-h/semana: Equivalentes metabólicos semanales. Un MET-h se define como la cantidad de energía empleada por el organismo estando sentado en 1 hora. Según el nivel de intensidad de la actividad física realizada, sus MET varían (algunos ejemplos: durante 1 hora, andar supone 4 MET; jugar al tenis, 7 MET; jugar al squash, 12 MET).

### 12.22.1.1. Regresión lineal

Deben crearse variables *dummy* «a mano» para todas aquellas variables cualitativas de más de dos categorías implicadas en el análisis. A través del menú:

**Transformar → Calcular**

En el caso del tabaco, como se trata de una variable cualitativa de tres categorías, habrá que crear dos *dummy*, por ejemplo *tabacodum1* (para fumadores) y *tabacodum2* (para exfumadores), considerando los no fumadores como categoría de referencia (*tabaco* = 0):

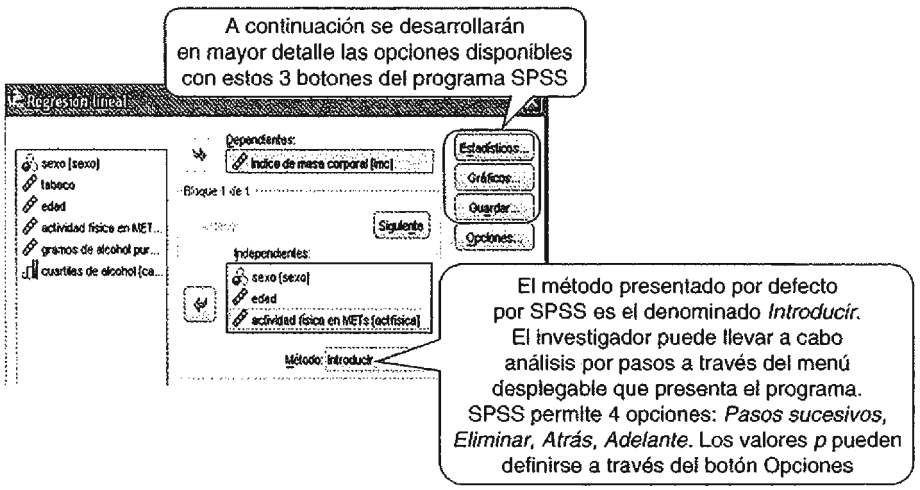



A través de instrucciones:

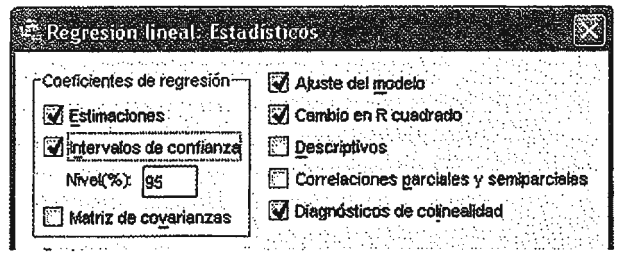
```
COMPUTE tabacodum1=0.

COMPUTE tabacodum2=0.

IF tabaco=1 tabacodum1=1.

IF tabaco=2 tabacodum2=1.
```

Para llevar a cabo la regresión propiamente dicha, se acude al menú:

**Analizar → Regresión → Lineal** (Lineales para versiones de SPSS más avanzadas).

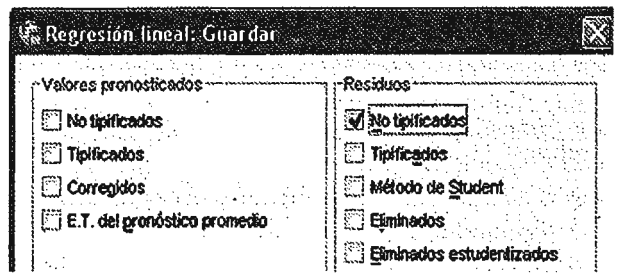


Pulsando el botón ESTADÍSTICOS y en la opción *Coeficientes de regresión* pueden obtenerse diferentes indicadores, como los propios de la regresión *(Estimaciones, Ajuste de modelo)*, los intervalos de confianza de los coeficientes de regresión *(Intervalos de confianza)*, los coeficientes de determinación ($R^2$) y los $R^2$ ajustados *(Cambio en R cuadrado)* o las medidas de multicolinealidad *(Diagnósticos de colinealidad)*.



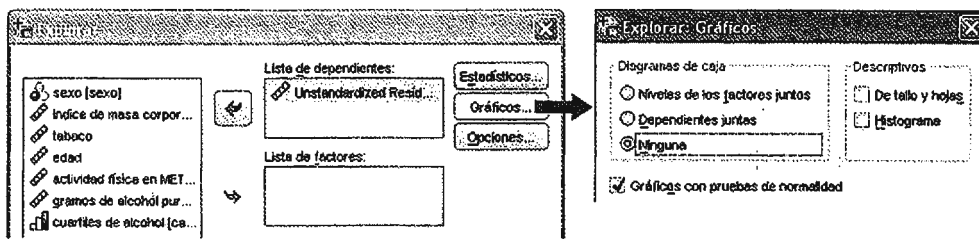

El botón GRÁFICOS permite la creación de gráficas de normalidad (*Q-Q, P-P*) y otras representaciones, como gráficos de dispersión, que permiten comprobar visualmente la adecuación del modelo de regresión.

La creación de residuales puede realizarse a través del botón GUARDAR. Debe marcarse la opción *Residuos* y la casilla *No tipificados.*



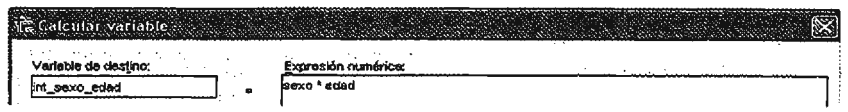

Una vez calculada esta variable (*RES_1;* nombre que por defecto da SPSS a esta nueva variable de la base de datos), el paso siguiente consiste en comprobar su normalidad a través de la instrucción:

**Analizar → Estadísticos descriptivos → Explorar**




Al igual que las *dummy,* las variables de interacción deben ser creadas previamente por el investigador e introducidas a continuación en el modelo de regresión. A través del menú:

**Transformar → Calcular**



A través de instrucciones:

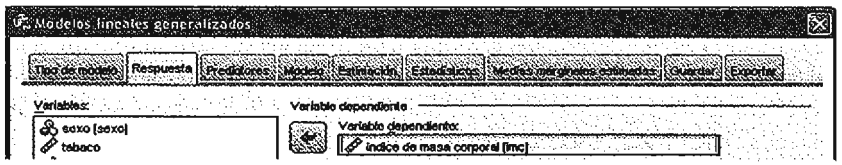
```
COMPUTE int_sexo_edad=sexo*edad.
```

El aspecto de los resultados es similar al obtenido con el programa STATA.

### 12.22.1.2. Modelos lineales generalizados

No es necesaria la creación de variables *dummy*. Este modelo es el más completo. Desde el menú:

**Analizar → Modelos lineales generalizados → Modelos lineales generalizados**



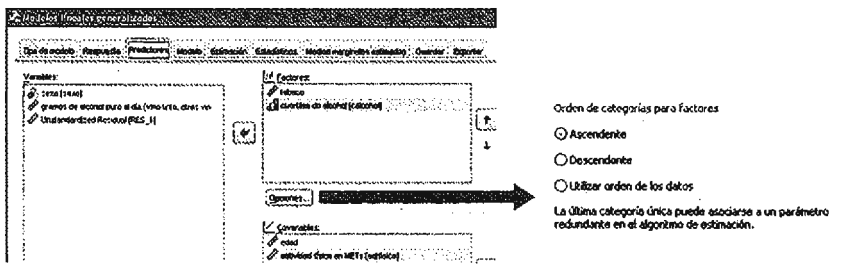

Este cuadro de diálogo resulta bastante complejo, con la presentación de diferentes lengüetas que el investigador debe rellenar. Sin embargo, se intentará simplificar y ajustarlo a las necesidades específicas del investigador.

**Lengüeta 1: Tipo de modelo.** En realidad, un modelo lineal generalizado permite trabajar con variables dependientes no solo cuantitativas, sino también cualitativas. Sin embargo, estas instrucciones no corresponden al presente capítulo. Debe elegirse *Respuesta de escala* y la casilla *Lineal* (es la que presenta SPSS por defecto).

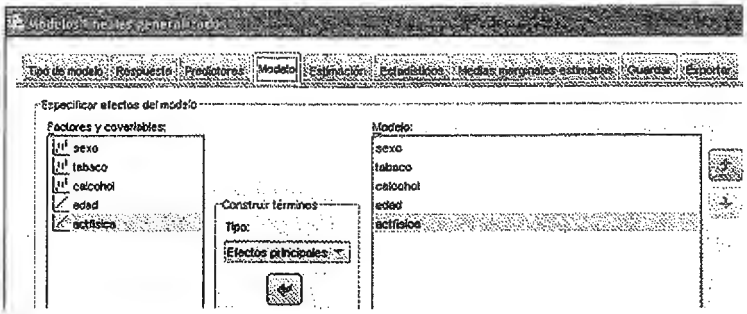
**Lengüeta 2: Respuesta.** Bajo el recuadro *Variable dependiente* debe colocarse Y y olvidarse del resto de apartados de esta sección.

**Lengüeta 3: Predictores.** SPSS solicita las variables independientes cualitativas *(Factores)* y las variables independientes cuantitativas *(Covariables)*.

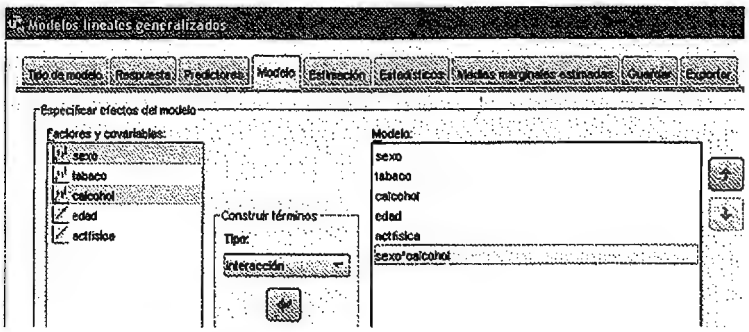
En este caso no es necesario crear anteriormente *dummy* para las variables cualitativas de más de dos categorías. Pueden introducirse directamente en el modelo. Lo que sí se necesita es determinar cuál de las categorías de la variable se considerará como categoría de referencia en los análisis. Generalmente, como categoría de referencia se utiliza la primera o la última de las categorías de la variable. Para ello puede usarse el botón OPCIONES. En la opción *Orden de categorías para factores* se elije si como referencia se desea seleccionar la primera (opción descendente) o la última categoría (opción ascendente). Por defecto, SPSS considera la última categoría de la variable cualitativa como referencia (Ascendente).




**Lengüeta 4: Modelo.** Deben seleccionarse todas las variables independientes y arrastrar el botón flecha bajo la opción *Modelo*. Con respecto al apartado *Construir términos*, por defecto SPSS presenta *Tipo: Efectos principales*. Debe mantenerse.




La especificación *Efectos principales* es válida para modelos sin términos de interacción. Si interesara valorar la posible interacción entre dos variables del modelo sobre la variable dependiente (p. ej., si el efecto del alcohol sobre el IMC se modifica según el sexo), entonces el tipo de modelo para elegir sería *Interacción*. Se seleccionarían las dos variables implicadas (*sexo* y *calcohol*) y se trasladarían con el botón flecha al recuadro de modelo, donde se incluirían como la variable *sexo*calcohol*.




**Lengüeta 5 y 6: Estimación-Estadísticos.** Deben mantenerse las opciones marcadas por defecto por SPSS. No tienen gran importancia para el investigador.

**Lengüeta 7: Medias marginales estimadas.** El ordenador realiza análisis de ANCOVA presentando medias de IMC ajustadas según categorías. Por ejemplo, calcula el IMC medio ajustado por sexo, edad, consumo de alcohol y práctica de deporte en no fumadores, fumadores y exfumadores.

**Lengüeta 8: Guardar.** SPSS permite guardar los residuales del modelo cuya normalidad es necesaria para considerar válido un modelo.

## 12.23. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS (regresión lineal) | SPSS (modelo lineal generalizado) |
|---|---|---|---|
| **Cálculo de coeficientes de regresión** | regress $y\ x_1\ x_2\ x_p$ | REGRESSION /STATISTICS COEFF /DEPENDENT $y$ /METHOD=ENTER $x_1\ x_2\ x_p$. | GENLIN $y$ BY $a_1\ a_2\ a_p\ b_1\ b_2\ b_p$ WITH $c_1\ c_2\ c_p$ /MODEL $a_1\ a_2\ a_p\ b_1\ b_2\ b_p\ c_1\ c_2\ c_p$ /PRINT SOLUTION. |
| **Intervalos de confianza** | p.d. | /STATISTICS CI(95) | p.d. |

## 12.23. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS *(cont.)*

| Procedimiento | STATA | SPSS (regresión lineal) | SPSS (modelo lineal generalizado) |
|---|---|---|---|
| **Cálculo de coeficientes de determinación** | p.d. | **/STATISTICS R** | — |
| **Creación de residuales** | **predict** ***res****, **residuals** | **/SAVE RESID** | **/SAVE RESID** |
| **Comprobación de normalidad de residuales (gráfica)** | **qnorm** ***res*** **pnorm** ***res*** | **/RESIDUALS NORMPROB(ZRESID)** | — |
| **Comprobación de normalidad de residuales (test)** | **swilk** ***res*** **sfrancia** ***res*** **sktest** ***res*** | **/SAVE RESID EXAMINE VARIABLES = RES_1* /PLOT NPPLOT.** | |
| **Tolerancia, factor de inflación de varianza y multicolinealidad** | **estat vif** | **/STATISTICS COLLIN TOL** | — |
| **Variable cualitativa con más de dos categorías (a), categoría de referencia, valor más bajo** | Pueden crearse previamente: **generate** ***dum1=1*** **if** ***a==1*** **generate** ***dum2=1*** **if** ***a==2*** **generate** ***dumn=1*** **if** ***a==n*** **mvencode** ***dum1 dum2 dumn*****, mv(*0*)** | Deben crearse previamente: **compute** ***dum1=0.*** **compute** ***dum2=0.*** **compute** ***dumn=0.*** | **GENLIN** ***y*** **BY** $a_1$ $a_2$ $a_p$ **(ORDER= DESCENDING)** |
| | O a través de la instrucción: **regress** ***y*** **i.*a*** | **if** $a_1$***=1 dum1=1.*** **if** $a_1$***=2 dum2=1.*** **if** $a_1$***=n dumn=1.*** | |
| **Modificación del efecto** (Sirve solo para variables cualitativas dicotómicas y cuantitativas en STATA y en regresión lineal en SPSS) | Debe crearse previamente: **generate** ***intx1x2=*** $x_1$***** $x_2$ | Debe crearse previamente: **compute** ***intx1x2=*** $x_1$***** $x_2$***.*** | **GENLIN** ***y*** **BY** ***a b*** **WITH** ***c*** **/MODEL** ***a b c a*b*** **/PRINT SOLUTION.** |
| (Sirve para todo tipo de variables independientes en los modelos lineales generalizados de SPSS) | **regress** ***y*** $x_1$ $x_2$ ***intx1x2*** | **REGRESSION /STATISTICS COEFF /DEPENDENT** ***y*** **/METHOD= ENTER** $x_1$ $x_2$ ***intx1x2.*** | |

*(Continúa)*

## 12.23. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS *(cont.)*

| Procedimiento | STATA | SPSS (regresión lineal) | SPSS (modelo lineal generalizado) |
|---|---|---|---|
| **Métodos automáticos** | | **REGRESSION /STATISTICS COEFF CI(95) /CRITERIA=PIN(#) POUT(#) /DEPENDENT *y*** | — |
| Hacia atrás (excluir variables) | **stepwise, pr(#): regress *y* $x_1$ $x_2$ $x_p$** | **/METHOD= BACKWARD $x_1$ $x_2$ $x_p$.** | |
| Hacia atrás (modelo jerárquico) | **stepwise, pr(#)hierarchical: regress *y* $x_1$ $x_2$ $x_p$** | — | |
| Hacia adelante (incluir variables) | **stepwise, pe(#): regress *y* $x_1$ $x_2$ $x_p$** | **/METHOD= FORWARD $x_1$ $x_2$ $x_p$.** | |
| Hacia adelante (modelo jerárquico) | **stepwise, pe(#)hierarchical: regress *y* $x_1$ $x_2$ $x_p$** | — | |
| Ambas (inclusión y exclusión de variables) | **stepwise, pr(#) pe(#):regress *y* $x_1$ $x_2$ $x_p$** | **/METHOD= STEPWISE $x_1$ $x_2$ $x_p$.** | |
| *Ejemplo del capítulo:* estimar el peso de un recién nacido (g) a partir del peso de la madre (kg), la presencia de HTA durante el embarazo (no/sí) y su hábito tabáquico (no fuma (0, ref.), exfumadora (1), fumadora (2)). | **regress *peso pesomadre HTA i.tabaco*** | **compute *tab1=0.* compute *tab2=0.* if *tabaco=1 tab1=1.* if *tabaco=2 tab2=1.*** | **GENLIN *peso* BY *HTA tabaco* (ORDER= DESCENDING) WITH *pesomadre* /MODEL *pesomadre HTA tabaco* /PRINT SOLUTION /SAVE RESID.** |
| | | **REGRESSION /STATISTICS COEFF CI(95) /DEPENDENT *peso* /METHOD=ENTER *pesomadre HTA tab1 tab2* /SAVE RESID.** | **EXAMINE VARIABLES= RES_1 /PLOT NPPLOT.** |
| Obtener coeficientes de regresión e intervalos de confianza al 95% y guardar residuales. Comprobar la normalidad de las residuales a través de un test de normalidad | **predict *res*, re swilk *res* sfrancia *res* sktest *res*** | **EXAMINE VARIABLES= RES_1 /PLOT NPPLOT.** | |

## 12.23. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS *(cont.)*

| Procedimiento | STATA | SPSS (regresión lineal) | SPSS (modelo lineal generalizado) |
|---|---|---|---|
| Valorar, a continuación, la interacción entre el tabaco y el peso de la madre con respecto al peso del niño. Ajustar por HTA | g *tab1=1* if *tabaco==1*<br>g *tab2=1* if *tabaco==2*<br>mvencode *tab1 tab2*, mv(*0*)<br>g *tab1pm=tab1*pesomadre*<br>g *tab2pm=tab2* pesomadre* | compute *tab1pm=tab1* pesomadre.*<br>compute *tab2pm=tab2* pesomadre.* | GENLIN *peso* BY *HTA tabaco* (ORDER= ASCENDING) WITH *pesomadre* /MODEL *pesomadre HTA tabaco tabaco* pesomadre* /PRINT SOLUTION. |
| | regress *peso pesomadre HTA tab1 tab2 tab1pm tab2pm* | REGRESSION /STATISTICS COEFF CI(95) /DEPENDENT *peso* /METHOD=ENTER *pesomadre HTA tab1 tab2 tab1pm tab2pm.* | |

*Debe especificarse a STATA el nombre que quiere dársele a la variable residual, por ejemplo *res*. Por defecto, SPSS crea la variable residual con etiqueta *RES_1*. Se deberá cambiar el nombre de la variable si así se desea (p. ej., por *res*). HTA, hipertensión arterial; p.d., por defecto (el programa calcula el parámetro sin añadir instrucciones); (#), nivel de significación estadística que se fija para excluir/incluir una variable en el modelo; *a*, variables independientes cualitativas de más de dos categorías; *b*, variables independientes cualitativas dicotómicas; *c*, variables independientes cuantitativas; *x*, variables independientes cuantitativas o cualitativas dicotómicas (*b* o *c*).

## REFERENCIAS


1. Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. J Epidemiol Community Health 2003;57(8):634-8.
2. De Irala J, Martínez-González MA, Guillen Grima F. ¿Qué es una variable de confusión? Med Clin (Barc) 2001;117:377-85. Fe erratas: Med Clin (Barc) 2001;117:775.
3. Sánchez-Villegas A, Toledo E, de Irala J, Ruiz-Canela M, Pla-Vidal J, Martínez-González MA. Fast-food and commercial baked goods consumption and the risk of depression. Public Health Nutr 2012;15(3):424-32.
4. De Irala J, Martínez-González MA, Guillén-Grima F. ¿Qué es una variable modificadora del efecto? Med Clin (Barc) 2001;117:297-302.
5. Corella D, Tai ES, Sorlí JV, Chew SK, Coltell O, Sotos-Prieto M, et al. Association between the APOA2 promoter polymorphism and body weight in Mediterranean and Asian populations: replication of a gene-saturated fat interaction. Int J Obes (Lond) 2011;35(5):666-75.

6. Estruch R, Martínez-González MA, Corella D, Salas-Salvadó J, Ruiz-Gutiérrez V, Covas MI, et al. PREDIMED Study Investigators. Effects of a Mediterranean-style diet on cardiovascular risk factors: a randomized trial. Ann Intern Med 2006;145(1):1-11.

7. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel; 2008.

8. Sánchez-Villegas A, Field AE, O'Reilly EJ, Fava M, Gortmaker S, Kawachi I, et al. Perceived and actual obesity in childhood and adolescence and risk of adult depression. J Epidemiol Community Health 2013;67(1):81-6.

9. Martínez-González MA, Guillén-Grima F, De Irala J, Ruiz-Canela M, Bes-Rastrollo M, Beunza JJ, et al. The Mediterranean diet is associated with a reduction in premature mortality among middle-aged adults. J Nutr 2012;142(9):1672-8.

10. Hosmer DW, Lemeshow S. Applied logistic regression. 3rd ed. Hoboken: John Wiley & Sons; 2013.

11. Jewell NP. Statistics for epidemiology. Boca Raton: Chapman & Hall/CRC Press; 2004.

12. Katz MH. Multivariable Analysis. 2nd ed. New York: Cambridge University Press; 2006.

13. Cox DR. Regression model and life tables. J Roy Statist Soc B 1972;34:187-220.

14. Collett D. Modelling survival data in medical research. London: Chapman & Hall; 1994.

15. Altman DG, Goodman SN. Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. JAMA 1994;272:129-32.

16. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. Berlin: Springer Verlag; 1997.

17. Hosmer DW, Lemeshow S. Applied Survival Analysis: Regression Modelling of Time to Event Data. New York: Wiley; 1999.

18. Healy MJR. Multiple regression. Arch Dis Child 1995;73:177-81.

19. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151-69.

20. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8:551.

21. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995;6:356-65.

22. Weinberg CR. How bad is categorization? Epidemiology 1995;6:345-7.

23. Motulsky H, Christopoulos A. Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting. Oxford: Oxford University Press; 2004.

24. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Multivariable Methods. Pacific Grove: Duxbury Press; 1998.

25. Chatterjee S, Hadi AS. Regression Analysis by Example. 4th ed. New York: Wiley; 2006.

26. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiologic confounding. Int J Epidemiol 1986;15:413-9.

27. Weinberg CR. Towards a clearer definition of confounding. Am J Epidemiol 1993;137:1-8.

28. Hernan MA. Conocimiento experto, confusión y métodos causales. Gac Sanit 2001;15(Suppl 4): S44-8.

29. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248-52.

30. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg: Aspen Publishers; 1999.

31. Cole SR, Hernan MA. Fallibility in estimating direct effects. Int J Epidemiol 2002;31:163-5.

32. Manson JE, Willett WC, Stampfer MJ, Colditz GA, Hunter DJ, Hankinson SE, et al. Body weight and mortality among women. N Engl J Med 1995;333:677-85.

33. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10:37-48.

34. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2003;14:300-6.

35. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176-84.

36. Miettinen OS. Causal and preventive interdependence. Elementary principles. Scand J Work Environ Health 1982;8:159-68.

37. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not p values. BMJ 1996;313:808.

38. Lagakos SW. The challenge of subgroup analyses – reporting without distorting. N Engl J Med 2006;354:1667-9.

39. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet 2005;365:1657-61.

40. Rothman KJ, Greenland S. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

41. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariable methods. Belmont: Duxbury Press; 1988.

42. De Irala J, Fernández-Crehuet Navajas R, Serrano del Castillo A. Abnormally broad confidence intervals in logistic regression: interpretation of results of statistical programs. Rev Panam Salud Publica 1997;1:230-4.

43. Feinstein AR. Multivariable analysis: an introduction. New Haven: Yale University Press; 1996.

44. Sánchez-Cantalejo E, Ocaña-Riola R. Actualizaciones en regresión: suavizando las relaciones. Gac Sanit 1997;11:24-32.

# REGRESIÓN LOGÍSTICA 13

*A. Sánchez-Villegas, M. Bes-Rastrollo, M. Á. Martínez-González*

## 13.1. INTRODUCCIÓN

Se usa la regresión logística binaria cuando se desea conocer el modo en que diversos factores (variables cuantitativas o categóricas) se asocian simultáneamente a una variable cualitativa o categórica dicotómica. Aunque hay otros modelos, como la regresión logística nominal para variables dependientes categóricas con más de dos categorías o la regresión logística ordinal, estas modalidades se usan muy poco en medicina.

### 13.1.1. Función logística

Si se clasifica el valor de la variable respuesta como 0 cuando no se presenta el suceso y con el valor 1 cuando está presente, y se pretende predecir o pronosticar la presencia del suceso o enfermedad a partir de un posible factor de riesgo, se podría caer en la tentación de utilizar un modelo lineal:

$$\textit{Valor pronosticado (suceso)} = a + b\ \textit{factor}$$

y estimar, a partir de los datos, por el procedimiento habitual de mínimos cuadrados, los coeficientes $a$ y $b$ de la ecuación. Aunque pudiese aplicarse desde el punto de vista fáctico, emplear en este caso el modelo lineal no sería válido. Al tratarse de una probabilidad, los únicos valores posibles que podría tomar la variable dependiente («suceso») en los datos serán 0 y 1. Tendrá valor 0 cuando el suceso no se produzca y valor 1 cuando esté presente. El problema que surgiría al emplear regresión múltiple sería que los valores que teóricamente podría adoptar la variable dependiente Y no estarían limitados al rango de 0 a 1, sino que oscilarían desde $-\infty$ hasta $+\infty$.

¿Cómo conseguir obtener una variable dependiente Y que esté comprendida entre $-\infty$ y $+\infty$ y que represente la probabilidad de presentar el suceso o enfermedad? Este dilema se resuelve a través del empleo de la función logística. Esta función describe la forma matemática en que está basado el modelo logístico para conseguir predecir un suceso (Y) a partir de un factor de riesgo X. Según el modelo logístico, la probabilidad «p» de que se diera un suceso sería:

$$\text{Valor pronosticado} = p = \frac{e^{(a+bx)}}{1+e^{(a+bx)}} = \frac{1}{1+e^{-(a+bx)}}$$

donde $(a + bx)$ puede tomar ya valores entre $-\infty$ y $+\infty$. Sin embargo, el rango de $p$ (valor pronosticado) puede oscilar únicamente entre 0 y 1. Es posible calcular también la probabilidad de que no se produzca el evento, que sería el complementario del valor pronosticado $(1 - p)$:

$$1-p = 1-\frac{1}{1+e^{-(a+bx)}} = \frac{1+e^{-(a+bx)}-1}{1+e^{-(a+bx)}} = \frac{1}{e^{(a+bx)}+1}$$

### 13.1.2. Transformación LOGIT

Esta fórmula compleja puede ser simplificada a través de una transformación algebraica en la siguiente expresión:

$$\ln\left(\frac{p(suceso)}{1-p(suceso)}\right) = a + bx$$

Puede apreciarse su parecido con la regresión múltiple. La diferencia reside en que se ha sustituido la variable dependiente Y por otra expresión. En la regresión logística, la variable dependiente no tiene un sentido numérico en sí misma, sino que es el logaritmo neperiano (ln) de la probabilidad *(p)* de que ocurra un suceso dividido por la probabilidad de que no ocurra $(1 - p)$. Al cociente $p/1 - p$ en inglés se le llama *odds,* que se ha traducido por «ventaja».

$$Odds = \frac{p}{1-p}$$

Así, la ecuación podría escribirse también como:

$$\ln(odds) = a + bx$$

La expresión de la variable dependiente $\ln(p/1 - p)$ se conoce por $\text{logit}(p)$. Por consiguiente:

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds) = \text{logit}(p)$$

La transformación logarítmica es necesaria para adaptarse a un fenómeno como la probabilidad cuyos límites teóricos son tan estrechos como 0 y 1. En cambio, los límites teóricos de $\ln(odds)$ oscilan desde $-\infty$ hasta $+\infty$.

## 13.2. CONCEPTOS DE *ODDS, ODDS RATIO* Y RIESGO RELATIVO

Una *odds* se calcula dividiendo el número de individuos que tienen una característica por el número de quienes no la tienen.

Imagine que en una muestra de 100 pacientes que han recibido un fármaco se ha alcanzado éxito en 75 de ellos. Si se divide el número de quienes se curaron (75) por el número de quienes no lo consiguieron (25), se obtendrá la *odds* de curación para ese tratamiento, que valdría 3. También se llegaría al mismo resultado al dividir las proporciones o tantos por ciento (*odds* = 75%/25% = 0,7 5/0,25 = 3). ¿Cómo se interpreta una *odds* = 3 en el ejemplo? Se entendería que, por cada paciente en que no se alcanzó el éxito terapéutico, hay tres en que se logró; es decir, con ese tratamiento la probabilidad de éxito es tres veces mayor que la de fracaso. La ventaja para curarse se cifra en 3. Esta ventaja es la *odds,* tal como se muestra en la figura 13.1.

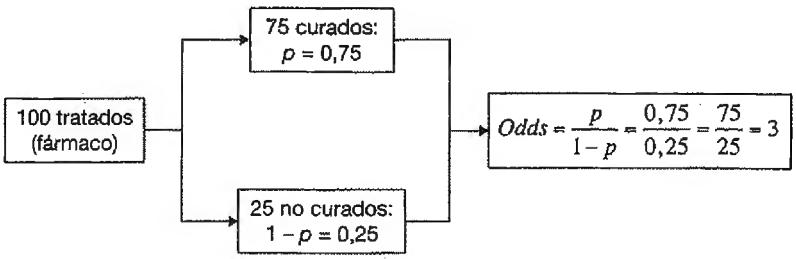



**Figura 13.1** Concepto de ventaja *(odds):* 75 curaciones en 100 pacientes tratados con un fármaco.

Para transformar una proporción *(p)* en una *odds,* o viceversa, se aplican los cálculos que se indican a continuación (fig. 13.2). Si la *odds* de curarse con un tratamiento (v. fig. 13.1) es de 3, la proporción sería:

$$\text{Proporción} = \frac{3}{1+3} = \frac{3}{4} = 0,75\ (75\%)$$

Las proporciones y las *odds* expresan lo mismo, pero en dos escalas numéricas distintas: las proporciones oscilan entre 0 y 1, y las *odds,* entre 0 e infinito. A veces interesa pasar de una escala a otra, para lo cual se utilizan las expresiones que se han señalado:

$$odds = p\,/\,(1-p)$$
$$p = odds\,/\,(1+odds)$$

Definido el concepto de *odds,* hay estudiar qué es una *odds ratio* (**1-7**). La traducción más lógica es *razón de odds o razón de ventajas.* No obstante, el término *odds ratio* ha recibido diversas traducciones al castellano: razón de oportunidades, razón de posibilidades, oportunidad relativa, razón de probabilidades o razón de productos cruzados, e incluso algo tan extraño como «razón de momios». Una buena opción que sirve para evitar confusiones y se ha hecho mayoritaria (**5**) consiste en aplicar directamente el término inglés y decir siempre *odds ratio* (abreviadamente, OR).

¿Qué es una OR? Un cociente entre dos *odds.* La división de una *odds* por otra *odds* es una razón de *odds* u *odds ratio.* En el ejemplo anterior (v. fig. 13.1), de 100 pacientes tratados médicamente con un fármaco se curaron 75 (*odds* = 75/25 = 3).

Imagine ahora que otros 90 pacientes se trataron quirúrgicamente y se alcanzó el éxito terapéutico en 81 de ellos. La *odds* esta vez sería de 9 (*odds* = 81/9 = 9), como muestra la figura 13.3.

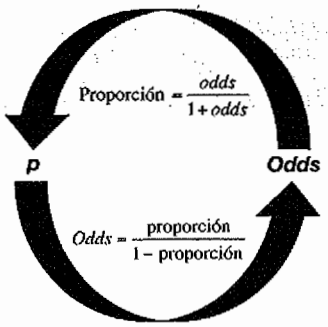



**Figura 13.2** Transformación proporción-*odds*; *odds*-proporción.

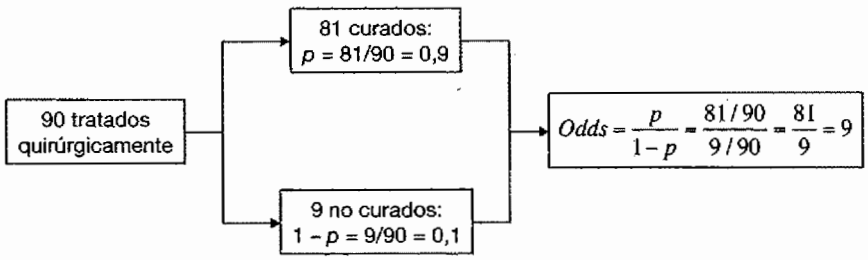



**Figura 13.3** *Odds* de curación si se producen 81 éxitos entre 90 pacientes tratados quirúrgicamente.

La OR se obtiene al dividir la *odds* de un tratamiento por la *odds* de otro:

$$OR = \frac{Odds_{T.\ QUIRÚRGICO}}{Odds_{FÁRMACO}} = \frac{9}{3} = 3$$

Se obtiene una OR = 3 para el éxito terapéutico del tratamiento quirúrgico con respecto al tratamiento con el fármaco, como muestra la figura 13.4. Una OR, por tanto, es el cociente o razón entre dos *odds* y carece de unidades de medida.

***Interpretación:*** *si la OR vale 3, se interpreta que el tratamiento quirúrgico ofrece una ventaja terapéutica tres veces superior al tratamiento con el fármaco.*

Para poder interpretar una OR, es necesario tener siempre en cuenta cuál es el factor o variable predictora que se estudia y cuál es el resultado o desenlace. En este caso, el factor es el tratamiento y la respuesta o desenlace es el éxito terapéutico. La OR no tiene interpretación absoluta, siempre es *relativa*. Una OR de 3 se interpreta como una ventaja tres veces superior de una de las categorías (la categoría quirúrgica en el factor tratamiento) *relativamente a la otra* categoría (fármaco) para alcanzar el desenlace o resultado (éxito terapéutico).

El valor nulo para la OR es el 1. Una OR = 1 implica que las dos categorías comparadas son iguales. El valor mínimo posible es 0 y el máximo teóricamente posible es infinito.

Una OR inferior a la unidad se interpreta como un caso en que el desenlace es menos frecuente en la categoría o grupo que se ha elegido como de interés *con respecto* al otro grupo o categoría de referencia. La *odds* del grupo de interés se debe colocar siempre en el numerador, y la de referencia, en el denominador.

El ejemplo de la figura 13.4 también podría representarse en forma de tabla, del modo que muestra la figura 13.5.

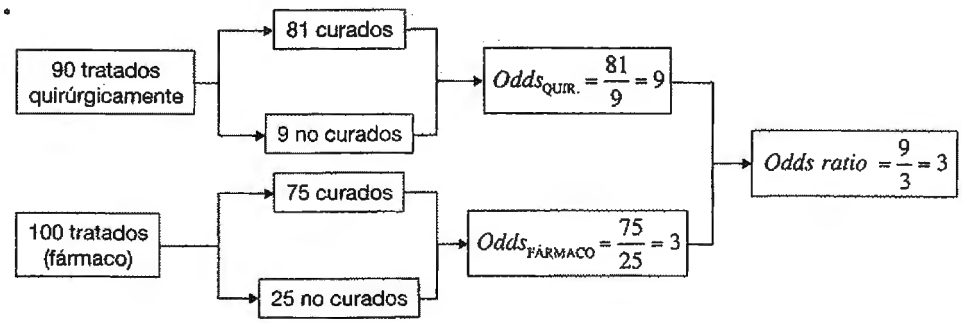



**Figura 13.4** Una *odds ratio* se obtiene al dividir una *odds* entre otra *odds*.

Numerador = 81 × 25

| | Se curan | No se curan | TOTAL |
|---|---|---|---|
| Tratamiento quirúrgico | 81 | 9 | 90 |
| Tratamiento con fármaco | 75 | 25 | 100 |
| TOTAL | 156 | 34 | 190 |

Denominador = 75 × 9

**Figura 13.5** El cálculo de una *odds ratio* se obtiene en una tabla por la razón de productos cruzados.

La OR se calcula por la razón de productos cruzados:

$$OR = \frac{(81)(25)}{(75)(9)} = 3$$

Generalizando, se podrían usar las notaciones de la tabla 13.1.
En esta disposición de la tabla, la OR se calcula por el producto cruzado:

$$OR = \frac{ad}{bc}$$

De todos modos, al manejar una OR se presenta una aparente incongruencia con nuestro modo habitual de pensar. ¿Hasta qué punto es verdad que el tratamiento quirúrgico es tres veces mejor que el farmacológico? El modo habitual de razonar es que, si el tratamiento quirúrgico ha curado al 90% y el farmacológico solo al 75%, existe una razón de probabilidades de curarse de 1,2:

$$\frac{90\%}{75\%} = \frac{0,9}{0,75} = 1,2$$

En epidemiología, este cociente, que surge de dividir proporciones ($p_A/p_B$), se conoce como «riesgo relativo» o «razón de riesgos» (RR).

$$RR = \frac{p_A}{p_B}$$

El RR es una medida de asociación entre una exposición y un desenlace que mide por cuánto se multiplica el riesgo de sufrir un evento por el hecho de estar expuesto al factor de estudio. Este indicador de riesgo es utilizado habitualmente en estudios epidemiológicos de cohortes (**8**). Como en el caso de la OR, el valor nulo para el RR es 1. Un RR = 1 implica que las dos categorías comparadas son iguales. El valor mínimo posible es 0 y el máximo teóricamente posible, infinito.

La OR solo se aproxima al RR cuando el suceso es raro y ocurre en menos del 10% de los sujetos ($p < 0,1$), por lo que su interpretación debe matizarse en función de lo frecuente que sea el suceso que se usa como respuesta o variable dependiente (**4-6**). Cuanto más común será un suceso, más se separará la OR del RR. En estos casos, la OR supone siempre una sobreestimación del RR.

En el ejemplo resumido en la tabla 13.2, el acontecimiento de desenlace o respuesta (variable dependiente) es el cáncer de páncreas. Se compara su ocurrencia en un grupo de 10.000 fumadores y en otro de 20.000 no fumadores. Afortunadamente, el cáncer de páncreas es un fenómeno raro, incluso entre los fumadores. Hubo 10 individuos entre los 10.000 fumadores que desarrollaron cáncer, y solo 10 de los 20.000 no fumadores.

***Tabla 13.1 Disposición de una tabla para el cálculo de una OR***

| | RESPUESTA | |
|---|---|---|
| FACTOR | SÍ | NO |
| Categoría A | a | b |
| Categoría B | c | d |

***Tabla 13.2 La OR se aproxima bien al RR solo cuando el fenómeno o evento (variable dependiente) es raro***

| | CÁNCER DE PÁNCREAS | NO CÁNCER DE PÁNCREAS | TOTAL |
|---|---|---|---|
| Fumadores | 10 | 9.990 | 10.000 |
| No fumadores | 10 | 19.990 | 20.000 |
| Total | 20 | 29.980 | 30.000 |

En esta tabla, la OR valdría:

$$OR = \frac{(10)(19.990)}{(9.990)(10)} = 2{,}001$$

El RR sería:

$$RR = \frac{10/10.000}{10/20.000} = 2{,}000$$

Puede comprobarse que apenas existen diferencias entre ambos estimadores, que en este caso son intercambiables. Esto se debe a que el fenómeno estudiado es raro.

Se ha hecho esta larga introducción sobre la OR porque es uno de los estimadores básicos que proporciona la regresión logística y el que más interés tiene para ser interpretado y aplicado (5).

La regresión logística se emplea habitualmente en uno de los diseños epidemiológicos más utilizados: los estudios de casos y controles. En los estudios de casos y controles se utiliza la regresión logística para calcular OR ajustadas por factores de confusión. Estas OR representan una aproximación razonable al RR, ya que los análisis de casos y controles suelen elegirse como el diseño más apto para estudiar enfermedades raras. Al ser rara la enfermedad bajo estudio, la OR es muy parecida al RR.

Sin embargo, es conveniente saber que hay un tipo de estudios de casos y controles, denominado casos y controles apareados, en el que no se debe aplicar la regresión logística convencional, sino que se ha de utilizar un tipo especial de regresión logística: la *regresión logística condicional* (v. apartado 13.16).

## 13.3. EJEMPLO RIDÍCULAMENTE SENCILLO DE REGRESIÓN LOGÍSTICA BINARIA UNIVARIANTE: INTERPRETACIÓN

### 13.3.1. Cálculo de una OR (manual)

En la tabla 13.3 se recogen los datos de un ejemplo en el que se valora si la exposición al alcohol tiene algún tipo de influencia en la probabilidad de accidente de tráfico.

Se calculará la OR de accidente tras beber. Puede obtenerse como un producto cruzado:

$$OR = \frac{24 \times 40}{4 \times 12} = 20$$

La *odds* de accidente tras haber bebido es 20 veces superior a la *odds* de accidente si no se ha bebido. También puede calcularse dividiendo una *odds* entre otra:

- La *odds* de accidente si se bebió alcohol es 24/12 = 2.
- La *odds* de accidente si *no* se bebió alcohol es 4/40 = 0,1.

***Tabla 13.3 Tabla de contingencia (2 × 2) de exposición a alcohol y accidente de tráfico***

| | | ALCOHOL | | |
|---|---|---|---|---|
| | | SÍ | NO | TOTAL |
| AT | Sí | 24 | 4 | 28 |
| | No | 12 | 40 | 52 |
| Total | | 36 | 28 | 80 |

- Por tanto:

$$\text{OR} = \frac{odds_{\text{AT}|\text{alcohol}}}{odds_{\text{AT}|\text{NO alc}}} = \frac{\left(\frac{24}{12}\right)}{\left(\frac{40}{4}\right)} = \frac{2}{0{,}1} = 20$$

### 13.3.2. Ecuación logística: interpretación

Si $Odds_{AT}$ es la *odds* de accidente, se puede ajustar el siguiente modelo de regresión logística:

$$\ln\left(\frac{p(suceso)}{1-p(suceso)}\right) = \ln(odds) = a + bx$$

$$\ln\left(p_{\text{AT}} / 1 - p_{\text{AT}}\right) = \ln\left(odds_{\text{AT}}\right) = a + b\,\text{alcohol}$$

Si la variable *alcohol* vale 1 en bebedores y 0 en no bebedores, los modelos serán:

Beben: $\ln(odds_{\text{AT}\,|\,\text{alcohol}}) = a + b*1 = a + b$

No beben: $\ln(odds_{\text{AT}\,|\,\text{NO alc}}) = a + b*0 = a$

Si ahora se restan ambas ecuaciones, se obtiene:

$$\ln(odds_{\text{AT}|\text{alcohol}}) - \ln(odds_{\text{AT}|\text{NO alc}}) = a + b - a = b$$

Es decir:

$$\ln\left(\frac{odds_{\text{AT}\,|\,\text{alcohol}}}{odds_{\text{AT}\,|\,\text{NO alc}}}\right) = b$$

Lo escrito dentro del paréntesis es precisamente la OR. Por lo tanto:

$$\ln(\text{OR}) = b$$

La interpretación más sencilla de la regresión logística es que cada coeficiente de regresión $b_i$ expresa el logaritmo neperiano de la OR de que ocurra un fenómeno por unidad de cambio de la variable independiente. En el ejemplo, una «unidad» de cambio es comparar a bebedores frente a no bebedores:

$$\text{OR} = \frac{odds_{\text{AT}\,|\,\text{alcohol}}}{odds_{\text{AT}\,|\,\text{NO alc}}}$$

$$b = \ln(\text{OR})$$

Tomando antilogaritmos, se obtendría:

$$\text{OR} = \text{antilog}(b) = e^b$$

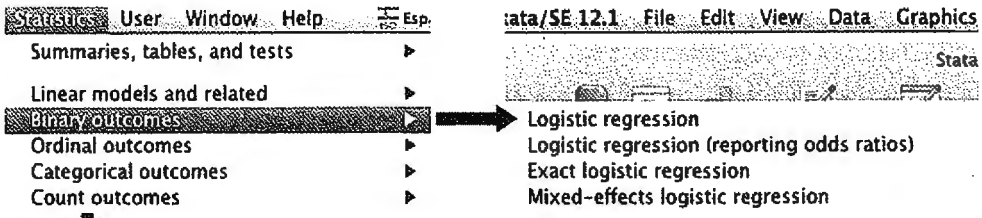
### 13.3.3. Estimaciones con STATA

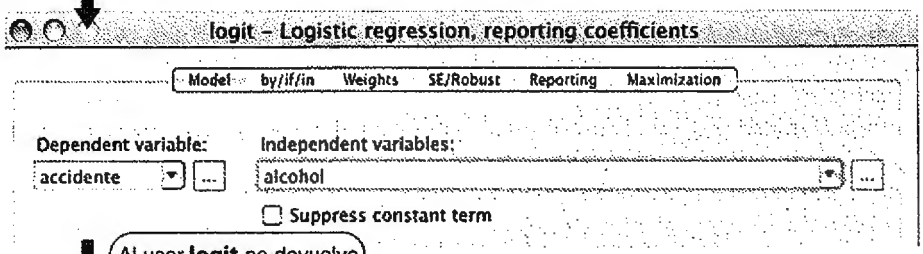
Cuando se ajusta un modelo de regresión logística con STATA, si hubiese más de una variable independiente, como suele suceder, el ordenador devolverá coeficientes $b_i$ para cada una de las variables independientes, $x_i$, que pueden considerarse predictores del suceso que constituye la respuesta o variable dependiente.

En el ejemplo presentado *solo se dispone de una* variable independiente, *alcohol*. La variable dependiente siempre es una sola, en ese caso es *accidente*. Se debe introducir la variable dependiente como 0 = no y 1 = sí. Una vez introducidos los datos, se pide al ordenador que ajuste un modelo logístico para pronosticar la probabilidad de accidente según se haya bebido o no.

Puede obtenerse a través del menú de STATA:

**Statistics → Binary outcomes → Logistic regression**




Al usar **logit** no devuelve OR, sino logaritmos

```
. logit accidente alcohol

Iteration 0:   log likelihood = -51.795731
Iteration 1:   log likelihood = -36.896952
Iteration 2:   log likelihood = -36.325057
Iteration 3:   log likelihood = -36.318501
Iteration 4:   log likelihood = -36.318498

Logistic regression                             Number of obs   =         80
                                                LR chi2(1)      =      30.95
                                                Prob > chi2     =     0.0000
Log likelihood = -36.318498                     Pseudo R2       =     0.2988

------------------------------------------------------------------------------
   accidente |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   2.995732   .6324554     4.74   0.000     1.756142    4.235321
       _cons |  -2.302584   .5244043    -4.39   0.000    -3.330398   -1.274771
------------------------------------------------------------------------------
```

*b* = logaritmo de la *odds ratio*

*a* = logaritmo de la *odds* en no expuestos

o a través de la instrucción:

```
logit accidente alcohol
```

STATA permite obtener, a través de un análisis de regresión logística, además de los coeficientes de regresión ($b_i$), los valores de OR ($e^b$) para la comparación entre dos categorías de una variable empleando la instrucción **logit** con la opción **or:**

```
logit accidente alcohol, or
```

Es más eficiente usar la instrucción **logistic.** También pueden obtenerse OR desde la instrucción del menú del programa:

**Statistics → Binary outcomes → Logistic regression (reporting odds ratios)**

En el ejemplo de la tabla 13.3, las salidas correspondientes a este análisis que proporciona la *odds ratio* con STATA sería:

```
. logistic accidente alcohol

Logistic regression                                Number of obs   =         80
                                                   LR chi2(1)      =      30.95
                                                   Prob > chi2     =     0.0000
Log likelihood = -36.318498                        Pseudo R2       =     0.2988

------------------------------------------------------------------------------
   accidente | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   19.99999   12.6491     4.74   0.000     5.790055    69.08388
       _cons |   .1000001   .0524405    -4.39   0.000     .0357789    .279495
------------------------------------------------------------------------------
```

STATA obtiene directamente los valores de las OR con la instrucción **logistic**

exp(a) = 0,10 es la ***odds*** **de accidente en los que no beben** 4/40 = 0,10

***Interpretación:*** *el coeficiente de regresión (b) (Coef.) para beber vale 2,996 y su error estándar es 0,632. Elevando el número e al coeficiente b, se obtiene la* OR:

$$OR = e^{b} = e^{2,996} = 20$$

Como *puede comprobarse, OR* = $e^{2,996}$ = *20 coincide con la OR que se había calculado antes. También coincide con la OR obtenida igualmente por STATA tras aplicar la instrucción* ***logistic****, y que se interpreta como que la odds de accidente después de beber es 20 veces superior que si no se bebe.*

### 13.3.4. Obtención de *odds*, probabilidades (riesgos absolutos) y riesgo relativo

Para entender a fondo el modelo logístico, se puede partir del logaritmo neperiano de la *odds:*

$$\ln(odds) = a + bx$$

Si se asume que la exposición (alcohol en el ejemplo) vale 0 en no expuestos y 1 en expuestos, los logaritmos neperianos de las *odds* en cada situación serán:

Expuestos: $\ln(odds) = a + b*1 = a + b$

No expuestos: $\ln(odds) = a + b*0 = a$

¿Cómo podría saberse la probabilidad de accidente tras haber bebido alcohol? ¿Y si no se había bebido? Para calcular estas dos probabilidades, el primer paso es calcular sus *odds.* En este ejemplo ridículamente simple, para cada sujeto solo se considera una variable independiente *(alcohol)* en el modelo. Por lo tanto:

$$\ln(odds) = -2,303 + 2,996 * alcohol$$

Se sustituye el valor 0 (si se trata de no bebedor) o el valor 1 (si se trata de un bebedor) para la variable *alcohol,* y se hacen los cálculos.

No bebedor: $\ln(odds) = \ln(4/40) = -2,303 = a$

Bebedor: $\ln(odds) = \ln(24/12) = 0,69 = -2,303 + 2,996 = a + b$

Se toman antilogaritmos y se obtiene:

No bebedor: $odds = \exp(-2,303) = 0,10$

Bebedor: $odds = \exp(0,69) = 2$

La probabilidad *(p)* es igual a *odds*/(1 + *odds*):

No bebedor: $p = odds/(1 + odds) = 0{,}1/1{,}1 = 0{,}091 = 4/44$

Bebedor: $p = odds/(1 + odds) = 0{,}69/1{,}69 = 0{,}408 = 24/36$

La probabilidad predicha de accidente si se ha bebido será del 40,8% y, si no se ha bebido, del 9,1%. Esto puede y debe comprobarse en la tabla 13.3. Mientras no se entienda bien y a fondo este ejemplo numérico en todos estos detalles, es mejor no seguir adelante y tratar de hacer los cálculos a mano para comprobar cómo funciona la regresión logística.

Téngase en cuenta que solo cuando la variable dependiente o efecto es poco frecuente (<10%) puede asumirse la interpretación de la OR como una razón de probabilidades. Desgraciadamente, en el ejemplo, que es real, se constata que el efecto (accidente) estaba presente en el 9,1% de un grupo y el 40,8% del otro. Por tanto, al interpretar la OR = 20 no puede afirmarse que la *probabilidad* de tener un accidente tras beber es 20 veces superior que si no se hubiese bebido. No es la probabilidad lo que se multiplica por 20, sino la *odds*.

La *probabilidad* solo se multiplica por 4,5 (40,8%/9,1%).

## 13.4. REGRESIÓN LOGÍSTICA BINARIA CON VARIABLE INDEPENDIENTE CUANTITATIVA: INTERPRETACIÓN

A continuación se presenta otro ejemplo de regresión logística binaria[1]. En este caso, la variable predictora independiente es cuantitativa. Se ha valorado la relación entre edad *(age)* y cardiopatía isquémica *(coronary heart disease, chd)* (tabla 13.4 y fig. 13.6).

Puede llevarse a cabo un análisis de regresión logística a través de la instrucción:

```
logit chd age
```

El ordenador proporciona el siguiente resultado:

```
Logistic regression                                Number of obs   =        100
                                                   LR chi2(1)      =      29.31
                                                   Prob > chi2     =     0.0000
Log likelihood = -53.676546                        Pseudo R2       =     0.2145

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1109211   .0240598     4.61   0.000     .0637647    .1580776
       _cons |  -5.309453   1.133655    -4.68   0.000    -7.531376   -3.087531
------------------------------------------------------------------------------
```

*b* (coeficiente de regresión asociado a la variable *age*)

*a* (constante)

***Tabla 13.4 Base de datos para valorar la asociación entre la edad (age) y la cardiopatía isquémica (coronary heart disease, chd)***

| AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD | AGE | CHD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 0 | 30 | 0 | 34 | 0 | 37 | 0 | 41 | 0 | 44 | 1 | 48 | 1 | 53 | 1 | 57 | 0 | 60 | 0 |
| 23 | 0 | 30 | 0 | 34 | 0 | 37 | 1 | 42 | 0 | 44 | 1 | 48 | 1 | 53 | 1 | 57 | 1 | 60 | 1 |
| 24 | 0 | 30 | 0 | 34 | 1 | 37 | 0 | 42 | 0 | 45 | 0 | 49 | 0 | 54 | 1 | 57 | 1 | 61 | 1 |
| 25 | 0 | 30 | 0 | 34 | 0 | 38 | 0 | 42 | 0 | 45 | 1 | 49 | 0 | 55 | 0 | 57 | 1 | 62 | 1 |
| 25 | 1 | 30 | 0 | 34 | 0 | 38 | 0 | 42 | 1 | 46 | 0 | 49 | 1 | 55 | 1 | 57 | 1 | 62 | 1 |
| 26 | 0 | 30 | 1 | 35 | 0 | 39 | 0 | 43 | 0 | 46 | 1 | 50 | 0 | 55 | 1 | 58 | 0 | 63 | 1 |
| 26 | 0 | 32 | 0 | 35 | 0 | 39 | 1 | 43 | 0 | 47 | 0 | 50 | 1 | 56 | 1 | 58 | 1 | 64 | 0 |
| 28 | 0 | 32 | 0 | 36 | 0 | 40 | 0 | 43 | 1 | 47 | 0 | 51 | 0 | 56 | 1 | 58 | 1 | 64 | 1 |
| 28 | 0 | 33 | 0 | 36 | 1 | 40 | 1 | 44 | 0 | 47 | 1 | 52 | 0 | 56 | 1 | 59 | 1 | 65 | 1 |
| 29 | 0 | 33 | 0 | 36 | 0 | 41 | 0 | 44 | 0 | 48 | 0 | 52 | 1 | 57 | 0 | 59 | 1 | 69 | 1 |

1 Esta base de datos se puede descargar de nuestra página web: www.unav.es/preventiva.

Este resultado puede ser expresado como función logarítmica. A continuación se muestra la función de probabilidad *(p)* de cardiopatía isquémica que se obtendría a partir de la edad de los participantes:

$$\text{Valor pronosticado} = p = \frac{1}{1 + e^{-(-5{,}3+0{,}111 age)}}$$

Si ahora se va aplicando en dicha función valores para diferentes edades, se obtendrán los valores pronosticados que aparecen calculados o representados en la figura 13.7.

Como se aprecia en la figura, la función logística sigue una distribución con forma de «S». Esta forma de la curva indica que el efecto del factor sobre el riesgo de un suceso es mínimo cuando el

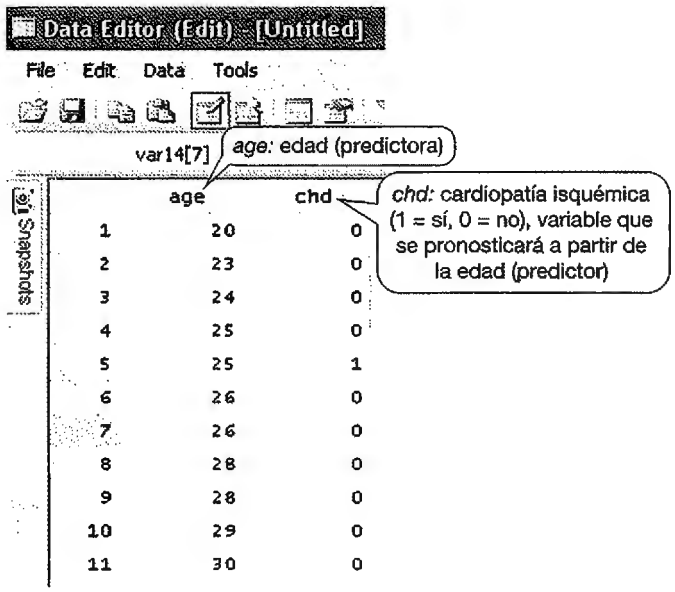



**Figura 13.6** Aspecto parcial de una base de datos en STATA.

| Edad | Valor predicho = 1/(1 + EXP(−(−5,3 + 0,111*age))) |
|---|---|
| 30 | 0,12 |
| 35 | 0,2 |
| 40 | 0,3 |
| 45 | 0,42 |
| 50 | 0,56 |
| 55 | 0,69 |
| 60 | 0,8 |

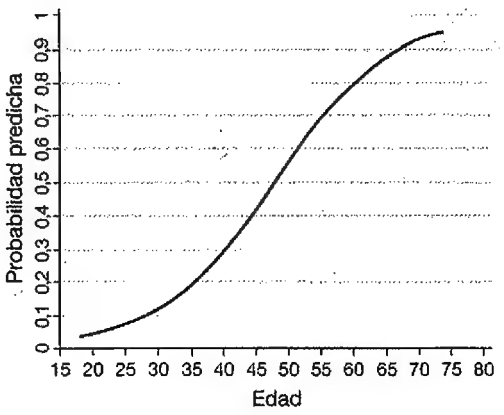



**Figura 13.7** Valores pronosticados o predichos calculados con la función logística en el ejemplo de la asociación entre la edad y la cardiopatía isquémica.

factor tiene un valor bajo hasta que se llega a un valor umbral a partir del cual el riesgo aumenta rápidamente a lo largo de varios valores intermedios hasta estabilizarse con valores pronosticados cercanos a la unidad en el caso de valores muy elevados del factor.

Las probabilidades pronosticadas se pueden obtener directamente de STATA. El programa crea una nueva variable en la base de datos de valor predicho o pronosticado a partir de la ecuación, que puede ser representada posteriormente en un gráfico de dispersión para representar la relación entre el factor y el riesgo (fig. 13.8), de forma similar a como se representó en la figura 13.7.

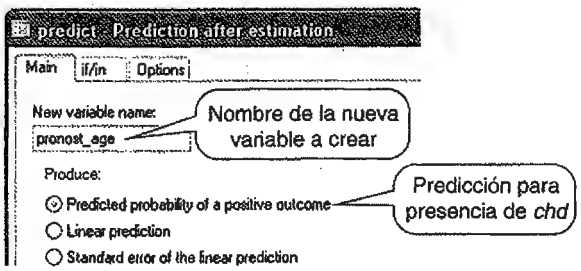
¿Cómo predice STATA la probabilidad de cardiopatía isquémica a partir de la edad?

Basta con darle la instrucción **predict** seguida por el nombre que se le quiere dar a esta nueva variable de la base de datos después de haber solicitado un **logit.** En el ejemplo anterior:

```
logit chd age
predict pronost_age
```

Esta instrucción puede obtenerse a través del menú:

**Statistics → Postestimation → Predictions, residuals, etc.**




La diferencia entre las figuras 13.7 y 13.8 es que, en la segunda, la gráfica ha sido realizada por STATA basándose exclusivamente en aquellos puntos para los que hay algún dato. En cambio, en la primera figura (realizada con Excel) se han extrapolado datos de predicción para cualquier edad, simplemente aplicando la ecuación a todas las posibles edades entre 18 y 75 años.

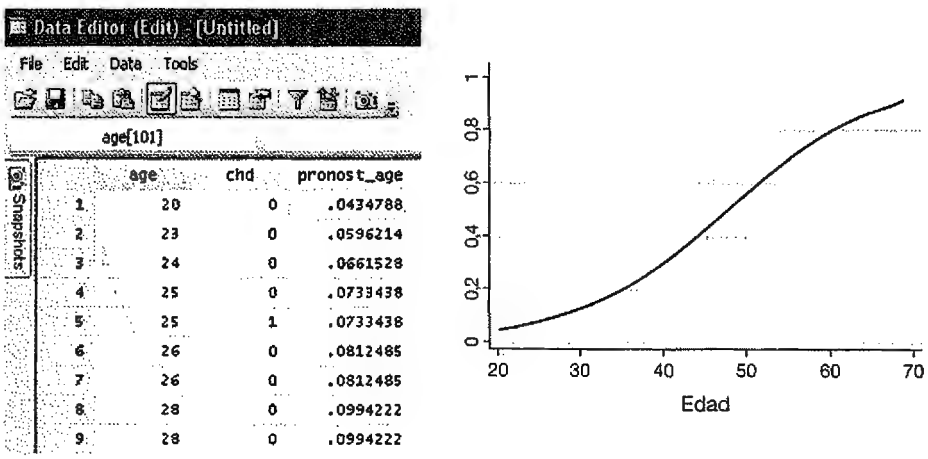



**Figura 13.8** Obtención de valores pronosticados y su gráfico de dispersión en STATA para el ejemplo de la asociación entre la edad y la cardiopatía isquémica.

Si se utilizara la instrucción **logistic,** STATA calcularía la OR de cardiopatía isquémica asociada al incremento en 1 año en la edad de los participantes.

```
. logistic chd age

Logistic regression                               Number of obs   =        100
                                                  LR chi2(1)      =      29.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.676546                       Pseudo R2       =     0.2145
```

| chd | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | 1.117307 | .0268822 | 4.61 | 0.000 | 1.065842 | 1.171257 |

***Interpretación:*** *la OR obtenida (1,117) se refiere al aumento de la odds por un solo año más de edad. Puede observarse cómo una OR de 1,12 corresponde al valor de* $e^{0,11}$.

*Para interpretar correctamente esta OR, es preferible muchas veces establecer un incremento de mayor magnitud en la variable independiente (edad). Por ejemplo, ¿cuánto valdría la OR al comparar una persona con 10 años más que otra? La OR valdría 3,03. Se calcula usando como exponente del número e el coeficiente b multiplicado por la diferencia de edad (10 años):*

$$OR(+10\,años) = EXP(0{,}111 \times 10) = e^{1{,}11} = 3{,}03$$

*Es decir, la instrucción* ***logistic*** *no podría ser utilizada para calcular el aumento de la odds de cardiopatía isquémica asociado a un incremento en 10 años en la edad de los participantes.*

## 13.5. REGRESIÓN LOGÍSTICA BINARIA CON UNA VARIABLE INDEPENDIENTE CON > 2 CATEGORÍAS: INTERPRETACIÓN

El modelo de regresión logística solo permite introducir como variables independientes variables de tipo cuantitativo o cualitativo *dicotómico.* Por tanto, cuando se desea introducir en un modelo de regresión logística una variable independiente cualitativa que tenga tres o más categorías, esta no puede introducirse en el modelo como tal.

La forma de actuar es transformar este variable cualitativa en otras distintas, todas cualitativas dicotómicas que la representen. Estas variables se denominan variables indicadoras o *dummy,* y presentan únicamente los valores 0 y 1. Sin embargo, surgen varias preguntas:

¿Cuántas variables *dummy* hay que crear?

Se crearán siempre tantas variables *dummy* como categorías tenga la variable original menos una. Por tanto, si la variable independiente tiene tres categorías (p. ej., *Hábito tabáquico*; 0 = no fumador, 1 = exfumador, 2 = fumador), se crearán dos *dummy.*

¿Cómo se crean?

Se elige siempre una categoría de referencia y se compara el resto con respecto a esta. En el ejemplo de la variable *hábito tabáquico,* puede elegirse como categoría de referencia al no fumador.

¿Cuál es la interpretación?

La creación de variables *dummy* permite la comparación por pares de las diferentes categorías de la variable independiente con respecto a la probabilidad *(odds)* de que se dé el fenómeno de estudio (variable dependiente). En el ejemplo del hábito tabáquico se realizarían dos comparaciones: exfumador/no fumador *(dummy1)* y fumador/no fumador *(dummy2).*

Véase con un ejemplo concreto. En una muestra de enfermeras (*Nurses' Health Study* II) se valoró si la obesidad en la edad infantil (cuantificada a través de la elección de una entre nueve siluetas corporales que definían a las participantes a los 5 años de edad) se relacionaba con la aparición de depresión en la edad adulta (9).

1. La incidencia de depresión en la edad adulta es la variable dependiente en este análisis, variable dependiente dicotómica (*depre;* 0 = no depresión; 1 = depresión), por lo que la prueba estadística de elección es la regresión logística.
2. La variable independiente imagen corporal en la infancia poseía nueve categorías. Dado el pequeño número de participantes que elegían las siluetas 8 y 9, estas dos categorías se solaparon en una sola. Finalmente, la imagen corporal a los 5 años *(silueta)* presentó ocho categorías (1 hasta 8-9). Así, fue necesario crear siete variables indicadoras.
3. Se eligió como categoría de referencia la silueta 3, considerada un índice de masa corporal (IMC) «normal» (*silueta* = 3). De esta forma, el resto de categorías se compararon con respecto a la categoría 3 (IMC normal).
4. Las nuevas variables creadas *(dummy)* recibieron el nombre *Imag1-Imag89,* y solo presentaban dos valores posibles (0 y 1). La *dummy Imag1* valía 1 cuando la participante presentaba la silueta 1 en la infancia y 0 en caso contrario. La *dummy Imag2* valía 1 cuando la participante presentaba la silueta 2 en la infancia y 0 en caso contrario, y así sucesivamente. El aspecto general de dichas variables en la base de datos se muestra en la tabla 13.5.

Puede observarse que no se ha creado la *dummy Imag3,* ya que dicha variable significaría comparar una participante consigo misma y no resultaría informativa.

Esta recodificación puede ser llevada a cabo con STATA gracias a la instrucción:

```
g Imag1=1 if silueta==1
.
g Imag2=1 if silueta==2
```

etc.

```
mvencode  Imag1  Imag2, mv(0)
```

(Esta instrucción permite transformar los valores faltantes en las variables *Imag1-Imag89* en valores 0.)

La forma de operar es idéntica a la explicada para variables independientes cualitativas dicotómicas o variables independientes cuantitativas. Las instrucciones que emplea STATA que permiten calcular OR o coeficientes de regresión ($b_j$) son:

***Tabla 13.5** Creación de variables* **dummy**

| SILUETA | IMAG1 | IMAG2 | IMAG4 | IMAG5 | IMAG6 | IMAG7 | IMAG89 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 8-9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

```
logistic depre imag1 imag2 imag4
         imag5 imag6 imag7 imag89

logit depre imag1 imag2 imag4 imag5
      imag6 imag7 imag89
```

(La instrucción **logistic** calcula OR, mientras que la instrucción **logit** calcula coeficientes de regresión.)

El aspecto del listado de salida de STATA tras aplicar estas instrucciones es:

| depre | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| imag1 | .9487778 | .1087381 | -0.46 | 0.646 | .7578958 | 1.187735 |
| imag2 | 1.063786 | .1174022 | 0.56 | 0.575 | .8568675 | 1.320673 |
| imag4 | 1.012836 | .1340943 | 0.10 | 0.923 | .781348 | 1.312907 |
| imag5 | 1.294748 | .1819386 | 1.84 | 0.066 | .9830466 | 1.705283 |
| imag6 | 1.393486 | .269363 | 1.72 | 0.086 | .954034 | 2.035359 |
| imag7 | 1.981571 | .5088521 | 2.66 | 0.008 | 1.19792 | 3.277869 |
| imag89 | 2.434501 | 1.054066 | 2.05 | 0.040 | 1.041994 | 5.687938 |

| depre | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| imag1 | -.0525806 | .1146086 | -0.46 | 0.646 | -.2772094 | .1720481 |
| imag2 | .0618346 | .1103625 | 0.56 | 0.575 | -.154472 | .2781412 |
| imag4 | .0127546 | .1323949 | 0.10 | 0.923 | -.2467346 | .2722438 |
| imag5 | .2583163 | .1405205 | 1.84 | 0.066 | -.0170988 | .5337314 |
| imag6 | .3318082 | .1933016 | 1.72 | 0.086 | -.0470559 | .7106724 |
| imag7 | .6838899 | .2567923 | 2.66 | 0.008 | .1805863 | 1.187193 |
| imag89 | .889742 | .4329701 | 2.05 | 0.040 | .0411361 | 1.738348 |
| _cons | -3.346478 | .0806892 | -41.47 | 0.000 | -3.504626 | -3.18833 |

Puede observarse cómo la imagen corporal 3 no ha sido introducida en el modelo, ya que se trata de la categoría de referencia y es comparada con respecto a sí misma; su OR valdría 1 y su coeficiente de regresión 0

***Interpretación:*** *aquellas participantes con sobrepeso-obesidad en la infancia (principalmente las figuras corporales 5 a 9) presentaron un incremento en el riesgo de sufrir depresión en la edad adulta comparadas con aquellas con peso normal en su niñez (silueta 3). Obsérvense las OR asociadas a dichas dummy: 1,29; 1,39; 1,98 y 2,43.*

*Este incremento fue particularmente importante entre las mujeres con una imagen 8 o 9 a los 5 años, en las que se observó una OR = 2,43 al compararlas con mujeres con IMC normal. Las mujeres que eligieron una imagen corporal a los 5 años muy obesa (silueta 8-9) presentaron 2,43 veces mayor odds de depresión que aquellas que eligieron una silueta normal. El valor de la OR podría hallarse igualmente a través de la segunda salida de STATA, aplicando antilogaritmos:* $e^{0,89}$.

En ocasiones, no resulta necesario crear las variables indicadoras a través de instrucciones dadas al programa estadístico. Tanto STATA como otros programas, como SPSS, permiten la creación directa de variables indicadoras o *dummy* sin necesidad de manipular las variables originales. En el programa STATA, la creación es directa siempre y cuando la categoría elegida como referencia sea la primera. Bastaría con incluir el término «**i.**» delante de la variable cualitativa que debe ser transformada. En el ejemplo presentado:

```
logistic depre i.silueta
```

En el caso del programa SPSS, el ordenador creará variables indicadoras directamente, siempre y cuando la categoría de referencia elegida sea la primera o la última.

## 13.6. REGRESIÓN LOGÍSTICA CON MÚLTIPLES VARIABLES INDEPENDIENTES

Lo expuesto hasta ahora sirve para introducir la regresión logística, aunque generalmente este modelo intenta explicar un fenómeno de salud (o enfermedad) (variable Y, dependiente) teniendo en consideración varias variables simultáneamente (variables X, independientes). La aplicación de un modelo de regresión logística responderá siempre a uno o varios de los siguientes objetivos de investigación:

1. Determinar los predictores de una determinada variable biosanitaria (variable Y) a partir de una lista más o menos amplia de posibles variables explicativas (variables X).
2. Construir un índice pronóstico (ecuación) para predecir una determinada condición (variable Y) a partir de los valores recogidos en otras variables (variables X).
3. Determinar el efecto de una variable $X_1$ sobre otra variable Y teniendo en cuenta otras características ($X_2$, $X_3$ ... $X_p$; factores de confusión) que pudieran distorsionar la verdadera asociación entre estas variables (**5**).
4. Detectar y describir fenómenos de interacción entre variables (modificación del efecto) sobre un determinado resultado. Es decir, si la presencia de una variable $X_2$ es capaz de modificar el efecto ejercido por la variable $X_1$ sobre la variable dependiente Y (**10**).

Por tanto, la función logística puede extenderse a la combinación de más de un factor predictor, X. Los diversos factores formarán una combinación lineal de variables:

$$\text{Valor pronosticado} = p = \frac{1}{1 + e^{-(a + b_1x_1 + b_2x_2 + \ldots b_px_p)}}$$

$$\ln\left(\frac{p(suceso)}{1 - p(suceso)}\right) = a + b_1x_1 + b_2x_2 + \ldots\ldots b_px_p$$

Si se define $f = a + b_1x_1 + b_2x_2 + \ldots b_px_p$, se demuestra que si $p = e^f/(1 + e^f)$ y $1 - p = 1/(1 + e^f)$, entonces, el valor de la *odds*($p/1 - p$):

$$\frac{p}{1-p} = \frac{e^f \Big/ \left(1 + e^f\right)}{1 \Big/ \left(1 + e^f\right)} = \frac{e^f\left(1 + e^f\right)}{\left(1 + e^f\right)} = e^f$$

$odds = e^f$ y, por tanto $\ln(odds) = f$.

Este es el fundamento de la regresión logística multivariable, que permite hacer predicciones más complejas (y más completas), ajustar por variables de confusión, valorar interacciones con términos de producto, etc.

En el ejemplo presentado para predecir la presencia de cardiopatía isquémica a partir de la edad de los participantes no solo debe tenerse en cuenta esta variable, sino otras, como el sexo o el índice de masa corporal. Los diversos factores (p. ej., edad, sexo e índice de masa corporal) formarían una combinación lineal de variables en el modelo de regresión logística:

$$\text{Valor pronosticado} = p = \frac{1}{1 + e^{-(a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots)}} = \frac{1}{1 + e^{-(a + b_1age + b_2sex + b_3bmi + \ldots)}}$$

## 13.7. CONTROL DE LA CONFUSIÓN EN REGRESIÓN LOGÍSTICA. LA CONFUSIÓN NO TIENE NADA QUE VER CON VALORES *p*

Se desea estudiar el efecto de la variable $X_1$ sobre la aparición de una determinada enfermedad Y. Sin embargo, se sospecha la presencia de un factor de confusión que podría sesgar la estimación que se calculará. Este factor de confusión o variable de control se denominará $X_2$. Para que una variable $X_2$

sea considerada factor de confusión, esta debe estar asociada independientemente tanto con la exposición ($X_1$) como con el desenlace (Y), y no ser un eslabón intermedio en la cadena causal (5).

La presencia de distorsión por una o más variables debe ser comprobada a partir de la creación de distintos modelos matemáticos en los que se van introduciendo las posibles variables de confusión paulatinamente. Un primer modelo de regresión incluirá únicamente la variable estudiada (variable independiente = $X_1$), pero no la variable de control (posible variable de confusión, $X_2$) (estimación cruda), y un segundo modelo en que se incluya tanto la variable predictora de interés ($X_1$) como el posible factor de confusión ($X_2$) (estimación ajustada).

Otra aproximación parte de modelos matemáticos complejos (con un número alto de variables independientes: variable(s) principal(es) de análisis y posibles variables de confusión), para ir eliminando variables del modelo hasta llegar al modelo más parsimonioso (que incluya el menor número relevante de variables).

Una vez obtenido el modelo matemático más complejo (con un mayor número de variables X incluidas), es posible establecer qué variables independientes pueden ser eliminadas del análisis observando el valor de significación estadística que llevan asociado o su intervalo de confianza. Si el valor *p* de significación es inferior a 0,05, la variable debe permanecer en el modelo, al tratarse de una variable predictora del suceso estudiado (se comporte o no como variable de confusión). Si el valor de significación estadística es superior a 0,25, la variable se eliminará del modelo, ya que no es una variable relevante (no se asocia con el desenlace o variable dependiente y no es, por tanto, un factor de confusión). Si su valor de significación estadística se sitúa entre 0,05 y 0,25, deberá observarse el cambio que se produce en la OR de la variable principal de análisis tras la inclusión/exclusión de la hipotética variable de confusión. Si el cambio es grande (en torno al 15-20%), la variable produce distorsión y debe permanecer en el modelo, es un factor de confusión que, de no tenerse en cuenta, sesgaría el resultado principal (5). Si el cambio de magnitud no es relevante, se optaría por el modelo más parsimonioso (esto es, por aquel con un menor número de variables independientes; la variable se eliminaría del análisis).

Puede observarse que decidir si un determinada variable independiente $X_2$ es o no un factor de confusión no depende de ninguna prueba estadística (u obtención de un determinado valor *p* de probabilidad), sino de un cambio observado en la magnitud del efecto (cambio de magnitud en el valor de la OR para la variable principal de análisis $X_1$) cuando se considera y no se considera esta segunda variable $X_2$ en el modelo matemático.

Suponga una investigación que intenta explicar los factores que intervienen en la probabilidad de conducir después de haber bebido. Puede considerarse que el sexo es una de estas variables, estableciendo que los hombres presentan una mayor probabilidad que las mujeres. Imagine que intervienen también otras variables en esta probabilidad, como la edad de los sujetos, el número de kilómetros conducidos o el estado civil.

Así, se ajustaría un modelo de regresión logística para predecir la probabilidad *(p)* de tener el hábito de conducir después de haber bebido alcohol, considerando como variables independientes el sexo (0 = mujer; 1 = varón), la edad (en siete grupos, comenzando por los menores de 25 años, hasta los que tienen 50 o más años), los kilómetros conducidos al año (en cinco categorías, la inferior para los de <1.000 y la superior para >50.000 km/año) y el estado civil (cuatro categorías: soltero, casado, viudo y otros). En el listado (tabla 13.6), para cada variable aparece una categoría menos que las originales. La categoría que no aparece es la de referencia, frente a la cual se comparan el resto, usando variables *dummy*.

***Interpretación:*** *la odds de conducir tras beber es 1,303 veces mayor en hombres que en mujeres, una vez ajustadas la comparación por los otros tres factores que aparecen en la tabla, es decir, a igualdad de edad, kilómetros conducidos y estado civil. La OR sería 1,303 para los varones y se podría escribir que es 1,00 (referencia) para las mujeres.*

*Tabla 13.6 Análisis de regresión logística de factores asociados con el hábito de beber antes de conducir (n = 16.393)*

| | COEFICIENTE DE REGRESIÓN | ERROR ESTÁNDAR | OR | *p* |
|---|---|---|---|---|
| Sexo (1 = varón) | 0,264 | 0,038 | 1,303 | <0,001 |
| **Grupos de edad** | | | | <0,001 |
| 25-29,9 | 0,742 | 0,077 | 2,099 | <0,001 |
| 30-34,9 | 0,925 | 0,081 | 2,521 | <0,001 |
| 35-39,9 | 1,025 | 0,086 | 2,786 | <0,001 |
| 40-44,9 | 0,954 | 0,090 | 2,596 | <0,001 |
| 45-49,9 | 0,761 | 0,093 | 2,140 | <0,001 |
| ≥50 | 0,588 | 0,088 | 1,800 | <0,001 |
| **Kilómetros conducidos** | | | | <0,001 |
| 1.000-10.000 | 0,502 | 0,074 | 1,652 | <0,001 |
| 10.001-20.000 | 0,736 | 0,075 | 2,088 | <0,001 |
| 20.001-50.000 | 0,753 | 0,077 | 2,122 | <0,001 |
| >50.000 | 0,700 | 0,116 | 2,014 | <0,001 |
| **Estado civil** | | | | 0,553 |
| Casado | –0,012 | 0,046 | 0,988 | 0,798 |
| Viudo | –0,180 | 0,195 | 0,835 | 0,355 |
| Separado/otros | 0,096 | 0,103 | 1,101 | 0,350 |
| **Constante** | –2,346 | 0,090 | 0,096 | <0,001 |

La máxima frecuencia de conducir tras beber se da en personas de 35 a 40 años (OR = 2,786) y la mínima en quienes tienen menos de 25 años (OR = 1,00, categoría de referencia). Todo esto con igualdad de sexo, kilómetros conducidos y estado civil.

Los que conducen de 20.000 a 50.000 km/año son quienes con más frecuencia se exponen a beber antes de ponerse al volante (OR = 2,122), comparados con los que conducen <1.000 km/año, que son la referencia (OR = 1,00).

No existen diferencias estadísticamente significativas en el hábito de beber antes de conducir según estado civil, ya que el test de significación aplicado (test de Wald; v. apartado 13.11) no resulta significativo para ninguna de las tres variables *dummy* que se han usado para valorar el estado civil ($p$ = 0,798; $p$ = 0,355 y $p$ = 0,350) frente a la categoría de referencia (solteros).

Puede comprobarse que, para las variables con más de dos categorías (grupos de edad, kilómetros conducidos y estado civil), el análisis de regresión logística proporciona primero un valor $p$ global ($p$ = 0,553 para estado civil y $p < 0,001$ para las otras dos), que indica si la predicción del suceso de interés (beber y conducir en el ejemplo) mejora significativamente al añadir esta variable en su conjunto.

A la luz del resultado para el estado civil, puede afirmarse que no es una variable independientemente asociada con el hecho de conducir tras haber bebido y, por tanto, tampoco actuaría como variable de confusión. Habría que suprimirla del modelo. Recuérdese que siempre se busca el modelo con menor número de variables (más parsimonioso).

Una vez eliminada del modelo la variable estado civil, este se vuelve ajustar y quedaría tal como se recoge en la tabla 13.7, a la que se han añadido los intervalos de confianza (v. apartado 13.11) y las categorías de referencia para asimilarla al modo en que se suelen presentar unos resultados de regresión logística en una publicación científica.

***Interpretación:*** *las OR para la edad, el sexo y el número de kilómetros prácticamente no se han modificado al dejar de ajustar por estado civil, lo que indica que esta variable no inducía confusión, como ya se intuía. Obsérvese que todas las variables del modelo se asocian de manera independiente al riesgo de conducir después de haber bebido. Con respecto a las mujeres, e independientemente de la edad y del número de kilómetros conducidos anualmente, los hombres presentan 1,3 veces mayor odds*

*Tabla 13.7 Modo de presentar los resultados de regresión logística en una publicación. Factores independientemente asociados con el hábito de beber antes de conducir (n = 16.393)*

| | OR (IC 95%) | p |
|---|---|---|
| **Sexo** | | |
| Mujer | 1 (ref.) | |
| Hombre | 1,30 (1,21-1,40) | <0,001 |
| **Grupos de edad** | | |
| <25 | 1 (ref.) | |
| 25-29,9 | 2,10 (1,80-2,44) | <0,001 |
| 30-34,9 | 2,51 (2,15-2,93) | <0,001 |
| 35-39,9 | 2,77 (2,37-3,24) | <0,001 |
| 40-44,9 | 2,58 (2,20-3,04) | <0,001 |
| 45-49,9 | 2,13 (1,80-2,51) | <0,001 |
| > = 50 | 1,78 (1,53-2,08) | <0,001 |
| **Kilómetros conducidos** | | |
| <1.000 | 1 (ref.) | |
| 1.000-10.000 | 1,65 (1,43-1,91) | <0,001 |
| 10.001-20.000 | 2,09 (1,81-2,42) | <0,001 |
| 20.001-50.000 | 2,13 (1,83-2,47) | <0,001 |
| >50.000 | 2,02 (1,61-2,54) | <0,001 |

Es conveniente añadir también el tamaño (n) de cada categoría.

*de conducir tras haber ingerido alcohol. También presentan una mayor odds los sujetos con más de 25 años y que conducen más de 1.000 km anuales.*

## 13.8. IDENTIFICACIÓN DE LA INTERACCIÓN EN REGRESIÓN LOGÍSTICA: TEST DE RAZÓN DE VEROSIMILITUD

Se denominan variables modificadoras del efecto aquellas que modifican la intensidad o el sentido de la relación entre el factor de estudio (variable independiente $X_1$) y el desenlace (Y) (**10**). La forma de operar a la hora de identificar posibles interacciones (variables modificadoras del efecto) en un análisis de regresión logística consiste en crear e introducir en el modelo matemático términos producto entre la variable principal de análisis $X_1$ y cada hipotética variable modificadora del efecto ($X_2$, $X_3$, etc.). Estas variables, hipotéticos modificadores del efecto, son seleccionadas a partir de las variables de confusión identificadas en análisis anteriores.

Se creará, por tanto, un modelo final, que debe ser jerárquico. El modelo jerárquico se define como un modelo tal que, si se elimina un término cualquiera, todos los términos de mayor orden en los que intervenga también deben ser eliminados. Inversamente, si se incluye un término cualquiera, todos sus términos de menor orden también deberán estar presentes en el modelo. Esto implica que si, por ejemplo, un modelo contiene la interacción $X_1*X_2*X_3$, también deberá contener las interacciones $X_1*X_2$ y $X_1*X_3$, los términos control $X_2*X_3$, $X_2$, $X_3$ y la variable de exposición $X_1$.

Se recomienda no crear interacciones demasiado complejas del tipo $X_1*X_2*X_3$, porque presentan dos problemas:

1. Son de difícil interpretación clínica.
2. Suelen dar problemas de colinealidad.

El modelo final, por tanto, tendrá un aspecto similar al siguiente:

$$\text{logit}(p) = a + b_1x_1 + b_2x_2 + b_3x_1 * x_2$$

Para comprobar si existe interacción entre la variable $X_1$ y $X_2$, el valor de significación estadística asociado al coeficiente de regresión de la variable producto ($b_3$ en el ejemplo) debe ser inferior a 0,05 según el test de razón de verosimilitud (v. apartados 13.10 y 13.11).

Si se sospecha la presencia de más de una interacción, en primer lugar deben analizarse todas ellas de forma conjunta en un modelo matemático máximo inicial.

$$\text{logit}(p) = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1 * x_2 + b_5x_1 * x_3$$

A través del test de razón de verosimilitud, se evalúa la significación del conjunto de interacciones del modelo máximo. Si el resultado del test fuera estadísticamente significativo, se aplicarían nuevas pruebas de significación estadística a cada interacción candidata de ser eliminada.

Para evaluar una interacción mediante el test de la razón de verosimilitud en STATA, se seguiría el siguiente procedimiento:

```
generate inter =x1 *x2
logistic y x1 x2
estimates store A   #(para guardar los resultados
del modelo sin interacción)
logistic  y x1 x2 inter
lrtest A.
```

siendo $y$ la variable dependiente y $x_1$ y $x_2$ las variables independientes (cuantitativas o cualitativas dicotómicas).

(STATA llama «.» a todos los resultados del último modelo con interacción.)

Un resultado de ejemplo podría ser el siguiente:

```
Likelihood -ratio test                    LR chi2(1)=8.70

(Assumption: A nested in.)          Prob > chi2 =0.0032
```

En este caso, la interacción sería estadísticamente significativa ($p$ = 0,0032).

## 13.9. SUPUESTO DE LINEALIDAD EN EL LOGIT Y USO DE TÉRMINOS POLINÓMICOS

### 13.9.1. Linealidad en el logit

Para introducir una variable cuantitativa como independiente, debe comprobarse su linealidad en el logit. Sucede así porque el modelo de regresión logística establece que la función se hace lineal en el logit a partir de la ecuación:

$$\text{logit}(p) = \ln(odds) = a + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

La comprobación de que la variable independiente cuantitativa se vuelve lineal en el logit puede hacerse categorizando esta variable en grupos excluyentes (por cuantiles o por puntos de corte

establecidos *a priori*) y calculando, a continuación, el logit (logit = ln[*odds*]) en cada categoría (o el logaritmo de la OR con respecto a una categoría de referencia). A continuación se representa de forma gráfica cada logit (eje *y*) frente al valor (eje *x*) de la mediana de cada cuantil (o del valor medio del intervalo si se utilizaron otros puntos de corte).

Si el tamaño de la muestra es muy grande, una alternativa muy utilizada por los investigadores para el tratamiento de las variables cuantitativas es agruparlas en categorías. Sin embargo, si se comprueba que la variable categorizada es lineal en el logit, puede (y debería) introducirse la variable de forma cuantitativa, en vez de ordinal. De hecho, esta solución es mejor, debido a que la variable categorizada tiene tantos grados de libertad como categorías −1, pero cuando se introduce de forma cuantitativa solo tiene un grado de libertad.

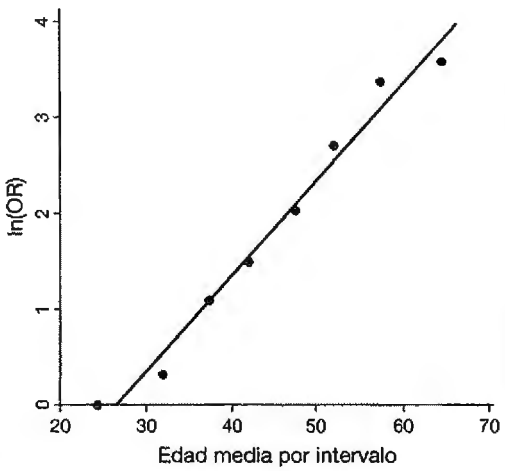
Como ejemplo, supóngase que se desea determinar el efecto de la edad sobre el infarto de miocardio. Se prefiere analizar la edad como variable cualitativa ordinal. Se decide categorizar esta variable en ocho grupos por puntos de corte establecidos *a priori (int_edad)* (tabla 13.8).

Se calcula el ln de la OR obtenida al comparar cada categoría con la de referencia, que será la correspondiente a las edades más jóvenes (20-29 años) (fig. 13.9).

***Interpretación:*** *se observa linealidad del ln(OR), por lo que puede introducirse la variable edad como variable cuantitativa continua.*

***Tabla 13.8 Categorización de la variable edad en ocho grupos por puntos de corte establecidos a priori (int_edad)***

| INTERVALOS | MEDIANA | DUM1 | DUM2 | DUM3 | DUM4 | DUM5 | DUM6 | DUM7 |
|---|---|---|---|---|---|---|---|---|
| 20-29 | 24,5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30-34 | 32 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 35-39 | 37,5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 40-44 | 42 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 45-49 | 47,5 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 50-54 | 52 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 55-59 | 57,5 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 60-69 | 64,5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |




| Intervalos de edad | Mediana del intervalo | ln (OR) |
|---|---|---|
| 20-29 | 24,5 | 0 |
| 30-34 | 32 | 0,325 |
| 35-39 | 37,5 | 1,099 |
| 40-44 | 42 | 1,504 |
| 45-49 | 47,5 | 2,043 |
| 50-54 | 52 | 2,708 |
| 55-59 | 57,5 | 3,376 |
| 60-69 | 64,5 | 3,584 |

**Figura 13.9** Representación gráfica del ln de la *odds ratio* de cardiopatía según intervalos de edad. Linealidad en el logit.

### 13.9.2. Comprobación de la linealidad en el logit con STATA. Test de tendencia lineal

La comprobación de la linealidad puede llevarse a cabo a través de la aplicación de un test de tendencia lineal sobre la variable de exposición categorizada (bien en grupos *a priori* o bien en cuantiles). Para ello, a todos los sujetos de cada categoría se les asignará el valor de la mediana específica de dicha categoría, intervalo o cuantil, y se tratará esta nueva variable como continua. El valor *p* para esta variable es el *p* de tendencia lineal, que contrasta la hipótesis nula de que no hay variación (ni aumento ni descenso) progresiva del riesgo a medida que se pasa de uno a otro intervalo, categoría o cuantil.

En el ejemplo de la asociación entre la edad y la cardiopatía isquémica, para obtener un test de tendencia lineal deben llevarse a cabo tres pasos que puede ejecutarse a través del uso de instrucciones:

1. Recodificar la variable *age* en la variable cualitativa ordinal *int_edad,* con ocho categorías.

```
recode age (20/29=0 "20-29 años") (30/34=1 "30-34 años")
(35/39=2 "35-39 años") (40/44=3 "40-44 años") (45/49=4
"45-49 años") (50/54=5 "50-54 años") (55/59=6 "55-59
años") (60/69=7 "60-69 años"), g (int_edad)
```

2. La variable categorizada *(int_edad)* será transformada en una variable cuantitativa (*edadlin* con 1 grado de libertad). A los sujetos de cada categoría se les asignará el valor de la mediana específica de dicha categoría[2].

```
g edadlin =24.5

replace edadlin =32 if int_edad==1

. replace edadlin =37.5 if int_edad==2

replace edadlin =42 if int_edad==3

replace edadlin =47.5 if int_edad==4

replace edadlin =52 if int_edad==5

replace edadlin =57.5 if int_edad==6

replace edadlin =64.5 if int_edad==7
```

Esto podría hacerse con un bucle, pero se ha especificado así para que se vea más claro el proceso. Si el coeficiente de esta nueva variable *(edadlin)* resulta significativo ($p < 0{,}05$), se rechaza la hipótesis nula y se dispondrá de evidencias para apoyar la hipótesis de un crecimiento lineal progresivo del riesgo a medida que aumenta la edad (se ha observado una OR superior a 1 en el quintil 5).

---

2 Podría hacerse también a través de las instrucciones:

```
egen int_edad= cut (age), at (0 30 35 40 45 50 55 60 200)
gen edadlin=24.5
replace edadlin =32 if int_edad==0
replace edadlin =37.5 if int_edad==30
replace edadlin =42 if int_edad==35
replace edadlin =47.5 if int_edad==40
replace edadlin =52 if int_edad==45
replace edadlin =57.5 if int_edad==50
replace edadlin =64.5 if int_edad==60
```

```
Logistic regression                                   Number of obs   =        100
                                                      LR chi2(1)      =      28.24
                                                      Prob > chi2     =     0.0000
Log likelihood = -54.212098                           Pseudo R2       =     0.2066

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     edadlin |   .1045586   .022946     4.56   0.000     .0595853    .1495319
       _cons |  -5.042559   1.086614    -4.64   0.000    -7.172284   -2.912834
------------------------------------------------------------------------------
```

El coeficiente de regresión asociado a la variable *edadlin* es significativo (asume linealidad)

### 13.9.3. Uso de términos polinómicos

A veces, la relación en el logit no es lineal (forma de recta), sino que adopta otras distribuciones. Se podrían aplicar los mismos principios de transformación de variables independientes vistos en el capítulo anterior para la regresión lineal. En estos casos, para la regresión logística, lo más habitual no es usar polinomios, sino que la variable cuantitativa independiente se transforma en una variable cualitativa ordinal para introducirse de dicha forma en el modelo de regresión logística (como varias variables *dummy*, habitualmente tres o cuatro). Otra forma de proceder se basaría en crear nuevas variables, como $X^2$, $X^3$, $1/X$ (según proceda) a través de transformaciones matemáticas sencillas para introducirlas así en el modelo.

Imagínese que quiere valorarse el efecto de la edad sobre el riesgo de obesidad. Se establece que dicha relación es probablemente cuadrática, con un mayor riesgo de presentar obesidad en edades intermedias de la vida, estando más protegidos los sujetos más jóvenes y los ancianos. El modelo final sería:

$$\text{logit}(p) = a + b_1 edad + b_2 edad^2$$

## 13.10. AJUSTE DE UN MODELO MULTIVARIABLE DE REGRESIÓN LOGÍSTICA

En el modelo de regresión lineal, los parámetros se estimaban a través del método de mínimos cuadrados. En la regresión logística, los parámetros son estimados mediante el método de máxima verosimilitud *(maximum likelihood)*. Para explicar este método, es importante introducir el concepto de la función de verosimilitud (*likelihood*, L).

Esta función representa la probabilidad de observar los datos que se han recogido. Puede entenderse con un ejemplo sencillo. Al lanzar una moneda al aire, existe una probabilidad poblacional *desconocida* ($\pi$) de obtener cara y una probabilidad ($1 - \pi$) de que resulte cruz. Esta probabilidad es desconocida, ya que nadie ha lanzado infinitas monedas al aire. El parámetro ($\pi$) es desconocido y solo se puede apostar por él. Imagine que lanza cuatro veces la moneda al aire y obtiene dos caras y dos cruces. La probabilidad de que salgan dos caras al lanzar cuatro veces la moneda al aire puede ser descrita con la distribución binomial. La expresión de la distribución binomial para este ejemplo es la siguiente:

$$\pi^{n.^{o}\ caras}(1-\pi)^{n.^{o}\ cruces} = \pi^2(1-\pi)^2$$

Esta expresión corresponde a la función de verosimilitud de este ejemplo.

La probabilidad de sacar cara al lanzar una vez la moneda puede oscilar entre 0 y 1, tal y como se recoge en la tabla 13.9. Como realmente se desconoce el valor del parámetro ($\pi$), se realizan diferentes estimaciones. Se trata de buscar aquel valor de $\pi$ que maximice la verosimilitud de los datos realmente observados (dos caras y dos cruces). En la tabla 13.9 se van sustituyendo diversos valores por $\pi$ en la ecuación:

$$L = \pi^2(1-\pi)^2$$

***Tabla 13.9 Verosimilitud de los datos según diversos valores de probabilidad poblacional ($\pi$)***

| $\pi$ (PARÁMETRO DESCONOCIDO) | VEROSIMILITUD (L) |
|---|---|
| 0,1 | 0,0081 |
| 0,2 | 0,0256 |
| 0,3 | 0,0441 |
| 0,4 | 0,0576 |
| 0,5 | 0,0625 |
| 0,6 | 0,0576 |
| 0,7 | 0,0441 |
| 0,8 | 0,0256 |
| 0,9 | 0,0081 |
| 1 | 0 |

L = *likelihood* (verosimilitud)

Se empieza apostando por $\pi = 0{,}1$, luego $\pi = 0{,}2$, etc., hasta que se alcanza «convergencia». Se alcanza «convergencia» cuando se obtiene un valor del parámetro ($\pi = 0{,}5$, en el ejemplo) que maximiza la verosimilitud (fig. 13.10).

El método de máxima verosimilitud probará posibles valores de los parámetros hasta encontrar el que maximice la función de verosimilitud (la distribución binomial en el ejemplo). La función logística presenta similitudes con el ejemplo del lanzamiento de la moneda. En la función logística, la probabilidad $\pi$ de que se produzca el suceso viene dada por la expresión:

$$\pi = \frac{e^{a+bx}}{1+e^{a+bx}}$$

La probabilidad de que el suceso *no* se produzca será:

$$1-\pi = \frac{1}{1+e^{a+bx}}$$

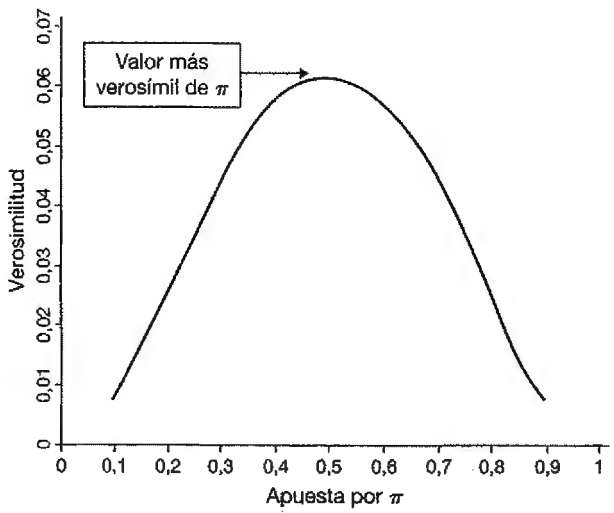



**Figura 13.10** Función de máxima verosimilitud.

En el escenario más sencillo, si se denomina Y a la variable dependiente (solo puede valer 1 o 0) y se llama X a una variable independiente que también sea dicotómica, se tendrá el resultado expresado en la tabla 13.10.

Si se sustituyen estos valores en la expresión de la distribución binomial:

$$\pi^{n.^{\circ}\ eventos}(1-\pi)^{n.^{\circ}\ no\ eventos}$$

$$Likelihood = \prod\left(\frac{e^{a+bx}}{1+e^{a+bx}}\right)^{y} \times \left(\frac{1}{1+e^{a+bx}}\right)^{1-y}$$

(Cociente 1.° *($\pi$):* solo entra en el producto para los eventos. Cociente 2.° ($1 - \pi$): solo entra en el producto en sujetos sin evento.)

Y = 1 para los eventos.

Y = 0 para los no eventos.

El multiplicatorio con el que comienza la ecuación supone que esta expresión de verosimilitud *(likelihood)* se irá multiplicando para todos los sujetos de la muestra. Para cada sujeto de la muestra se aplica una probabilidad de convertirse en caso ($\pi$, primer cociente) o en no caso ($1 - \pi$, segundo cociente). Obsérvese que el primer cociente se eleva a *y*, por lo tanto se elevará a 1 si el sujeto es caso y a 0 si no lo es. Elevar a 0 equivale a valer 1 y supone que no se aporta nada a la multiplicación. El segundo cociente ($1 - \pi$) se eleva a $1 - y$; por lo tanto, se elevará a 0 (no entrará) en los casos y a 1 (entrará) en los controles.

Imagine que el primer sujeto de una base de datos es un caso, el segundo un control y el tercero un caso. La función de verosimilitud empezaría así:

$$L = \left[\left(\frac{e^{(a+bx)}}{1+e^{(a+bx)}}\right)^{1}\left(\frac{1}{1+e^{(a+bx)}}\right)^{0}\right] \times \left[\left(\frac{e^{(a+bx)}}{1+e^{(a+bx)}}\right)^{0}\left(\frac{1}{1+e^{(a+bx)}}\right)^{1}\right] \times \left[\left(\frac{e^{(a+bx)}}{1+e^{(a+bx)}}\right)^{1}\left(\frac{1}{1+e^{(a+bx)}}\right)^{0}\right] \times \ldots$$

$$L = \left(\frac{e^{(a+bx)}}{1+e^{(a+bx)}}\right) \times \left(\frac{1}{1+e^{(a+bx)}}\right) \times \left(\frac{e^{(a+bx)}}{1+e^{(a+bx)}}\right) \times \ldots = \pi(1-\pi)\pi\ldots$$

Para el primer sujeto, que es un caso, solo entra el primer cociente, pues el segundo cociente en ese primer sujeto estaría elevado a 0 ($1 - 1 = 0$). Para el segundo sujeto solo entraría el segundo cociente, ya que, al ser un control, el primer cociente está elevado a 0 (Y = 0), etc. Puede apreciarse que al final ($\pi(1 - \pi)\ \pi\ldots$) hay cierto parecido con el ejemplo anterior de lanzar la moneda dos veces.

***Tabla 13.10 Fórmulas para el cálculo de probabilidades de acuerdo con una variable de exposición (independiente) X y una variable de desenlace (dependiente) Y***

| VARIABLE INDEPENDIENTE O EXPOSICIÓN (X) | VARIABLE DEPENDIENTE (Y) | |
|---|---|---|
| | Y = 1 (EVENTOS) | Y = 0 (NO EVENTOS) |
| X = 1 (expuestos) | $\pi = \frac{e^{a+b}}{1+e^{a+b}}$ | $1-\pi = \frac{1}{1+e^{a+b}}$ |
| X = 0 (no expuestos) | $\pi = \frac{e^{a}}{1+e^{a}}$ | $1-\pi = \frac{1}{1+e^{a}}$ |

De manera intuitiva, puede entenderse que, a partir de esta función, se obtienen los coeficientes. Sucede así porque, una vez fijada la función de verosimilitud (*likelihood,* L), el ordenador va probando reiteradamente (como si hiciese ensayos y errores repetidos) con distintos valores para los parámetros desconocidos *a* y *b* hasta que, después de varias iteraciones, alcanza convergencia. En aquellos valores en que se alcance convergencia, el ordenador decidirá que están los coeficientes *a* y *b* del modelo de regresión logística. Como puede comprenderse, este proceso es mucho más complicado que el de una regresión lineal simple. Habitualmente no se maximiza la función de verosimilitud, sino su logaritmo, debido a que resulta más sencillo manejar los logaritmos, pues las cifras absolutas de verosimilitud suelen ser ínfimas.

La comparación de dos cantidades correspondientes a –2 ln (verosimilitud) para dos modelos sigue una distribución ji cuadrado ($\chi^2$) y permite comparar estadísticamente los modelos. La $\chi^2$resultante tendrá tantos grados de libertad como la diferencia en el número de parámetros entre los dos modelos que se comparan. A este test se le ha dado el nombre de test de la razón de verosimilitud (*–2 loglikelihood* o –2 log verosimilitud).

## 13.11. SIGNIFICACIÓN ESTADÍSTICA EN LA REGRESIÓN LOGÍSTICA

Los coeficientes del modelo de regresión y sus correspondientes *odds ratios* (OR) son estimadores obtenidos a través de una muestra procedente de una población con parámetros desconocidos. Por ello, llevan asociado cierto grado de variabilidad expresada a través de sus errores estándar. Esto hace necesario emplear técnicas de inferencia para estimar los parámetros poblacionales. Para comprobar la significación estadística de estas estimaciones realizadas según un modelo de regresión logística, pueden emplearse test de hipótesis o intervalos de confianza de los parámetros poblacionales.

Con respecto a los primeros, se utilizan dos test de hipótesis, que se indican a continuación.

### 13.11.1. Prueba de la razón de verosimilitud

Imagine tres modelos de regresión logística. El primero está formado por tres variables independientes (Logit$(p) = a + b_1x_1 + b_2x_2 + b_3x_3$) y presenta una función de máxima verosimilitud (L3). El segundo modelo, formado por dos de estas variables, presenta una función L2 y, finalmente, un modelo formado por una sola de estas tres variables independientes tiene una función de máxima verosimilitud L1.

Como cuantos más parámetros tenga un modelo mejor se ajustan los datos, la función de máxima verosimilitud del modelo 3 debe ser mayor o igual que la del modelo 2, y esta, a su vez, será mayor o igual a la del modelo 1: $\mathbf{L3 \geq L2 \geq L1.}$

Si la función de máxima verosimilitud es expresada a través de logaritmos neperianos, entonces: $\mathbf{lnL3 \geq lnL2 \geq lnL1.}$

Sin embargo, los programas estadísticos no emplean el valor del logaritmo neperiano, sino el valor –2ln[L], es decir –2 por el logaritmo neperiano de la verosimilitud. El signo menos hace que se invierta el sentido de las diferencias: $\mathbf{-2lnL3 \leq -2lnL2 \leq -2lnL1.}$

Los cambios de significación estadística que se producen al eliminar una o más variables de un modelo de referencia o previo con el que se compara el nuevo se estiman con el test de razón de verosimilitud. La hipótesis nula se define estableciendo un valor 0 para aquellos parámetros poblacionales que no están presentes en el modelo más reducido.

En un ejemplo:

*Modelo 3:* $Logit(p) = a + b_1x_1 + b_2x_2 + b_3x_3$.

*Modelo 1:* $Logit(p) = a + b_1x_1$.

$H_0 \equiv b_2 = b_3 = 0$.

$$\chi^2 \approx -2\ln\left(\frac{\text{L Modelo 1}(a+b_1x_1)}{\text{L Modelo 3}(a+b_1x_1+b_2x_2+b_3x_3)}\right)$$

con 2 grados de libertad.

Para evaluar con STATA diferencias estadísticas entre distintos modelos mediante el test de verosimilitud, se realizaría el mismo procedimiento que para evaluar la interacción en la regresión logística explicado en el apartado 13.8.

### 13.11.2. Test de Wald

Se emplea normalmente cuando solo se está probando un único parámetro (eliminación de una única variable del modelo largo).

Por ejemplo:

*Modelo 3:* $Logit(p) = a + b_1x_1 + b_2x_2 + b_3x_3$

*Modelo 2:* $Logit(p) = a + b_1x_1 + b_2x_2$

$H_0 \equiv b_3 = 0.$

Se calcula dividiendo el parámetro por su error estándar:

$$\text{Wald}^2 = \left(\frac{b_3}{\text{Error estándar}(b_3)}\right)^2$$

Para muestras grandes, el test de Wald sigue una distribución $z$, es decir, se puede contrastar con la distribución normal tipificada. De la misma forma, $z^2$ corresponde aproximadamente a una distribución ji cuadrado con un grado de libertad.

La significación estadística (valor $p$) de cada variable que aparece en el modelo es la obtenida a través del test de Wald.

Ambas pruebas suelen llegar aproximadamente al mismo resultado con muestras grandes, mientras que con muestras pequeñas pueden existir diferencias. Cuando existan diferencias, es preferible usar el test de la razón de verosimilitud.

### 13.11.3. Intervalo de confianza de la *odds ratio*

El cálculo del intervalo de confianza al 95% para la OR se realizará como se muestra a continuación:

$$IC95\%(OR) = e^{b\pm1{,}96(EE)}$$

STATA permite obtener directamente el intervalo de confianza para cada OR a través de la instrucción:

```
logistic
```

En el ejemplo de la cardiopatía isquémica y la edad:

```
logistic chd age
```

En este caso, el intervalo de confianza asociado a la edad oscilará entre 1,07 y 1,17 (con una confianza del 95%)

| chd | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | 1.117307 | .0268822 | 4.61 | 0.000 | 1.065842 | 1.171257 |

Desde la opción:

```
logit
```

```
logi chd age
```

El coeficiente de regresión vale 0,11 y el error estándar 0,024, así: IC95%($OR_{age}$) = $e^{0,11 \pm 1,96 \times 0,024}$ = (1,07 – 1,17)

| chd | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| age | .1109211 | .0240598 | 4.61 | 0.000 | .0637647 | .1580776 |
| _cons | -5.309453 | 1.133655 | -4.68 | 0.000 | -7.531376 | -3.087531 |

Desde esta misma opción, incorporando la subinstrucción **or** pueden obtenerse los mismos resultados hallados con la instrucción **logistic:**

```
logi chd age, or
```

***Interpretación:*** *la edad parece predecir la cardiopatía isquémica. Su OR vale 1,117. Como la edad se ha introducido como variable cuantitativa, esto significa que por cada año más de edad se incrementa relativamente la odds de cardiopatía isquémica en un 11,7%. El intervalo de confianza hallado presenta un límite inferior de 1,07 y un límite superior de 1,17. Por tanto, no incluye el valor nulo, es estadísticamente significativo.*

## 13.12. TEST DE HOSMER-LEMESHOW: BONDAD DE AJUSTE O CALIBRACIÓN

Una vez obtenida la ecuación logística debe dársele solo un carácter *preliminar,* ya que antes de considerarla como definitiva debe evaluarse cómo se ajusta el modelo a los datos. Ajustarse bien supone que, globalmente, las predicciones no andarán lejos de la realidad. Existen diferentes técnicas para verificar la bondad de ajuste, pero, para modelos que incluyan al menos una variable cuantitativa, la prueba más aceptada es el test de Hosmer-Lemeshow (**11**).

El test consiste en observar detalladamente cómo se separan los valores observados (número de sujetos con el suceso y número de sujetos sin el suceso) con respecto a los esperados según las predicciones del modelo de regresión logística. Se comparan observados y esperados en diferentes categorías consideradas *a priori,* generalmente cada valor individual de la variable que recoge la probabilidad de evento predicha por el modelo de predicción o en cada decil de dicha variable. Si el ajuste es bueno, es de esperar que haya coincidencia entre observados y esperados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de ji cuadrado con $k - 2$ grados de libertad, siendo $k$ el número de categorías creadas *a priori* (p. ej., deciles). Si con esta prueba se obtiene un resultado *significativo,* existirá una *falta de ajuste* del modelo a los datos.

Ejemplo: para predecir la cardiopatía isquémica a partir de la edad, se halló el modelo logístico antes comentado: logit *(chd)* = –5,309 + 0,111 *age* ¿Se puede afirmar que el modelo se ajusta bien a los datos? La respuesta viene dada por el test de Hosmer-Lemeshow. Este test ordena los sujetos según las predicciones del modelo de regresión logística. Además, recoge los valores observados (Obs) y los predichos por el modelo (Exp) para cada valor de predicción.

Probabilidad predicha de *chd* según edad ordenada de menor a mayor

Observado con *chd*

Esperado sin *chd*

| Group | Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
|---|---|---|---|---|---|---|
| 1 | 0,0435 | 0 | 0,0 | 1 | 1,0 | 1 |
| 2 | 0,0596 | 0 | 0,1 | 1 | 0,9 | 1 |
| 3 | 0,0662 | 0 | 0,1 | 1 | 0,9 | 1 |
| 4 | 0,0733 | 1 | 0,1 | 1 | 1,9 | 2 |
| 5 | 0,0812 | 0 | 0,2 | 2 | 1,8 | 2 |
| 6 | 0,0994 | 0 | 0,2 | 2 | 1,8 | 2 |
| 7 | 0,1098 | 0 | 0,1 | 1 | 0,9 | 1 |
| 8 | 0,1211 | 1 | 0,7 | 5 | 5,3 | 6 |
| 9 | 0,1468 | 0 | 0,3 | 2 | 1,7 | 2 |
| 10 | 0,1612 | 0 | 0,3 | 2 | 1,7 | 2 |

Dos sujetos con una predicción del 8,12% de sufrir *chd* (a partir de la edad)

Existirán tantos grupos como valores de probabilidad se calculen. Se ha simplificado el listado presentando solo los 10 primeros valores.

Los valores observados de *chd* (Obs) y los predichos (Exp, *expected* en inglés) según la variable *edad* son contrastados según diferentes categorías, en este ejemplo, el contraste Obs y Exp es llevado a cabo para cada valor de probabilidad.

La $\chi^2$ que se corresponde a estos datos es:

$$\chi^2 = \sum \frac{(o+e)^2}{e} = \frac{(0-0{,}1)^2}{0{,}1} + \frac{(1-0{,}9)^2}{0{,}9} \cdots + \frac{(1-0{,}9)^2}{0{,}9} + \frac{(1-0{,}1)^2}{0{,}1} = 21{,}31$$

Resultado obtenido con STATA:

```
      number of observations =       100
number of covariate patterns =        43
      Pearson chi2(41)       =     21,31
      Prob > chi2            =    0,9953
```

El valor de la ji cuadrado = 21,31 con 41 grados de libertad claramente no es significativo. La no significación indica que los observados son parecidos a los esperados. Más importante que la significación estadística de la ji cuadrado es *valorar una a una las categorías de riesgo y ver si hay disparidad* entre observados y esperados. Esa observación atenta indicará en qué regiones de la predicción el modelo se ajusta peor a los datos. En el ejemplo, todo indica que el ajuste a los datos es excelente.

Cuando no se ajuste bien el modelo a los datos, deberían especificarse de otra manera las variables independientes, recategorizarlas, plantearse introducir interacciones (términos de producto) o términos cuadráticos, y repetir de nuevo el test de Hosmer-Lemeshow.

STATA realiza el test de Hosmer-Lemeshow para la bondad de ajuste a través de la instrucción:

```
estat gof
```

Puede solicitarse la tabla de frecuencias de esperados y observados a través de la subinstrucción:

```
table
```

El resultado presentado (comparación Obs-Exp para cada valor de probabilidad predicha por el modelo) es la comparación que realiza STATA por defecto. Se podría solicitar que el programa realizara la comparación no por valores predichos (hasta 43), sino por deciles (10 grupos). Para ello debe incluirse la subinstrucción:

`group()`

En el ejemplo presentado, la instrucción sería:

```
estat gof, t g(10)
```

Logistic model for chd, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

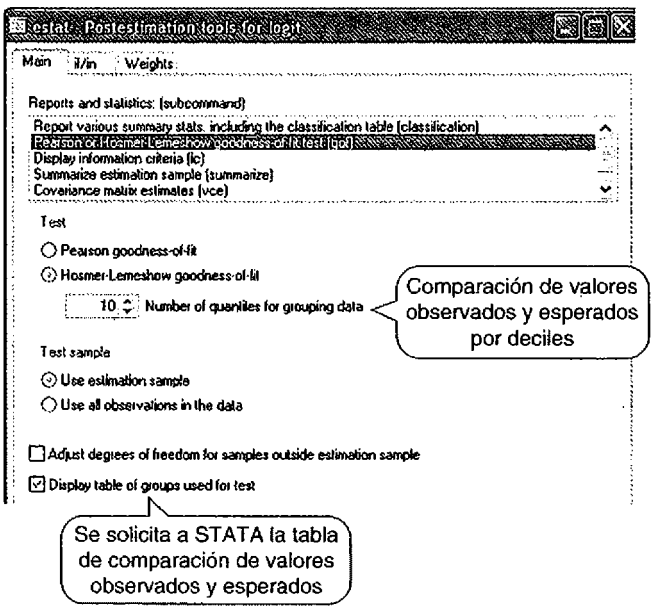
| Group | Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
|---|---|---|---|---|---|---|
| 1 | 0,1155 | 1 | 0,8 | 9 | 9,2 | 10 |
| 2 | 0,1690 | 1 | 1,3 | 9 | 8,7 | 10 |
| 3 | 0,2210 | 2 | 1,9 | 8 | 8,1 | 10 |
| 4 | 0,3183 | 3 | 3,0 | 8 | 8,0 | 11 |
| 5 | 0,3944 | 4 | 4,1 | 7 | 6,9 | 11 |
| 6 | 0,5037 | 5 | 4,7 | 5 | 5,3 | 10 |
| 7 | 0,6256 | 3 | 4,5 | 5 | 3,5 | 8 |
| 8 | 0,7336 | 12 | 10,5 | 3 | 4,5 | 15 |
| 9 | 0,7841 | 4 | 3,8 | 1 | 1,2 | 5 |
| 10 | 0,9125 | 8 | 8,4 | 2 | 1,6 | 10 |

```
number of observations =       100
      number of groups =        10
Hosmer-Lemeshow chi2(8) =      2,22
           Prob > chi2 =    0,9734
```

El valor de la ji cuadrado obtenida es diferente al caso anterior, así como el aspecto de la tabla. En este caso, cada grupo es un decil.

STATA realiza el test de bondad de ajuste también a través de la instrucción del menú:

- **Statistics → Postestimation → Reports and statistics**



## 13.13. CURVAS ROC: DISCRIMINACIÓN

A través de la construcción de curvas ROC se muestra el poder predictivo y de discriminación del modelo. Antes deben definirse los siguientes términos sensibilidad, especificidad y curva ROC.

**Sensibilidad** es la capacidad que posee un test (en nuestro caso, un modelo logístico de predicción) para identificar correctamente a los sujetos con la enfermedad o suceso. Corresponde a la probabilidad de que un sujeto realmente enfermo o con el suceso (evento observado = 1) sea clasificado como tal por el modelo de predicción porque su valor pronosticado *(p)* esté por encima del punto de corte para el que se calcula la sensibilidad. Por ejemplo, puede calcularse la sensibilidad para un punto de corte $p = 0,5$. Si se observa que el 67% de los que realmente tuvieron el evento tenían una $p \geq 0,5$ la sensibilidad será 0,67. La tabla de clasificación para un corte a $p = 0,5$ es proporcionada rutinariamente por STATA a través del menú:

**Statistics → Postestimation → Reports and statistics**

o con la instrucción:

```
estat classification
```

```
Logistic model for chd

                    True
Classified |         D            ~D |      Total
-----------+--------------------------+-----------
         + |        29            12 |         41
         - |        14            45 |         59
-----------+--------------------------+-----------
     Total |        43            57 |        100

Classified + if predicted Pr(D) >= .5
True D defined as chd != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   67,44%
Specificity                     Pr( -|~D)   78,95%
Positive predictive value       Pr( D| +)   70,73%
Negative predictive value       Pr(~D| -)   76,27%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   21,05%
False - rate for true D         Pr( -| D)   32,56%
False + rate for classified +   Pr(~D| +)   29,27%
False - rate for classified -   Pr( D| -)   23,73%
--------------------------------------------------
Correctly classified                        74,00%
```

***Interpretación:*** *se observa que de un total de 43 sujetos (14 + 29) con cardiopatía isquémica, el punto de corte 0,5 clasifica correctamente a 29 de ellos. La sensibilidad para un punto de corte de 0,5 será:*

$$\text{Sensibilidad} = \frac{\text{verdaderos positivos}}{\text{total eventos}} = \frac{29}{29+14} = 0,674$$

**Especificidad** es la capacidad que dispone el modelo para clasificar correctamente los sujetos *sin* evento. Se traduce como la probabilidad de que un sujeto sin evento (evento observado = 0) sea clasificado como tal por el modelo logístico (probabilidad predicha o esperado < punto de corte de *p*). En el ejemplo, para un corte de $p = 0,5$, la especificidad sería:

$$\text{Especificidad} = \frac{\text{verdaderos negativos}}{\text{total sin evento}} = \frac{45}{45+12} = 0,789$$

Se ha considerado solo un punto de corte, pero podrían calcularse sensibilidad y especificidad con otros puntos de corte. La curva ROC es un modo de considerarlos todos.

La **curva ROC** con un eje X y otro eje Y representa la sensibilidad (eje Y), pero no representa la especificidad, sino su complementario (1 – especificidad) en el eje X. Por tanto, es una forma útil de presentar ambas características (sensibilidad y especificidad) del modelo logístico cuando el punto de corte va cambiando. Según varía este punto de corte, se modificarán los valores de sensibilidad y especificidad del modelo logístico. A medida que varía el punto de corte, se obtendrán diferentes valores de sensibilidad y de (1 – especificidad), que son las coordenadas para representar cada punto gráficamente. La unión de estos puntos conforma la curva ROC.

La manera de cuantificar el poder discriminatorio del modelo es a través de la determinación del *área bajo la curva ROC (area under curve, AUC).* El poder discriminatorio es la capacidad de clasificar correctamente a los sujetos según su evento, equivale a la capacidad de conseguir que los sanos sean clasificados como sanos y los enfermos sean clasificados como enfermos[3].

Ejemplo: se pretende analizar el poder discriminatorio de un modelo de regresión logística para predecir la presencia de cardiopatía isquémica a partir de la edad de los sujetos, usando la misma base de datos anterior. La distribución de la probabilidad de sufrir *chd* predicha por el modelo se contrasta con la realidad de haber sufrido o no la *chd.* Para ello se guardan en STATA los valores predichos para cada sujeto en la regresión logística[4].

Los valores de sensibilidad y 1 – especificidad cambian a medida que se usan para hacer predicciones uno u otro punto de corte en la probabilidad pronosticada por la regresión logística. Por eso, la curva ROC tiene distintos puntos, cada uno con unas coordenadas para sensibilidad y 1 – especificidad. El valor representado por la diagonal correspondería a una capacidad de discriminación totalmente *nula* para distinguir entre quienes tienen y quienes no tienen *chd.* El área bajo la curva en esa situación sería AUC = 0,5. Esta es la hipótesis nula.

En STATA, la curva ROC puede obtenerse a través del menú:

- **Statistics → Binary outcomes → Postestimation → ROC curve after logistic/logit/probit/ivprobit**

o con la instrucción:

```
lroc
```

En el ejemplo se ha encontrado un área bajo la curva de 0,8 (fig. 13.11). Se concluiría que la edad discrimina bien entre eventos y no eventos de cardiopatía isquémica. Su poder de discriminación es el 80% del máximo posible.

## 13.14. CRITERIOS DE CONSTRUCCIÓN DE MODELOS EN REGRESIÓN LOGÍSTICA

Los criterios de construcción de modelos de regresión logística son similares a los empleados en la construcción de modelos de regresión múltiple. Existen indicaciones precisas y más detalladas sobre la construcción de modelos logísticos (**11**).

### 13.14.1. Construcción de gráficas dirigidas (DAG, *Directed Acyclic Graphs*)

Las gráficas dirigidas (DAG, **12**) pueden servir para establecer posibles asociaciones entre variables y para detectar factores de confusión. Sirven como síntesis de posibles asociaciones entre las variables que integran un análisis. Estas gráficas hacen explícitas las creencias existentes sobre relaciones causales entre variables y se usan para seleccionar el conjunto mínimo de variables que

3 Es equivalente a la probabilidad de clasificar correctamente a los sujetos, cada uno en su grupo, que se vio en el test de la *U* de Mann-Whitney.

4 A través de la instrucción **predict.**

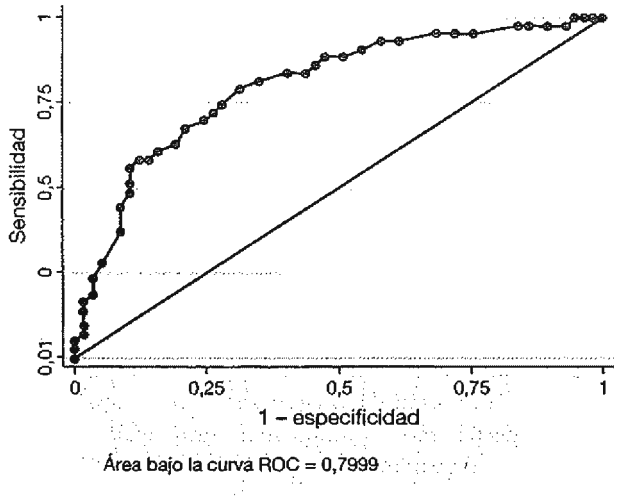



**Figura 13.11** Curva ROC obtenida en el ejemplo de la asociación entre la edad y la cardiopatía isquémica.

se deben considerar candidatas a su inclusión en el modelo. Deben basarse en el conocimiento científico existente sobre las relaciones entre variables y también han de tener en cuenta el diseño del estudio para situar temporalmente las variables.

Por ejemplo, una variable medida en el estudio después que otra no puede ser su causa. Una variable que se sitúa como eslabón intermedio en la cadena causal no debe ser considerada como factor de confusión, sino como mecanismo de acción. Se identificarán como factores de confusión aquellas variables que pueden ser causa común de la exposición y del efecto.

Imagine cinco variables (A, B, C, D y E). Una posible relación entre las variables es la que se recoge en la figura 13.12.

La construcción de gráficas dirigidas permitirá sospechar la presencia de posibles factores de confusión que distorsionarían la asociación real entre una exposición (X) y un desenlace (Y) (v. apartado 13.7). Los factores de confusión deberán ser tenidos en cuenta en los análisis en los que se detectará si finalmente actúan o no como factores de confusión y deben ser o no introducidos en los modelos finales.

### 13.14.2. Análisis estratificado

El análisis estratificado puede utilizarse como paso preliminar para identificar posibles factores de confusión e interacciones antes de la construcción de un modelo de regresión logística multiva-

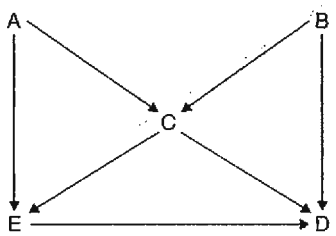



**Figura 13.12** Ejemplo de gráfica dirigida con cinco variables. A es causa de E y de C. A su vez, E es causa de D. B tiene influencia sobre D y sobre C. C, a su vez, la tiene sobre E y D.

riante. Se calculará la OR ponderada de Mantel-Haenzsel, que promedia la OR de cada estrato. Se estratificará por los factores de confusión potenciales.

Para solicitar una OR ponderada de Mantel-Haenzsel para cada estrato de la variable 2 *($var_2$)*, siendo *y* la variable dependiente (casos) y *var1* la variable de exposición, en STATA se debe indicar la siguiente instrucción:

```
cc y var1, by(var2)
```

(cc procede de la instrucción **case-control**).

### 13.14.3. Regresión logística univariante

Antes de llevar a cabo un análisis multivariante, resulta muy útil realizar un análisis univariante para establecer la asociación de cada variable independiente (predictora) con la variable dependiente (evento). En el caso de las variables cuantitativas, debe comprobarse si son lineales en el logit categorizándolas a través de la creación de categorías excluyentes, preferiblemente cuantiles, y comprobar si la OR crece o decrece monotónicamente. Una variable independiente que presente un valor $p < 0{,}25$ en el análisis univariante de regresión logística sería, en principio, candidata a formar parte del análisis multivariante (v. apartado 13.7).

### 13.14.4. Regresión logística multivariante

Se creará un modelo de regresión multivariante provisional sin incluir los términos de interacción. Solo se construirá el *modelo de efectos principales*. Se probará a ir eliminando del modelo, una a una, todas aquellas variables con coeficientes de regresión que presenten pruebas de razón de verosimilitud sin significación estadística y para las que no exista una razón biológica irrefutable que lleve a la necesidad de ajustar por ellas. Antes de decidir definitivamente eliminarlas, debe tenerse muy en cuenta cómo cambia la OR de la variable principal cuyo efecto se valora al eliminar alguna de estas variables (pueden ser factores de confusión, aunque no sean significativas).

### 13.14.5. Valorar posibles interacciones

Tal y como se indicó en el apartado 13.8, solamente aquellas interacciones significativas y fácilmente interpretables serán incluidas en el modelo final.

### 13.14.6. Comprobar la bondad de ajuste

Normalmente se realiza a través del test de Hosmer-Lemeshow. Se deben verificar las diferencias en cada una de las casillas consideradas entre observados y esperados, y no solo valorar que la ji cuadrado no es significativa. Si la ji cuadrado fuese significativa o hubiese clara disparidad entre observados y esperados, debe especificarse mejor el modelo (cambiar categorizaciones, probar términos cuadráticos o interacciones).

### 13.14.7. Construcción de una curva ROC en modelos predictivos

Solo si se pretende utilizar el modelo logístico con carácter predictivo debe construirse una curva ROC y hallar el área bajo la curva con su correspondiente límite de confianza. Esta área proporciona la capacidad de discriminación del modelo logístico.

## 13.15. REGRESIÓN LOGÍSTICA CONDICIONAL

La regresión logística condicional es un procedimiento estadístico aplicado en los estudios de casos y controles emparejados o apareados, donde cada caso es «emparejado» con un control con el que comparte algunas características. Estas características suelen ser el sexo o la edad[5] (**8**).

5 Téngase en cuenta que en este tipo de análisis no es necesario ajustar por las variables de emparejamiento.

| | par | deporte | infarto |
|---|---|---|---|
| 1 | 1 | No | No |
| 2 | 1 | No | Sí |
| 3 | 2 | No | No |
| 4 | 2 | Sí | Sí |
| 5 | 3 | No | No |
| 6 | 3 | No | Sí |
| 7 | 4 | Sí | No |
| 8 | 4 | No | Sí |
| 9 | 5 | No | No |
| 10 | 5 | No | Sí |
| 11 | 6 | Sí | No |
| 12 | 6 | No | Sí |
| 13 | 7 | Sí | No |
| 14 | 7 | Sí | Sí |

**Figura 13.13** Aspecto parcial de una base de datos en STATA para análisis de regresión logística condicional.

Imagine un estudio de casos y controles en el que se quiere valorar el efecto de la práctica deportiva *(deporte*; 0 = no; 1 = sí) sobre la ocurrencia de infarto *(infarto*; 0 = no; 1 = sí). Para ello se selecciona un grupo de casos (infarto = 1) y un grupo de controles (infarto = 0) y se pregunta a sus integrantes por su práctica deportiva en el pasado. Los casos y los controles están emparejados por edad y sexo (por tanto, a un caso le corresponde un control, pertenecen al mismo par). El aspecto de la base de datos en STATA se muestra en la figura 13.13.

Las instrucciones en STATA para llevar a cabo el análisis son las siguientes:

```
clogit  y x1 x2  xp, group(variable de emparejamiento)

clogit y x1x2  xp, group(variable de emparejamiento) or
```

En el ejemplo:

```
clog infarto deporte, gr(par)
```

| infarto | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| deporte | -.405465 | .9128709 | -0.44 | 0.657 | -2.194659 | 1.383729 |

```
clog infarto deporte, gr(par) or
```

| infarto | Odds Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| deporte | .6666667 | .6085807 | -0.44 | 0.657 | .1113965 | 3.989752 |

***Interpretación:*** *aunque la práctica de actividad física se asocia a una reducción del riesgo de infarto, el resultado obtenido en este estudio de casos y controles no es estadísticamente significativo (OR = 0,67; IC 95% = 0,11-3,99).*

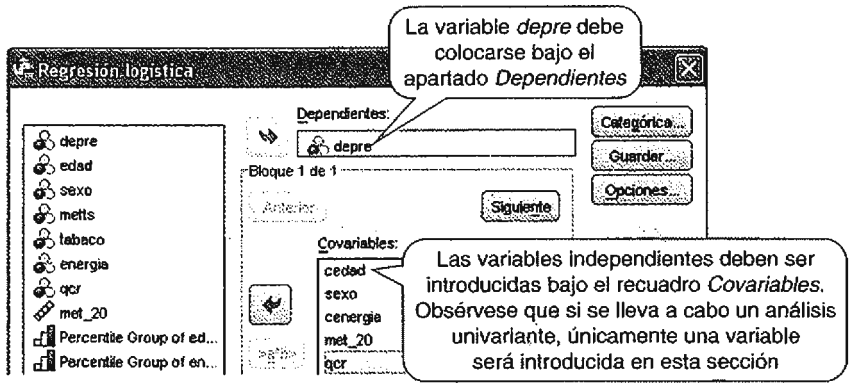
## 13.16. REGRESIÓN LOGÍSTICA EN SPSS

SPSS permite, como STATA, llevar a cabo un análisis de regresión logística desde el menú o con instrucciones.

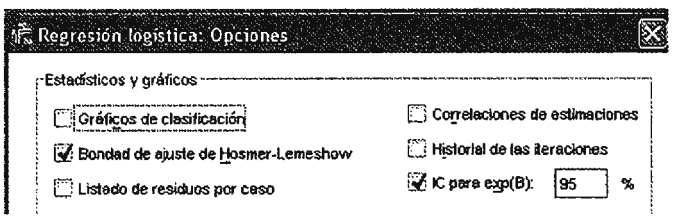
En el siguiente ejemplo se trata de valorar el efecto del consumo de comida rápida (quintiles) *(qcr)* sobre el riesgo de depresión *(depre)* independientemente de otros factores, como la edad (cuartiles) *(cedad)*, el sexo, el nivel de actividad física (incremento en 20 MET-h/semana[6]*(met_20)* o el consumo energético total (cuartiles) *(cenergía)* (13,14).

Desde el menú:

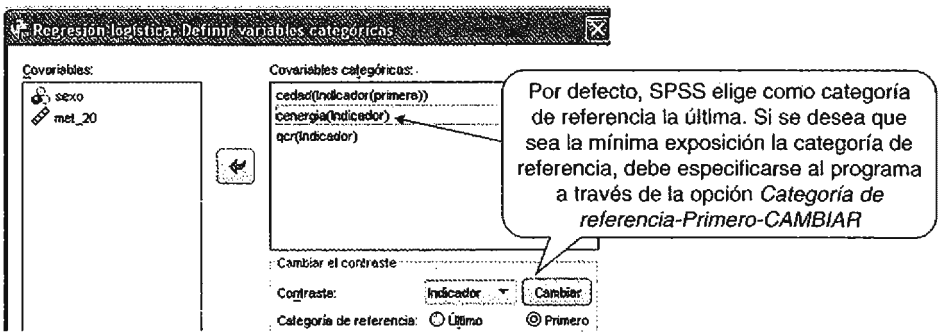
**Analizar → Regresión → Logística Binaria**




Este mismo cuadro de diálogo presenta tres botones: CATEGÓRICA, GUARDAR y OPCIONES. Desde el botón OPCIONES puede solicitarse a SPSS el test de Hosmer-Lemeshow y la visualización de los intervalos de confianza de las OR calculadas para cada variable independiente. En este programa, los intervalos de confianza no aparecen por defecto en los resultados.

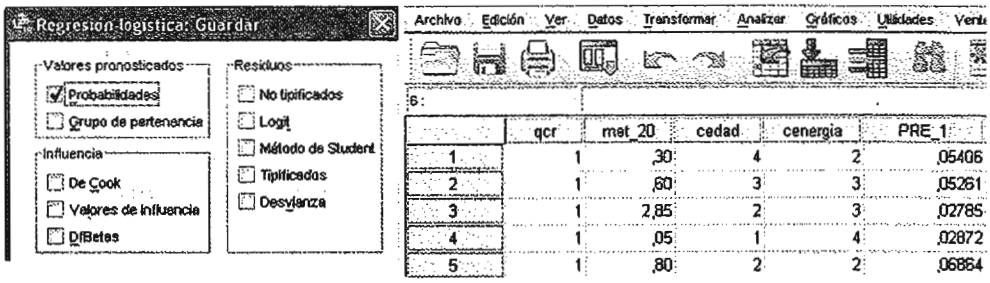



Desde el botón CATEGÓRICA pueden crearse variables indicadoras o *dummy*:



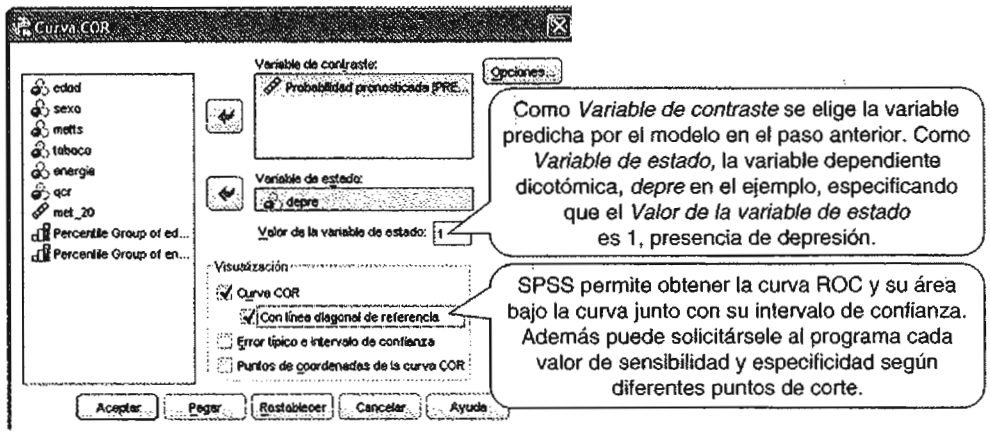

6 MET-h/semana: equivalentes metabólicos semanales. Un MET-h se define como la cantidad de energía empleada por el organismo estando sentado en 1 hora.

Desde el botón GUARDAR, SPSS es capaz de crear la variable *PRE_1*, que representa el valor predicho de depresión que tendría un individuo en concreto de acuerdo con los valores que tome el resto de variables del modelo.

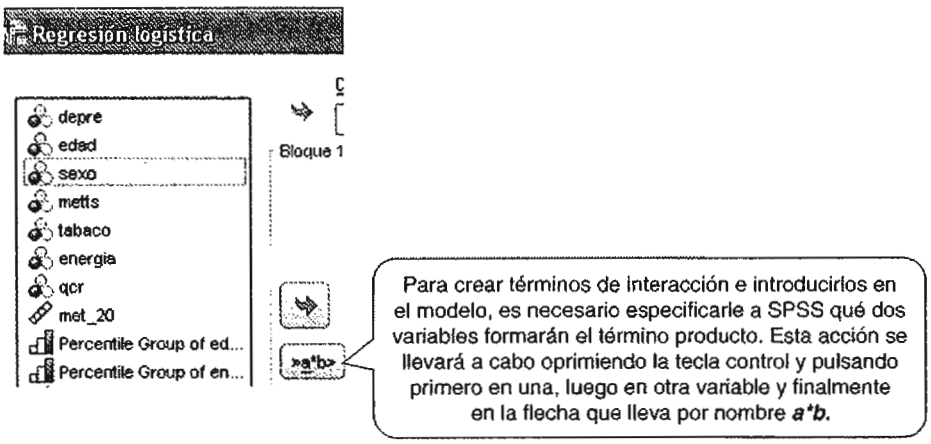



Una vez guardada esta variable de predicción, puede calcularse la curva ROC. En este caso es necesario acudir a otro menú del programa:

**Analizar → Curva COR**




Por último, ¿cómo se crean los términos producto en SPSS?




El aspecto de los resultados es similar al obtenido con el programa STATA.

## 13.17. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| Cálculo de coeficientes de regresión | **logit** *y* $x_1$ $x_2$ $x_3$ $x_p$ | **LOGISTIC REGRESSION VARIABLES** *y* **/METHOD=ENTER** $x_1$ $x_2$ $x_3$ $x_p$. |
| Cálculo de OR | **logit** *y* $x_1$ $x_2$ $x_3$ $x_p$, **or** **logistic** *y* $x_1$ $x_2$ $x_3$ $x_p$ | p.d. |
| IC 95% de la OR | p.d. | **/PRINT= CI(95)** |
| Variable cualitativa con más de dos categorías (a), categoría de referencia, valor más bajo | **i.***a* | **/CONTRAST (***a***)=Indicator(***1***)** |
| Interacción entre dos variables ($x_1$ y $x_2$) | • **generate** *inter*=$x_1$*$x_2$ **logistic** *y* $x_1$ $x_2$ **estimates store** *A* (Para guardar los resultados del modelo sin interacción) | Para interacción entre dos variables cuantitativas o cualitativas dicotómicas: **LOGISTIC REGRESSION VARIABLES** *y* **/METHOD=ENTER** $x_1$ $x_2$ $x_1$*$x_2$. Cuando una variable cualitativa con más de dos categorías se ve implicada: **LOGISTIC REGRESSION VARIABLES** *y* **/METHOD=ENTER** $x_1$ $a_1$ $x_1$*$a_1$ **/CONTRAST (***a***)=Indicator(1).** |
| | **logistic** *y* $x_1$ $x_2$ *inter* **lrtest** *A*. (STATA llama «.» a todos los resultados del último modelo con interacción) | |
| Valores predichos | **predict** *pred*[a] | **/SAVE=PRED** |
| Bondad de ajuste (Hosmer-Lemeshow), presentación de tabla de deciles | **estat gof, t g(***10***)** | **/PRINT=GOODFIT** |
| Sensibilidad y especificidad | **estat classification** | p.d. |
| Curva ROC con línea de referencia en diagonal | **lroc** | **ROC** *pred*[a] **BY** *y*(1) **/ PLOT=CURVE(REFERENCE) /PRINT=SE** |

| Procedimiento | STATA | SPSS |
|---|---|---|
| Intervalo de confianza de la curva | STATA no lo realiza[b] | |
| *Ejemplo:* valorar el efecto del consumo de comida rápida (quintiles) sobre la depresión a igualdad de sexo, edad (cuartiles), actividad física y consumo de energía (cuartiles); predecir la probabilidad de depresión según los valores tomados por estas variables *(pred_depre)* y valorar la validez del modelo | **logistic *depre i.qcr sexo i.cedad met_20 i.cenergia***<br>**predict *pred_depre*** | **LOGISTIC REGRESSION VARIABLES *depre***<br>**/METHOD=ENTER *qcr sexo cedad met_20 cenergia***<br>**/CONTRAST (*qcr*)=Indicator(*1*)**<br>**/CONTRAST (*cedad*)=Indicator(*1*)**<br>**/CONTRAST (*cenergia*)=Indicator(*1*)**<br>**/PRINT=GOODFIT CI(95)**<br>**/SAVE=PRED.**<br>**VARIABLE LABELS PRE_1 '*pred_depre*'.** |

[a]Debe especificarse al programa estadístico el nombre que quiere dársele a la variable predicha (p. ej., *pred*). Por defecto, SPSS crea la variable predicha con etiqueta *PRE_1*. Se deberá cambiar el nombre de la variable si así se desea (p. ej., *pred*).
[b]Debe acudirse a la instrucción **roctab *y pred.***
p.d., por defecto (el programa calcula el parámetro sin añadir instrucciones); *a*, variables independientes cualitativas de más de dos categorías; *x*, variables independientes cuantitativas o cualitativas dicotómicas; *y*, variable dependiente.

## REFERENCIAS


1. Martín-Moreno JM. Oportunidad relativa: reflexiones en torno a la traducción del término "odds ratio". Gac Sanit 1990;4(16):37.
2. Bautista LE. "Razón relativa" y "tasa relativa" como traducciones de odds ratio y de hazard ratio. Bol Ofic Sanit Panam 1995;119(3):278-80.
3. Lachenbruch PA. The odds ratio. Control Clin Trials 1997;18(4):381-2.
4. McNutt LA, Hafner JP, Xue X. Correcting the odds ratio in cohort studies of common outcomes. JAMA 1999;282(6):529.
5. De Irala J, Martínez-González MA, Guillén Grima F. ¿Qué es una variable de confusión? Med Clin (Barc) 2001;117(10):377-85. Fe de erratas: Med Clin (Barc) 2001;117(20):775.
6. Bland JM, Altman DG. Statistics notes. The odds ratio. BMJ 2000;320(7247):1468.
7. Jewell NP. Statistics for epidemiology. Boca Raton: Chapman & Hall/CRC Press; 2004.
8. De Irala J. Diseños en investigación en epidemiología. En: De Irala J, Martínez-González MA, Seguí-Gómez M, editores. Epidemiología aplicada. 2.ª ed. Madrid: Ariel Ciencias Médicas; 2008. p. 179-272.
9. Sánchez-Villegas A, Field AE, O'Reilly EJ, Fava M, Gortmaker S, Kawachi I, Ascherio A. Perceived and actual obesity in childhood and adolescence and risk of adult depression. J Epidemiol Community Health 2013;67(1):81-6.
10. De Irala J, Martínez-González MA, Guillén Grima F. ¿Qué es una variable modificadora del efecto? Med Clin (Barc) 2001;117(8):297-302.

11. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. Hoboken: John Wiley & Sons; 2000.

12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999;10(1):37-48.

13. Sánchez-Villegas A, Toledo E, de Irala J, Ruiz-Canela M, Pla-Vidal J, Martínez-González MA. Fast-food and commercial baked goods consumption and the risk of depression. Public Health Nutr 2012;15(3):424-32.

14. Sánchez-Villegas A, Ara I, Guillén-Grima F, Bes-Rastrollo M, Varo-Cenarruzabeitia JJ, Martínez-González MA. Physical activity, sedentary index, and mental disorders in the SUN cohort study. Med Sci Sports Exerc 2008;40(5):827-34.

# ASPECTOS AVANZADOS DE REGRESIÓN DE COX 14

*M. Ruiz-Canela, E. Toledo, J. López-Fidalgo, M. Á. Martínez-González*

## 14.1. INTRODUCCIÓN: CONTEXTO Y UTILIDAD

Cuando se trata a un enfermo no solo importa que se cure, sino que lo haga lo antes posible. El modelo de regresión de Cox tiene en cuenta ambas cosas: si se produce o no el resultado esperado (curación, en este caso) y cuánto tiempo tarda en ocurrir (**1,2**). Esta consideración es importante en el ámbito clínico, especialmente en enfermedades graves.

Otra ventaja de la regresión de Cox (como sucedía con el método de Kaplan-Meier; v. capítulo 11) es que todos los sujetos del estudio cuentan, incluso los que se pierden sin saber si sanaron o no. Ocurre así porque los modelos de regresión de Cox manejan el tiempo de todos los participantes que están en riesgo de enfermar en cada posible momento o instante de fallecer durante todo el seguimiento. Solo se consideran, no obstante, los tiempos o instantes en que de hecho falleció alguien.

## 14.2. CONCEPTOS DE *HAZARD* Y *HAZARD RATIO*

En la regresión de Cox hay dos conceptos fundamentales: *hazard* y *hazard ratio* (**3**). El *hazard* o tasa instantánea muestra el riesgo de experimentar un evento en un instante del estudio. Es un concepto semejante al de la velocidad instantánea que marca el cuentakilómetros de un vehículo en un momento determinado. Para aproximar esa velocidad instantánea, se usa la información sobre lo que ha ocurrido en un entorno de dicho punto temporal. Así, una vez elegido un incremento temporal, bastará con medir la distancia recorrida. El cuenta kilómetros calcula la velocidad instantánea teniendo en cuenta que será aproximadamente idéntica al cociente del espacio recorrido dividido entre el tiempo transcurrido (incremento temporal). Cuanto menor sea ese incremento, más cerca se trabajará del momento puntual en cuestión y, por tanto, mejor será la aproximación a la estimación de una velocidad instantánea.

En la regresión de Cox, el riesgo de un paciente en un momento determinado podría aproximarse por medio del cociente de la probabilidad de que el paciente sobreviva en un intervalo o incremento temporal, sabiendo que ha sobrevivido hasta ese momento, dividido por dicho incremento temporal. Por ejemplo, para conocer el riesgo de muerte de un paciente que ha sobrevivido durante 90 días después de la operación, podría calcularse la probabilidad de que no fallezca en los 5 días siguientes o bien, con un incremento menor, en los 2 días siguientes. El concepto matemático de límite proporciona una expresión para este riesgo cuando ese incremento temporal se hace tender a 0. En particular, para cada tiempo $t$, el *hazard* se calcula dividiendo el número de eventos ocurridos en ese preciso instante entre el número total de personas en riesgo:

$$hazard_t = \lambda_t = \frac{d_t}{n_t}$$

En la figura 14.1 se representa un ejemplo de un sencillo estudio con seis sujetos. El valor del *hazard* ($\lambda$) en el instante $t = 2$ es de 1/5. El denominador incluye solo cinco, porque el primer participante abandonó el estudio antes del segundo año y no se incluye en el total de sujetos en riesgo en ese instante *(risk set)*. De este modo, el *hazard* representa la probabilidad condicional

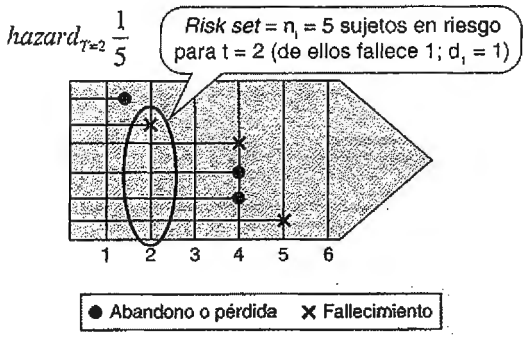



**Figura 14.1** *Hazard* o tasa instantánea a tiempo $t = 2$ en el seguimiento de seis pacientes.

*instantánea* de presentar el evento en ese momento, con la condición de que no haya aparecido antes de ese instante. Esta medición instantánea es útil, porque permite tener en cuenta los cambios del riesgo a lo largo del tiempo. Como se muestra a modo de ejemplo en la figura 14.2, la función del riesgo *(hazard)* puede variar de formas muy diversas durante el período de seguimiento.

Cuando se comparan dos grupos con diferente exposición, puede obtenerse un *hazard* para cada grupo en el mismo tiempo *t*. La función *hazard ratio* es el cociente de los riesgos instantáneos o *hazards* de cada grupo. Una *hazard ratio* indica si existe una relación entre una exposición y un evento en un tiempo *t*. Si el valor de la *hazard ratio* es superior a 1, significa que la tasa de riesgo instantáneo es mayor en el grupo expuesto que en el de no expuestos. Si valiese 1, correspondería a la igualdad de riesgo en ambos grupos; si fuera inferior a 1, se trataría de una exposición protectora.

## 14.3. EJEMPLO RIDÍCULAMENTE SENCILLO DE REGRESIÓN DE COX UNIVARIANTE

Supóngase que durante 48 meses se compara la supervivencia entre un grupo de 10.000 fumadores actuales con 10.000 personas que nunca han fumado.

Como se observa en la tabla 14.1, hay 12 muertes en los fumadores: 4 ocurrieron a los 14 meses, 4 a los 17 meses y 4 a los 28 meses. Los 9.988 fumadores restantes, observados hasta cumplir

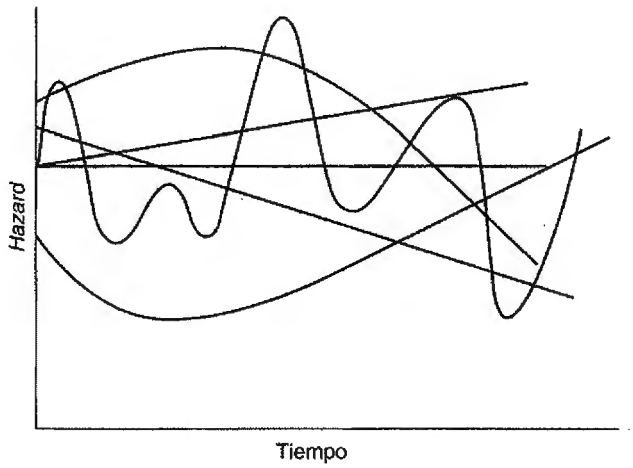



**Figura 14.2** Posibles cambios del *hazard* en función del tiempo.

*Tabla 14.1 Ejemplo ridículamente simple de regresión de Cox*

| | MUERTES A LOS... | | | CENSURADOS (+) |
|---|---|---|---|---|
| | 14 MESES | 17 MESES | 28 MESES | 44 MESES |
| Fumadores actuales | 4 | 4 | 4 | 9.988+ |
| Nunca fumadores | 1 | 1 | 1 | 9.997+ |
| *Hazard* fumadores | 4 | 4 | 4 | |
| | 10.000 | 9.996 | 9.992 | |
| *Hazard* no fumadores | 1 | 1 | 1 | |
| | 10.000 | 9.999 | 9.998 | |
| *Hazard ratio* fumadores frente a no fumadores | 4 | 4,001 | 4,002 | |

44 meses, fueron *censurados* (v. capítulo 11), pues no habían muerto durante el seguimiento. Entre los nunca fumadores solo hubo 3 muertes, una a los 14 meses, otra a los 17 meses y otra más a los 28. El resto de los sujetos, hasta completar los 10.000, es decir, 9.997, seguían vivos al final de los 44 meses de seguimiento y fueron censurados. ¿Cuánto valdrá la *hazard ratio* de los fumadores comparados con los nunca fumadores?

Intuitivamente es fácil deducir que el riesgo o *hazard* instantáneo de mortalidad es siempre aproximadamente cuatro veces superior en los fumadores que en los no fumadores, pues para todos los tiempos, ya sea a los 14 meses, a los 17 o a los 28, siempre hay aproximadamente cuatro veces más muertes en los fumadores. A los 44 meses no se debe calcular nada, ya que no murió nadie en ese último período. Si se ajustase con estos datos un modelo de regresión de Cox, el ordenador produciría en la salida un valor estimado de *hazard ratio* ligeramente superior a 4.

## 14.4. LA ECUACIÓN DE LA REGRESIÓN DE COX

La ecuación de la regresión de Cox se expresa del siguiente modo:

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + b_1x_1 + \ldots + b_px_p$$

Externamente, este modelo se parece mucho a la regresión logística, pero contiene dos diferencias fundamentales:

1. El término de la parte izquierda de la ecuación incluye el logaritmo neperiano de un *hazard* (en vez de una *odds*, como ocurre con la regresión logística). Por tanto, en la regresión de Cox, la variable dependiente es el logaritmo neperiano de la *hazard*, $\ln(\lambda_t)$. Este *hazard* es una función variable con el tiempo.
2. El primer término de la parte derecha incluye otro logaritmo neperiano, el de $\lambda_{0t}$ *(baseline hazard)*, que es el riesgo basal instantáneo. Se trata también de una función que cambia con el tiempo y corresponde a aquellos sujetos para los cuales todas las variables independientes valen 0, es decir, para quienes el resto de la parte derecha de la ecuación $(b_1x_1 + \ldots + b_px_p)$ es igual a 0.

La continuación de la ecuación que sigue al *baseline hazard* corresponde a las covariables cuyo efecto en el tiempo hasta el evento se quiere valorar, lo que incluye la exposición principal y los factores de confusión potenciales.

La ecuación se puede expresar también como una función exponencial:

$$\lambda_t = \lambda_{0t} * e^{b_1x_1+\ldots+b_px_p}$$

## 14.5. INTERPRETACIÓN DE LOS COEFICIENTES DE LA REGRESIÓN DE COX UNIVARIANTE

En un modelo de Cox univariante solo habrá una variable independiente en la ecuación:

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + bx$$

La covariable podría ser numérica o cualitativa binaria. Convencionalmente suele tener el valor de 1 si hay exposición y 0 cuando no la hay. Si fuese cualitativa con tres o más valores, se deberían incluir variables auxiliares binarias *(dummy)* y ya no sería univariante en sentido estricto. En general, la mayor parte del *software* estadístico que hoy se utiliza generará las variables *dummy,* siempre que se le indique así.

En el ejemplo del apartado anterior, la variable independiente era el tabaco (x). Por simplicidad, se asumen solo dos posibilidades: fumadores (x = 1) y no fumadores (x = 0).

El *hazard* para cada grupo en un tiempo *t* sería:

Fumadores: $\ln(\lambda_{tFUMA}) = \ln(\lambda_{0t}) + b_1 x_1 \rightarrow \lambda_{tFUMA} = \lambda_{0t} * e^{b_1}$.

No fumadores: $\ln(\lambda_{tNOFUMA}) = \ln(\lambda_{0t}) \rightarrow \lambda_{tNOFUMA} = \lambda_{0t}$.

La ha*zard ratio* (HR) se calcula dividiendo ambos *hazards:*

$$\textit{Hazard ratio}_{\text{FUMA vs. NO FUMA}} = \frac{\lambda_{tFUMA}}{\lambda_{tNOFUMA}} = \frac{\lambda_{0t} * e^{b_1}}{\lambda_{0t}} = e^{b_1}$$

En la figura 14.3 se muestra el resultado de la regresión de Cox obtenido con STATA.

La HR coincide con el exponencial del coeficiente $b_1$ ($b_1$ = 1,387). El número $e^{b_1}$ muestra la cantidad por la que se multiplica el riesgo cuando $x_1$ = 1. En este ejemplo, $e^{1,387}$ es igual a 4, aproximadamente. Es decir, en los fumadores, el riesgo de fallecer se multiplica por cuatro.

En la figura 14.3 se muestra también el valor *z,* que permite comprobar si la asociación del tabaco con la mortalidad es estadísticamente significativa. Se observa que la probabilidad de encontrar un valor *z* superior a 2,15 es estadísticamente significativa (valor *p* igual a 0,032, para una prueba a dos colas). Este test sigue una distribución normal y corresponde al concepto de test de Wald, explicado en el capítulo 10. Si se encuentra un resultado en el entorno de *p* = 0,05, se recomienda no fiarse de este test de Wald y basarse preferiblemente en un test de razón de verosimilitud, tal como se explicó en el capítulo 13. El procedimiento para obtener el test de razón de verosimilitud (*likelihood ratio test* o *LR*) es idéntico al que se usaba para la regresión logística.

En la figura 14.4 se representa, con otro ejemplo simple y ficticio, el logaritmo del *hazard* (eje *y*) en función del tiempo de seguimiento para dos grupos, fumadores y no fumadores. La escala logarítmica permite una modelización de los *hazards,* como si se tratase de ecuaciones de una recta. Se observa que los *hazards* no son constantes, sino que se incrementan con el tiempo.

En la gráfica se aprecia un hecho que el modelo de Cox asume: la diferencia entre las dos rectas es constante (sin olvidar que estamos en escala logarítmica) y siempre valdrá la misma cantidad. Para el grupo 1, el logaritmo del *hazard* será siempre superior en una constante b al del grupo 0. Al traducir este concepto a escala lineal en vez de logarítmica, lo que será constante es el cociente entre ambos *hazards,* es decir, se asume la proporcionalidad de los *hazards;* de ahí que el modelo de Cox sea conocido como *propotional hazards model.*

| _t | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| fum | 1.387 | 0.645 | 2.15 | 0.032 | 0.121 | 2.652 |
| | $b_1$ | EE($b_1$) | $z = b_1/EE(b_1)$ | | | |

**Figura 14.3** Resultados de la regresión de Cox obtenidos con STATA.

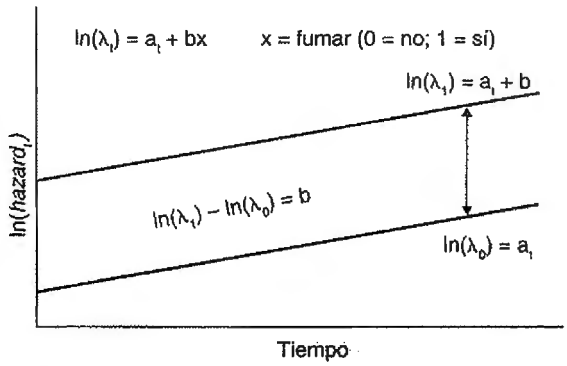



**Figura 14.4** Modelización del logaritmo de los *hazards* en función del tiempo.

Este es el fundamento de la regresión de Cox, que supone que la razón entre un *hazard* y otro es constante. Esto equivale a decir que los *hazards* son proporcionales, o que la diferencia de sus logaritmos es constante.

## 14.6. COMPARACIÓN DE CURVAS DE SUPERVIVENCIA CON LA REGRESIÓN DE COX

En la figura 14.5 se representa la curva de supervivencia con los datos del anterior ejemplo ridículamente sencillo. Se aprecia el número de sujetos en riesgo que se incluyen en el denominador para calcular el valor del *hazard* en cada tiempo *t* en el que se produce una muerte, tanto en fumadores como en no fumadores. Por ejemplo, al cabo de 17 meses, el número total en riesgo es 9.999 en los no fumadores y 9.996 en los fumadores.

En la parte inferior de la curva de Kaplan-Meier se muestra la salida de la regresión de Cox realizada con STATA. Se han marcado los estadísticos que permiten comprobar la significación global del modelo. *Iteration 0* muestra el valor inicial de verosimilitud, que se corresponde con la función de riesgo basal. La segunda verosimilitud que se ha destacado corresponde a la del modelo estimado tras incluir el tabaco como variable independiente. Si ambos valores se multiplican por –2, la diferencia entre ellos corresponderá al test de razón de verosimilitud (LR chi2) y su significación estadística. En este caso existe un grado de libertad, porque el tabaco tiene dos categorías (fumador, no fumador).

Se podría haber obtenido un resultado muy similar con el test del *log-rank* (v. capítulo 11). La ventaja de usar regresión de Cox es que, además, el análisis de Cox proporciona una medida de la magnitud del efecto, que es la *hazard ratio*. Como se había comprobado antes, la *hazard ratio* vale 4 (exactamente, será una cantidad ligeramente superior a 4, porque los denominadores de los fumadores se van reduciendo a medida que fallecen con más rapidez que los no fumadores). Una desventaja de la regresión de Cox con respecto al test del *log-rank*, solo aplicable a una única covariable cualitativa, es que ciertas desviaciones del supuesto de proporcionalidad de los *hazards* pueden inducir a veces mayor sesgo.

## 14.7. REGRESIÓN DE COX CON VARIABLE INDEPENDIENTE CUANTITATIVA

En la ecuación de regresión de Cox también se pueden introducir variables independientes cuantitativas, una opción muy frecuente en la investigación biomédica. El tiempo de aparición de numerosos desenlaces (muerte, infarto, deterioro cognitivo, etc.) depende de variables cuantitativas como la edad, el índice de masa corporal o la presión arterial. También, en la clínica se utilizan

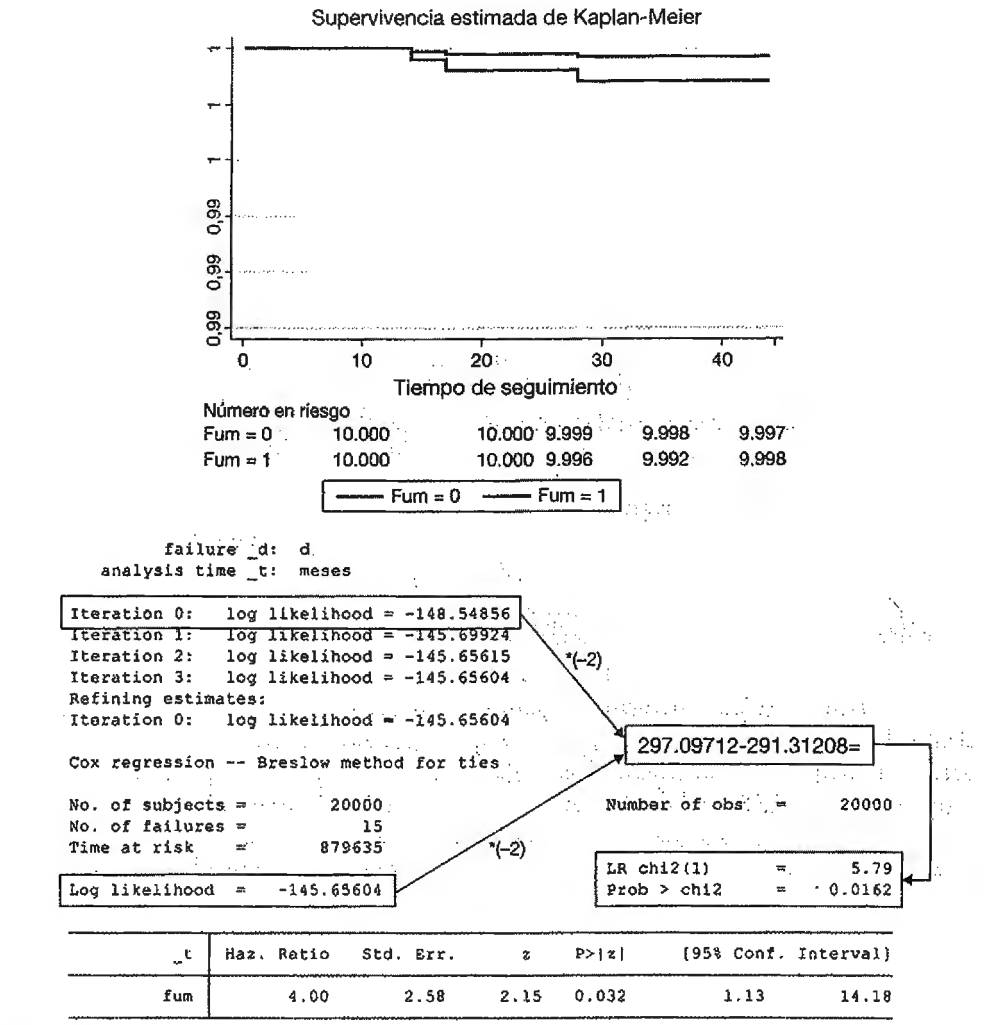



**Figura 14.5** Comparación de curvas de supervivencia utilizando una regresión de Cox.

habitualmente marcadores biológicos que son variables continuas, como, por ejemplo, la proteína C reactiva para predecir el riesgo de enfermedad cardiovascular.

La ecuación de la regresión se representa del mismo modo, si bien $x_1$ es una variable cuantitativa (discreta o continua). Según ello, teóricamente sus valores pueden ser infinitos, ya que se trata de una variable continua:

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + b_1 x_1$$

Cuando se introducen variables cuantitativas en un modelo de regresión de Cox, es importante no perder nunca de vista que lo ideal es cuantificar esa variable en la unidad de cambio que sea relevante desde un punto de vista clínico (**2**). Por ejemplo, tendría poco sentido modelizar el cambio del *hazard* asociado a la enfermedad cardiovascular por cada minuto más de vida. Clínicamente, sería interesante observar el cambio del *hazard* cada año o incluso, de forma preferible, en intervalos de 5 años.

## 14.8. INTERPRETACIÓN DE LOS COEFICIENTES DE VARIABLES INDEPENDIENTES CUANTITATIVAS

La *hazard ratio* para un valor concreto de una variable cuantitativa se calcula del siguiente modo:

$$HR = e^{x_1 * b_1}$$

El coeficiente $b_1$ se multiplica por el valor correspondiente de la variable cuantitativa ($x_1$).

Veámoslo con un ejemplo en el que se estudia la asociación entre arteriopatía periférica y consumo de ácido fólico (3). En el estudio se realizó un seguimiento de 46.000 varones durante 12 años, y se contabilizó un total de 308 casos de arteriopatía periférica. En la regresión de Cox se encontró una HR de 0,79 (IC 95%: 0,64-0,96) por cada incremento de 400 µg/día de ingesta de ácido fólico. Como la HR es inferior a la unidad, se observa que la ingesta de ácido fólico puede ser un factor protector frente a la arteriopatía periférica. Esta cantidad, 400 µg, tiene una aplicación práctica en la clínica, porque es una dosis frecuente para administrar el ácido fólico. Los autores de este estudio hicieron bien en cambiar la escala de medida del ácido fólico para que, en vez de ir de 1 µg en 1 µg, se contabilizase de 400 en 400. Como interpretación se indicaría que por cada incremento de 400 µg en la ingesta de ácido fólico se reduce el riesgo de arteriopatía periférica en un 21% en *términos relativos*.

## 14.9. REGRESIÓN DE COX CON UNA VARIABLE INDEPENDIENTE CON >2 CATEGORÍAS

Las variables cualitativas con más de dos categorías también se pueden introducir en un modelo de regresión de Cox. En este caso, si la variable inicial tiene *k* categorías, hay que establecer una categoría de referencia y el resto de categorías se comparan con ella. En el modelo se introducen *k* – 1 variables *dummy*. La *hazard ratio* que se obtiene para cada una de estas variables *dummy* compara el *hazard* de cada categoría con la categoría de referencia, y tiene en cuenta todo el período de seguimiento.

## 14.10. INTERPRETACIÓN DE COEFICIENTES DE VARIABLES *DUMMY*

En la tabla 14.2 y en la figura 14.6 se muestran los resultados de un estudio sobre el riesgo de desarrollar arteriopatía periférica en relación con el hábito tabáquico. El tabaco es una variable independiente con tres categorías posibles: no fumador, fumador en la actualidad y exfumador.

En la tabla 14.2 y en la figura 14.6 se representa la salida de la regresión de Cox realizada con STATA. En el recuadro inferior se encuentra la HR para las categorías de fumador actual (3,69) y exfumador (2,23). Estos valores se obtienen al comparar el *hazard* instantáneo de estas categorías con la de referencia (no fumadores, en este ejemplo).

La salida de STATA también muestra los valores *z* del test de Wald que se obtienen dividiendo los coeficientes de regresión (1,31 para fumadores y 0,80 para exfumadores) entre el error estándar correspondiente (0,26 en ambos casos).

## 14.11. REGRESIÓN DE COX CON MÚLTIPLES VARIABLES INDEPENDIENTES

En el apartado 14.3 se presentó la ecuación de la regresión de Cox con un número indeterminado de covariables o variables independientes:

*Tabla 14.2 Regresión de Cox*

| | $b_1$ | EE($b_1$) | *z* (valor *p*) | HR | IC 95% HR |
|---|---|---|---|---|---|
| **No fumadores** | | | | | **Ref.** |
| Fumador actual | 1,307 | 0,257 | 5,09 (< 0,001) | 3,69 | 2,23-6,11 |
| Exfumador | 0,802 | 0,257 | 3,12 (0,002) | 2,23 | 1,35-3,69 |

| _t | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| fum | | | | | | |
| 1 | 1.31 | 0.26 | 5.09 | 0.000 | 0.80 | 1.81 |
| 2 | 0.80 | 0.26 | 3.12 | 0.002 | 0.30 | 1.31 |

Fumadores actuales

Exfumadores

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| fum | | | | | | |
| 1 | 3.69 | 0.95 | 5.09 | 0.000 | 2.23 | 6.11 |
| 2 | 2.23 | 0.57 | 3.12 | 0.002 | 1.35 | 3.69 |

**Figura 14.6** Regresión de Cox con una variable independiente de tres categorías realizada con STATA.

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + b_1 x_1 + \ldots + b_p x_p$$

Los modelos multivariantes son los que habitualmente se presentan en las publicaciones científicas. Es lógico, ya que la probabilidad de que ocurra un evento (muerte, curación, recidiva, etc.) en un período de tiempo dependerá casi siempre de múltiples factores. Así ocurre, por ejemplo, con la arteriopatía periférica, una enfermedad del sistema circulatorio que, además de asociarse al tabaco, se ha vinculado también con otros múltiples factores de riesgo, como diabetes, hipertensión o dislipemia.

## 14.12. CONTROL DE LA CONFUSIÓN EN LA REGRESIÓN DE COX

La gran ventaja que presenta la regresión de Cox es que permite obtener las estimaciones *ajustadas* por factores de confusión. Esto supone aplicar el procedimiento multivariante explicado en el apartado anterior. En el análisis no solo se incluye la exposición principal, sino también todos los posibles factores de confusión como otras tantas variables independientes. Esta actuación hace posible estimar el efecto de una exposición *ante la igualdad de los otros posibles factores* causales del resultado o evento que se valora y, por tanto, permite pasar de una simple asociación a una relación causa-efecto, pues apunta más hacia una explicación verdaderamente causal de la exposición en estudio sobre el desenlace o efecto.

Esta situación se aplica perfectamente al ejemplo de la arteriopatía periférica. Supóngase que se desea comprobar la eficacia de seguir una dieta mediterránea para prevenir la arteriopatía periférica (4). Para ello se lleva a cabo un ensayo aleatorizado donde se compara un grupo de intervención, al que se le recomienda que siga una dieta mediterránea, y un grupo control. Si se encuentra una HR protectora (HR < 1), cabría afirmar que la dieta mediterránea podría ayudar en la prevención de esta enfermedad. Pero ¿se puede estar seguro de que el resultado se debe a la dieta y no a otras características asociadas al cumplimiento de esa dieta? Podrían existir factores relacionados con la arteriopatía periférica que se distribuyeran de manera desigual entre el grupo de intervención (asignado al azar de ser educado en dieta mediterránea) y el grupo control (asignado al azar de seguir una dieta baja en grasa, en este ejemplo). Por ejemplo, en el grupo control podría haber más fumadores.

En la figura 14.7 se presentan tres modelos de regresión de Cox, un modelo univariante en el que solo se introduce la variable de exposición (dieta mediterránea), el modelo multivariante en el que se ajusta por tabaco, y un segundo modelo multivariante en el que también se tiene en cuenta la edad y el sexo de los participantes.

Se observa que el valor de la HR para la dieta mediterránea en comparación con el grupo control se mantiene constante en los tres modelos. Era esperable, al haber usado un diseño con reparto o asignación al azar a una u otra dieta. Se confirma así que es muy verosímil que exista una relación

**Modelo univariante**

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | 0.42 | 0.09 | -4.12 | 0.000 | 0.28 | 0.63 |

Variable de confusión: tabaco

**Modelo multivariante 1**

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | 0.41 | 0.09 | -4.18 | 0.000 | 0.27 | 0.62 |
| fum | | | | | | |
| 1 | 3.72 | 0.95 | 5.11 | 0.000 | 2.25 | 6.15 |
| 2 | 2.28 | 0.59 | 3.20 | 0.001 | 1.38 | 3.77 |

Variables de confusión: tabaco, edad, sexo

**Modelo multivariante 2**

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | 0.41 | 0.09 | -4.18 | 0.000 | 0.27 | 0.62 |
| fum | | | | | | |
| 1 | 2.66 | 0.83 | 3.16 | 0.002 | 1.45 | 4.89 |
| 2 | 1.42 | 0.44 | 1.14 | 0.256 | 0.77 | 2.62 |
| edad0 | 1.05 | 0.02 | 2.82 | 0.005 | 1.01 | 1.09 |
| sexo | 0.43 | 0.13 | -2.88 | 0.004 | 0.25 | 0.77 |

**Figura 14.7** Modelos de regresión de Cox para comprobar el efecto protector de la dieta mediterránea sobre la arteriopatía periférica.

causa-efecto y que la dieta mediterránea a igualdad de esos otros factores ejerza un efecto protector sobre la arteriopatía periférica. También se observa que los valores de las HR para el tabaco (modelo multivariante 1) disminuyen tras un ajuste por sexo y edad en el modelo multivariante 2. Esto demuestra que el sexo y la edad de los participantes eran factores de confusión que explican parte del riesgo que se había atribuido erróneamente al tabaco en el primer modelo multivariante. El tabaco es un factor de riesgo, pero con una HR no tan alta como la que mostraba el modelo anterior (además, desaparece la significación estadística para los exfumadores).

## 14.13. INTERVALOS DE CONFIANZA PARA LA *HAZARD RATIO* EN EL MODELO DE REGRESIÓN DE COX

El intervalo de confianza de la HR se obtiene a partir de la siguiente fórmula:

$$IC95\% = e^{b_i \pm 1{,}96 * EE(b_i)}$$

Siguiendo el ejemplo ridículamente sencillo y los resultados de la figura 14.3, el intervalo de confianza de la HR se calcularía como:

$$IC95\%\ \text{inferior} = e^{1{,}387 - 1{,}96 * 0{,}645} = 1{,}1$$

$$IC95\%\ \text{superior} = e^{1{,}387 + 1{,}96 * 0{,}645} = 14{,}2$$

De este modo, puede afirmarse que la HR de la población se encuentra con un 95% de confianza entre los valores 1,1 y 14,2.

## 14.14. INTERACCIÓN (MODIFICACIÓN DEL EFECTO) EN REGRESIÓN DE COX Y TEST DE RAZÓN DE VEROSIMILITUD

En ocasiones, el efecto de una variable independiente sobre el evento (mortalidad, curación...) puede aumentar o disminuir por la interacción con otra variable. Por ejemplo, podría plantearse si el efecto protector de la dieta mediterránea aumenta o disminuye con la edad de los participantes. Dicho de otro modo, si la edad modifica el efecto protector de la dieta o, lo que es lo mismo, si existe una interacción dieta-edad. Para resolver esta cuestión se comparan dos modelos de regresión de Cox. En uno se introducen solo las covariables (dieta y edad) y en otro se añade, además, el término de la interacción. Este término es el producto de las dos variables independientes. Para concluir si existe interacción, se calcula la diferencia de verosimilitud entre ambos modelos y se aplica el test de razón de verosimilitudes. Se obtiene la prueba de la razón de verosimilitud (LR ji cuadrado) y el valor *p* de significación estadística. Cuando la *p* sea inferior a 0,05, podrá decirse que probablemente existe interacción. Cuando no se encuentre significación en el LR test, es posible que exista interacción, pero muchas veces faltarían evidencias (poca potencia) para demostrarla.

En la figura 14.8 se muestra la salida del análisis realizado con STATA. La prueba de la razón de verosimilitud es ji cuadrado = 1,83, y esto, con un grado de libertad, se corresponde con un valor *p* superior a 0,05. Por tanto, no hay pruebas en estos datos para demostrar una interacción significativa entre dieta y edad.

## 14.15. INTERPRETACIÓN DEL RIESGO BASAL *(BASELINE HAZARD)*

Como se ha mencionado anteriormente, y volviendo a la ecuación, en el modelo de regresión de Cox no existe una ordenada en el origen de carácter constante, como en otros modelos de regresión. Cuando todas las covariables valen 0, se obtiene el riesgo basal *($\lambda_{0t}$ baseline hazard)*. Se llama riesgo basal aunque no corresponde al riesgo cuando el tiempo es 0, sino que equivale al riesgo cuando todas las covariables tienen un valor exactamente igual a 0. Por consiguiente, es una función que cambia con el tiempo y que, a veces, corresponderá a supuestos absurdos (p. ej., nadie puede tener un peso corporal = 0 kg).

A veces sí es interpretable en términos de la vida real. Por ejemplo, si se obtuviese una ecuación que contiene el número de cigarrillos como variable cuantitativa y el sexo (hombre = 0, mujer = 1):

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + b_1\text{sexo} + b_2\text{cigarrillos}$$

El riesgo basal sería el riesgo que tiene un hombre que fuma 0 cigarrillos. Como se puede entender, este riesgo irá cambiando con el tiempo.

Como ocurre en otros modelos de regresión, el riesgo basal no será válido fuera de los límites de lo observado en el estudio ni cuando corresponda a valores implausibles.

## 14.16. REGRESIÓN DE COX ESTRATIFICADA

La regresión de Cox permite estimar las *hazard ratios* estratificando por distintas variables (hasta un máximo de cinco, en STATA). Volviendo al ejemplo ridículamente sencillo, supóngase que se desea tener en cuenta el sexo además del tabaco. Esta variable puede ser un factor de confusión, porque el sexo se asocia con el tabaco (p. ej., que los hombres fumen más) y con la mortalidad (los hombres mueren antes que las mujeres). Una posibilidad es introducir el sexo en el modelo de regresión de Cox:

$$\ln(\lambda_t) = \ln(\lambda_{0t}) + b_1x_1 + b_2x_2$$

$x_1$ = 0 para no fumadores, $x_1$ = 1 para fumadores.

$x_2$ = 0 para hombres, $x_2$ = 1 para mujeres.

**Modelo de regresión sin el término de interacción**

```
        failure _d:  PA_conf == 1
  analysis time _t:  foll_ap

Iteration 0:   log likelihood = -741.31699
Iteration 1:   log likelihood = -717.06057
Iteration 2:   log likelihood = -716.91125
Iteration 3:   log likelihood = -716.91125
Refining estimates:
Iteration 0:   log likelihood = -716.91125

Cox regression -- Breslow method for ties

No. of subjects =         7225                 Number of obs    =      7225
No. of failures =           89
Time at risk    =  31800.79945
                                               LR chi2(3)       =     48.81
Log likelihood  =   -716.91125                 Prob > chi2      =    0.0000
```

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | .4053385 | .0861458 | -4.25 | 0.000 | .2672475 | .6147834 |
| edad0 | 1.040185 | .0175724 | 2.33 | 0.020 | 1.006308 | 1.075203 |
| sexo | .3007579 | .0692775 | -5.22 | 0.000 | .1914912 | .4723733 |

```
. est store A

. capt drop interact

. g interact = edad0*sexo

. stcox MedDiet edad0 sexo interact
```

**Modelo de regresión con el término de interacción**

```
        failure _d:  PA_conf == 1
  analysis time _t:  foll_ap

Iteration 0:   log likelihood = -741.31699
Iteration 1:   log likelihood = -717.10664
Iteration 2:   log likelihood = -716.00017
Iteration 3:   log likelihood = -715.99851
Refining estimates:
Iteration 0:   log likelihood = -715.99851

Cox regression -- Breslow method for ties

No. of subjects =         7225                 Number of obs    =      7225
No. of failures =           89
Time at risk    =  31800.79945
                                               LR chi2(4)       =     50.64
Log likelihood  =   -715.99851                 Prob > chi2      =    0.0000
```

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | .4073264 | .0866222 | -4.22 | 0.000 | .2684886 | .6179586 |
| edad0 | 1.025796 | .0202847 | 1.29 | 0.198 | .986799 | 1.066333 |
| sexo | .0087433 | .023079 | -1.80 | 0.073 | .0000495 | 1.543549 |
| interact | 1.052476 | .0397807 | 1.35 | 0.176 | .9773253 | 1.133405 |

```
. lrtest A

Likelihood-ratio test                          LR chi2(1)  =      1.83
(Assumption: A nested in .)                    Prob > chi2 =    0.1767
```

**Figura 14.8** Comprobación de la interacción dieta-edad en una regresión de Cox.

| _t | Haz. Ratio | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MedDiet | 0.41 | 0.09 | -4.17 | 0.000 | 0.27 | 0.62 |
| fum | | | | | | |
| 1 | 2.48 | 0.79 | 2.86 | 0.004 | 1.33 | 4.63 |
| 2 | 1.53 | 0.48 | 1.36 | 0.174 | 0.83 | 2.83 |
| edad0 | 1.05 | 0.02 | 2.96 | 0.003 | 1.02 | 1.09 |
| sexo | 0.44 | 0.13 | -2.78 | 0.005 | 0.25 | 0.78 |

Stratified by nodo

**Figura 14.9** Modelos de regresión de Cox estratificada.

Este modelo asume que el *hazard* asociado al tabaco entre hombres y mujeres es proporcional. Podría ocurrir que no fuese así, y que, en realidad, el riesgo tenga una evolución diferente para hombres y mujeres.

Un modo de solventar este problema consistiría en hacer dos regresiones separadas, para hombres y para mujeres. Sin embargo, se obtendrían así dos estimaciones del riesgo asociado al tabaco, una para cada sexo; cada una de ellas contendría menos sujetos y menos desenlaces, y la potencia estadística sería baja. Tal vez se desee obtener un indicador del riesgo general y, por tanto, esta opción no sirva.

Existe una tercera alternativa que consiste en estratificar según la variable sexo dentro de un único modelo de regresión de Cox. Esta opción permite asumir que el riesgo cambie de manera distinta entre hombres y mujeres. El cálculo de la *hazard ratio* para el tabaco se realiza estratificando por sexo y, al final, se obtiene un único indicador del riesgo donde se ha ajustado muy bien por sexo.

Con esta última alternativa disminuye el número de grados de libertad en el test de máxima verosimilitud. El inconveniente es que no se obtiene el valor de las *hazard ratios* para la variable sexo. Sin embargo, no supone un problema cuando la variable por la que se estratifica es un factor ya conocido de sobra como fuertemente asociado al desenlace y para el que no resultaría relevante estimar y publicar una asociación, porque tal asociación no representaría ningún descubrimiento. Sucede lo mismo con factores intrínsecos y particulares de un estudio, como, por ejemplo, el centro de reclutamiento de participantes en estudios multicéntricos. No resulta interesante valorar el *efecto* de un centro comparado con otro, aunque debe ser un factor que se controle fuerte y eficazmente.

En la figura 14.9 puede verse otro ejemplo de estratificación. En este caso se estratificó por centro de reclutamiento en el estudio de la arteriopatía periférica. Como se ha explicado, no se obtiene una estimación para esta variable, aunque no importa, porque no es un factor relevante para la ciencia. Sin embargo, es necesario tenerlo en cuenta, porque en la estimación del riesgo podrían influir las características de cada nodo o centro de reclutamiento. En la salida de STATA se indica a pie de tabla la variable por la cual se obtiene la estratificación.

## 14.17. TIEMPO DE SEGUIMIENTO EN LA REGRESIÓN DE COX

Como se ha visto, la originalidad de la regresión de Cox reside en tener en cuenta el tiempo que tarda en producirse el evento y en considerar todo el período de seguimiento en estudios longitudinales. Es preferible usar regresión de Cox en vez de regresión logística cuando los períodos de observación de cada sujeto son variables. Por ejemplo, no tendría sentido utilizar como desenlace la supervivencia solo a 3 años como variable dicotómica (sobrevivió/no sobrevivió) y aplicar después una regresión logística. Así se perdería mucha información. Por ejemplo, al utilizar una regresión logística, en vez de usar una de Cox, se situaría erróneamente en el extremo de *óptimo* pronóstico al paciente que falleció a los 3,01 años, y en el extremo *pésimo* al que murió a los 2,99 años, cuando en realidad su pronóstico era básicamente el mismo.

Además, con la regresión de Cox se dispone ventajosamente de varios puntos de referencia para medir el momento en el que cada participante empieza a estar en riesgo:

- *Tiempo de seguimiento en el estudio.* En este caso, el contador de tiempo empieza cuando el participante *entra* en el estudio. Para todos los sujetos se considera un tiempo inicial igual a 0 desde el momento en el que empiezan a participar en el estudio.
- *Tiempo desde que se realizó el diagnóstico de la enfermedad.* Este tiempo puede ser *anterior* al momento de entrar en el estudio. En muchos casos será interesante considerar este tiempo, porque el inicio de la enfermedad marca el comienzo de estar en riesgo. El tiempo 0 para cada participante no es el inicio del estudio, sino que señala el momento en que se obtuvo el diagnóstico. En STATA se utiliza la opción *origin* para indicar la fecha de diagnóstico o cualquier otra fecha para la que se considere que marca el comienzo de estar en riesgo.
- *Tiempo desde el nacimiento.* La *edad* de los participantes es otro factor temporal que puede ser decisivo para estimar el riesgo. Si se valora el riesgo de fallecer, será muy distinto si se ha nacido en 1920 o en 1994. El tiempo inicial no es 0, sino que a cada participante se le asocia la edad antes del diagnóstico. En STATA se utiliza la opción *origin* para señalar la fecha de nacimiento cuando se desea hacer un ajuste fino por edad. A su vez, la opción *enter* distingue el momento del diagnóstico o la entrada en un estudio.

## 14.18. REGRESIÓN DE COX CON COVARIABLES DEPENDIENTES DEL TIEMPO

Hasta ahora se ha mantenido el supuesto de asumir que el riesgo asociado a un factor de exposición se mantiene constante a lo largo del tiempo. Sin embargo, con frecuencia, en medicina no ocurre así. Cuando el tiempo de exposición es largo, es probable que se produzca un cambio en la exposición (aumenta o disminuye el consumo de tabaco). En tales casos conviene realizar una regresión de Cox *con exposición cambiante.* El tiempo de exposición introducido en el modelo se ajustará según la información disponible para cada participante sobre el tiempo durante el cual ha estado expuesto.

También puede suceder que, a partir de un tiempo *t*, tenga lugar un cambio en el riesgo asociado a una variable de exposición (niveles de calcio, tensión arterial, etc.). Supóngase, por ejemplo, que se considera que se produce un cambio mayor después de estar fumando al menos durante 5 años. El tiempo de exposición (más de 5 años) provocará una modificación del efecto del tabaco, una situación que recuerda, por analogía, el concepto de interacción.

En estos casos es preciso introducir una covariable dependiente del tiempo en el modelo de regresión de Cox. En STATA se utiliza la opción *tvc (time-varying covariate)* para indicar la variable dependiente del tiempo y *texp* para señalar el tiempo en el que se produce el cambio en el nivel de riesgo (**5-6**). Este tipo de análisis consiste en fragmentar el tiempo de seguimiento de los participantes para tener en cuenta esos cambios en el riesgo.

## 14.19. MODELOS DE TIEMPOS DE FALLO ACELERADOS

Aunque el modelo de regresión de Cox se ha utilizado y se sigue empleando a menudo en análisis de supervivencia, el supuesto o hipótesis de riesgos proporcionales con el mantenimiento continuo de esta proporcionalidad a lo largo del tiempo es, con frecuencia, demasiado restrictivo y limita el modelo a unas cuantas distribuciones que lo satisfacen. Esta hipótesis no se cumple cuando el efecto de una covariable cambia con el tiempo, algo frecuente. Aunque la regresión de Cox estratificada relaja esta hipótesis, solo es aplicable a variables cualitativas que, a veces, no son las de mayor interés para el estudio. La inclusión de variables dependientes del tiempo es una solución más eficiente. Por otro lado, se suele aplicar la verosimilitud parcial para aproximar los estimadores máximo-verosímiles, lo que exige, además, el cumplimiento de nuevas hipótesis.

Además, se trata de un procedimiento semiparamétrico, en el que el riesgo basal se estima por separado de modo no paramétrico. En ocasiones se utiliza también la modelización paramétrica del riesgo basal. En el campo de la ingeniería se han usado desde hace tiempo *modelos de tiempos de fallo acelerados*, que ofrecen una alternativa a los de riesgos proporcionales. Una de sus ventajas es que modelizan directamente la función de supervivencia, en lugar del riesgo, lo que resulta más intuitivo y fácilmente interpretable. Cada vez es más frecuente su empleo en medicina. Kay y Kinnersley utilizaron un modelo de este tipo en un estudio sobre la gripe (7). Esta solución es especialmente interesante cuando se busca acelerar, o también retrasar, como sería más apropiado en estudios clínicos, la ocurrencia del evento. Lambert et al. (2004) aplicaron un modelo con efectos aleatorios a datos de supervivencia en el trasplante de riñón (8). Los modelos de este tipo admiten el uso de una gran diversidad de funciones de distribución de probabilidad para la supervivencia, como son la exponencial, de Weibull, log-logística, log-normal o gamma, por citar algunas de las más frecuentes.

## 14.20. RELACIÓN ENTRE *HAZARD RATIO* Y RAZÓN DE DENSIDADES DE INCIDENCIA

Aunque su interpretación pueda ser parecida, el riesgo relativo (RR) y la *hazard ratio* (HR) no deben confundirse (3). El RR no tiene en cuenta el tiempo en el que se producen los sucesos (incidencia en expuestos/incidencia en no expuestos). En cambio, el tiempo es absolutamente decisivo para la HR. Por lo tanto, ambos estimadores pueden proporcionar resultados diferentes. El RR y la HR únicamente serían equivalentes si el tiempo de seguimiento coincidiera exactamente en todos los sujetos de la muestra, lo cual es muy inusual, ya que es difícil que el evento de interés se dé a la vez en todos los sujetos.

En general, el RR será siempre el más cercano a la unidad, y la HR está más alejada (aunque no tanto como la *odds ratio* [OR]). La divergencia entre RR y HR es directamente proporcional al tiempo de seguimiento, la incidencia de eventos y el RR (las diferencias serán importantes cuando el RR sea superior a 2,5). Sabiendo esto, se podrían hacer combinables RR y HR en un metaanálisis y para calcular los riesgos atribuibles y el número de pacientes que es necesario tratar.

## 14.21. SIMILITUDES Y DIFERENCIAS ENTRE REGRESIÓN DE COX Y REGRESIÓN LOGÍSTICA

La regresión de Cox presenta numerosas similitudes con la regresión logística, aunque hay algunas diferencias importantes a la hora de aplicarla. Entre las *semejanzas* destacan:

- Tanto en la regresión de Cox como en la regresión logística, la medida de asociación entre la exposición y el efecto es un cociente. Por lo tanto, ambos modelos trabajan en escala *multiplicativa* y no en escala aditiva. Este cociente se obtiene al elevar el número e al coeficiente de la regresión.
- Otra similitud procede de que estas dos regresiones utilizan los métodos de *máxima verosimilitud* para calcular los parámetros. Además, en ambas se puede aplicar como test de hipótesis para comparar modelos el test de razón de verosimilitudes, que sigue una ji cuadrado con tantos grados de libertad como variables de más existan en el modelo más amplio con respecto al más reducido.
- Las dos regresiones permiten estudiar variables independientes cuantitativas, aunque exigen que se compruebe si la forma de su relación con el efecto sigue un patrón monotónico (en caso contrario, habría que introducir categorización). Además, permiten introducir en el modelo variables independientes cualitativas, con la creación de variables *dummy* o indicadoras. Mediante el uso de modelos multivariables, con ambas regresiones se puede controlar de manera semejante la confusión (añadiendo la variable en el modelo) y valorar la interacción (añadiendo el término producto en el modelo).

- Ambas regresiones permiten la aplicación de algoritmos de selección de variables (*stepwise*, etc.), aunque, en general, no están muy aconsejadas para la selección automática de posibles confusores. Especialmente en epidemiología, la selección debe tener en cuenta prioritariamente el conocimiento experto de la materia y las relaciones causales entre variables (9).
- Por último, los dos modelos requieren aproximadamente que hayan ocurrido unos 10 eventos por cada valor de cada variable independiente que se desee introducir.

Entre las *diferencias,* pueden destacarse:

- La principal diferencia reside en la variable dependiente. Mientras que en la regresión logística es de tipo cualitativo dicotómico (evento), en la de Cox es un conglomerado de dos variables: una cualitativa dicotómica (evento) y una cuantitativa (tiempo de seguimiento).
- La medida de asociación que se obtiene es la *hazard ratio* en la regresión de Cox y la OR en la logística. Aun siendo medidas diferentes, ambas pueden estimar el riesgo relativo, aunque con cierta sobreestimación, mayor en el caso de la OR. En cuanto a la ecuación del modelo, en la regresión de Cox no hay ordenada en el origen, al contrario que en la regresión logística.
- La regresión de Cox permite un ajuste alternativo más fino de algunas variables confusoras, aunque no cumplan la proporcionalidad de los *hazards* en el tiempo, con un análisis estratificado (10).
- Existen modelos de regresión de Cox en los que pueden incluirse covariables dependientes del tiempo, lo cual otorga a estos modelos cierta flexibilidad.
- La regresión de Cox se puede emplear en estudios de casos y controles emparejados para sustituir a la regresión logística condicional. Se forman estratos por la pareja (o relación caso-control), se crea un valor de seudotiempo que sea superior en los controles que en los casos y, como evento, se usa la categoría correspondiente a los casos. Esto permite preservar el emparejamiento (al introducirlo en los estratos) y es equivalente a la regresión logística condicional. Esta es la solución aplicada en SPSS para analizar estudios de casos y controles emparejados, ya que SPSS carece de regresión logística condicional.

## 14.22. POSIBILIDADES Y OPCIONES DE LA REGRESIÓN DE COX CON STATA

La familia de instrucciones para el análisis de supervivencia en STATA se llama **st** *(survival time).* En primer lugar, y de manera imprescindible, hay que indicar a STATA con la orden **stset** cuáles son las dos variables clave:

- La que recoge el tiempo hasta el evento.
- La que corresponde al evento.

Por tanto, esta orden **stset** es fundamental y debe ser la primera que se introduce. Si se omite, lo demás no funcionará. La estructura general más básica de **stset** debería ser:

```
stset tiempo, failure(evento==1)
```

Hay que introducir la variable tiempo y, como opción, indicar la variable del evento y el valor que corresponde a los que sí tienen el evento (si no se indica, por defecto, entiende evento==1). Tal vez en lugar de tener una sola variable de tipo tiempo se disponga de dos variables en formato fecha que recojan la entrada y salida del estudio. En este caso, la instrucción quedaría de la siguiente manera:

```
stset fecha_fin, f(evento==1) enter(fecha_inicio) ///
scale(365.25)
```

Esta forma ahorra la necesidad de calcular previamente la variable que recoge la duración del tiempo de seguimiento. Por defecto, cuando se introducen las dos fechas, STATA calcula el tiempo de seguimiento en días desde una fecha hasta la otra. Si se desea el resultado en otra unidad, habría que indicarlo con la opción **scale.** En el ejemplo, se introduce 365.25, para que el tiempo de seguimiento se exprese en años (las cifras decimales, .25, tienen en cuenta los años bisiestos).

En muchas ocasiones es la edad, y no la entrada en el estudio, la que establece fundamentalmente el tiempo en riesgo de una persona. Esto afectará en mayor medida a los desenlaces que tengan mayor relación con la edad. Una aproximación para solucionar este problema es el ajuste multivariable. Otra aproximación complementaria consistiría en establecer cuál es la fecha en la que empieza a estar en riesgo una persona cuando esta sea distinta de la de entrada en el estudio (generalmente, la fecha de nacimiento), con la opción **origin.**

```
stset fecha_fin , failure(evento==1) origin(fecha_nac) ///
    enter(fecha_inicio) scale(365.25)
```

Por último, si hay más de una medida por sujeto (es decir, cada sujeto aporta varias filas a la base de datos), se indicará con la opción **id** y, entre paréntesis, la variable que identifica a cada sujeto. La base de datos tiene que estar en formato alargado (cada fila corresponde a una observación).

```
stset fecha_fin, f(evento==1) id(id) o(fecha_nac) ///
en(fecha_inicio) sc(365.25)
```

Una vez establecidas las bases del análisis de supervivencia, pueden describirse los datos con las instrucciones **stdescribe, sts list** y **stsum.** Es interesante utilizar **stsum,** seguido de la opción **by,** para obtener el seguimiento total en personas-año y la tasa de incidencia de cada categoría de la exposición.

```
stsum, by(exp_cat)
```

La instrucción **sts gen var=s** genera una nueva variable que contiene las estimaciones de supervivencia acumulada, y con **sts graph** se representarían las curvas de Kaplan-Meier. Es interesante añadir la opción **by** cuando se quieran representar varias curvas. Además, permite la opción **cumhaz** (equivalente a **na**) para representar curvas de riesgo acumulado y no de supervivencia. Estas curvas de incidencia acumulada suben con el tiempo, en vez de bajar, como hacen las de supervivencia, y son conocidas como curvas de Nelson-Aalen. Estas representaciones gráficas y otras se pueden encontrar también en la instrucción **stcurve.**

Cuando existen datos repetidos (y se incluye en **stset** la opción **id**), para describir tiempos de seguimiento es posible usar la instrucción **stptime.** También es interesante ver el patrón que siguen las covariables, en cuanto a si cambian a lo largo del tiempo o si están perdidas. Para ello se utiliza la instrucción **stvary.** Cuando se trata de un diseño de medidas repetidas, tal vez interese rellenar las covariables que están perdidas, lo cual se puede hacer automáticamente con la instrucción **stfill,** a la que hay que añadir la opción **baseline** o **forward** en función de si se debe rellenar el valor perdido con la información basal o la última disponible.

La fase final del análisis correspondería al ajuste de la regresión de Cox. La instrucción correspondiente es **stcox.** Como STATA ya ha almacenado las variables que hacen referencia al tiempo de seguimiento y al evento, únicamente habría que añadir, detrás de **stcox,** la lista de variables independientes (indicando con ***i.*** delante de la variable si se trata de una variable cualitativa). Esta instrucción sería la misma si se tratase de un diseño con medidas repetidas.

```
stcox indep1 indep2
```

Esta instrucción devolverá los resultados en términos de *hazard ratios*. Sin embargo, si se considera interesante obtenerlos en coeficientes, basta con añadir la opción **nohr.**

Además de introducir en el modelo todas las variables de ajuste que sean necesarias, STATA permite realizar un ajuste más fino (en el que se podrían tener en cuenta relaciones no lineales de la variable de ajuste con el desenlace) mediante el análisis por estratos para obtener un único resultado. Se aplicaría la opción **strata,** seguida entre paréntesis de la variable cualitativa en cuyas categorías se estratificará el análisis.

Otra opción disponible es **vce,** que sirve para cambiar los estimadores de la varianza. Se pueden pedir estimadores robustos, con **vce(robust),** o estimadores calculados con métodos de remuestreo (v. capítulo 19), con **vce(jackknife)** o **vce(bootstrap).**

Para su validez, los modelos de regresión de Cox requieren que los riesgos sean proporcionales a lo largo del tiempo. En STATA existen tres instrucciones con las cuales se puede comprobar este requisito: **stphplot, stcoxkm** y **estat phtest.** Las dos primeras son métodos gráficos y la tercera es un test de hipótesis basado en el método de los residuales de Schoenfeld (**2,6,10,11**).

Por último, un aspecto interesante es el cálculo de tamaño muestral para este tipo de estudios. En este ejemplo se piden los distintos tamaños muestrales que serían necesarios para observar distintos escenarios de HR, desde 0,1 hasta 0,9 (en saltos de 0,2 en 0,2), para potencias del 80 y el 90%:

```
stpower cox, hratio(0.1(0.2)0.9) power(0.8 0.9) hr
```

STATA devolverá el número de personas que han de incluirse y el de eventos que se espera observar. Si se conoce o puede estimarse la probabilidad del evento, podría añadirse la opción **failprob(#)** para tenerla en cuenta. Con esta instrucción puede averiguarse, asimismo, la potencia del estudio, añadiendo como opción el tamaño de muestra (**n**) en vez de la opción **power.**

## 14.23. RESUMEN DE LAS INSTRUCCIONES EN STATA, SPSS Y R

| Concepto | Instrucción | Ejemplo |
|---|---|---|
| **STATA** | | |
| Establecer las bases | stset | stset seguimiento, f(evento==1) |
| Describir | stsum<br>stdescribe | stsum, by(cat_exposicion) |
| Kaplan-Meier | st graph | st graph, by(cat_exposicion) |
| Regresión de Cox | stcox | stcox i.cat_exposicion, strata (estrato) |
| **SPSS** | | |
| Regresión de Cox | COXREG seguimiento<br>/STATUS=evento(1)<br>/STRATA=estrato<br>/print=CI<br>/METHOD=ENTER cat_<br>exposicion. | |
| **R** | | |
| Regresión de Cox | library(survival)<br>summary(coxph(Surv(seguimiento,evento)~cat_exp + strata(estrato),data=) | |

## REFERENCIAS


1. Katz MH. Multivariable analysis. A practical guide for clinicians. 2nd ed. Cambridge: Cambridge University Press; 2006.
2. Hosmer DW, Lemeshow S, May S. Applied survival analysis. Regression modelling of time-to-event data. 2nd ed. Hoboken: John Wiley & Sons; 2008.
3. Martínez-González MA, Alonso A, López Fidalgo J. ¿Qué es una hazard ratio? Nociones de análisis de supervivencia. Med Clin (Barc) 2008;131:65-72.
4. Merchant AT, Hu FB, Spiegelman D, Willet WC, Rimm EB, Ascherio A. The use of B vitamin supplements and peripheral arterial disease risk in men are inversely relate. J Nutr 2003;133:2863-7.
5. Ruiz-Canela M, Estruch R, Corella D, Salas-Salvadó J, Martínez-González MA. Mediterranean diet inversely associated to Peripheral Artery Disease: the PREDIMED randomized trial. JAMA 2014;311(4):415-7
6. Cleves M, Gould W, Gutiérrez R, Marchenko Y. An introduction to survival analysis using Stata. 2nd ed. College Station: Stata Press Publication; 2008.
7. Kay R, Kinnersley N. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. Drug Inf J 2002;36:571-9.
8. Lambert P, Collett D, Kimber A, Johnson R. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. Stat Med 2004;23:3177-92.
9. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155(2):176-84.
10. Kleinbaum DG. Survival analysis: a self-learning text. New York: Springer-Verlag; 1995.
11. Collet D. Modelling Survival Data in Medical Research. 2nd ed. Boca Raton: CRC Press; 2009.

# ANÁLISIS DE CONCORDANCIA, VALIDEZ Y PRONÓSTICO

15

*M. Á. Martínez-González, E. Toledo, A. Sánchez-Villegas*

## 15.1. CONCEPTOS Y DEFINICIONES

### 15.1.1. Validez

La *validez* de una medición es su capacidad de lograr la identidad exacta con la verdad que se pretende o se afirma medir. Un procedimiento es válido en la medida en que captura la realidad con exactitud. El *sesgo* (error sistemático) es una alteración de la validez de una medición, y consiste en una tendencia permanente a alejarse de la verdad. Un sesgo concreto consistentemente tenderá a desviarse en la misma dirección en cada medición. La validez exige un requisito previo: la reproducibilidad.

### 15.1.2. Fiabilidad (reproducibilidad)

La *fiabilidad* o *reproducibilidad* es la capacidad de poder obtener un mismo valor cuando una medición se repite en la misma muestra. Hay dos tipos de *reproducibilidad:*

1. Si se utiliza un instrumento de medida varias veces por el mismo observador en la misma muestra: reproducibilidad *intraobservador.*
2. Se valoran distintas mediciones realizadas en la misma muestra, pero por procedimientos o personas diferentes: reproducibilidad *entre observadores.*

Los términos *consistencia, concordancia y acuerdo* parecen sinónimos de reproducibilidad, pero, cuando se trata de variables cuantitativas, se debe matizar una diferencia:

- La *consistencia* consiste en mantener el mismo orden o posición *(ranking)* de las observaciones al medirlas dos o más veces. Es decir, las distintas series de mediciones realizadas sobre la misma muestra ordenarán a los sujetos u observaciones de igual manera.
- El *acuerdo o concordancia* consiste en que los valores obtenidos en las distintas mediciones son idénticos.

Por lo tanto, siempre que haya acuerdo habrá consistencia, pero no al contrario.

Por ejemplo, sobre cinco muestras sanguíneas, el colesterol total puede ser 150, 170, 190, 210 y 230 mg/dl al medirlo con un método A, y 160, 175, 202, 206 y 231 mg/dl con otro método B. Hay consistencia, ya que la posición relativa de cada medición es la misma, pero no hay acuerdo o concordancia, porque los valores no coinciden.

### 15.1.3. Precisión

La *precisión* se refiere a la ausencia de error aleatorio, no de error sistemático (v. apartado 4.1). El error aleatorio interfiere en la valoración de la reproducibilidad y de la validez (**1**).

## 15.2. CONSIDERACIONES GENERALES SOBRE ESTUDIOS DE VALIDACIÓN DE PRUEBAS DIAGNÓSTICAS

La validez es la propiedad más importante de un test o prueba. En psicología se suele diferenciar entre validez de contenido, de constructo y de criterio.

La validez de *contenido* consiste en que el test recoge una representación adecuada de los contenidos que pretende evaluar y no presenta omisiones. Suele medirse por un juicio de expertos.

El término *constructo* se refiere a conceptos teóricos que no son directamente observables o medibles (autoestima, neuroticismo, inteligencia, etc.) y que se pretenden hacer operativos mediante un instrumento de medida, como un cuestionario. Los distintos métodos estadísticos de consistencia y acuerdo, e incluso el análisis factorial, pueden aplicarse para valorar la validez del constructo.

La validez de *criterio* se refiere a que existe un criterio externo que es una variable distinta del test y que indicará de modo cierto si aquello que el test pretendía medir se ha medido realmente o no.

En el ámbito médico, el valor clínico de una nueva prueba diagnóstica depende de que contribuya a mejorar el pronóstico del paciente gracias a la información que se obtiene al aplicarla. Este es el mejor *criterio*. Debe tenerse en cuenta que las pruebas diagnósticas, especialmente si son invasivas, conllevan el riesgo de efectos adversos que se prevendrían en su totalidad si no se realizase la prueba. Debe ponderarse siempre el balance de beneficios y riesgos. En los riesgos se añadirán también los secundarios a los tratamientos aplicados cuando la prueba diagnóstica conduce a un falso positivo. Lamentablemente, suele existir en el ámbito clínico una creencia muy arraigada de que aplicar más pruebas diagnósticas será *siempre* mejor, porque permitirá un tratamiento más precoz, cuando la enfermedad esté menos avanzada, y así se mejorará el pronóstico. Ahora bien, esto no puede darse por supuesto sin demostrarlo. Debe demostrarse que el tratamiento al que conduce un resultado positivo de la prueba realmente mejorará el pronóstico de la enfermedad. Lo ideal sería validar la prueba diagnóstica mediante *ensayos clínicos* aleatorizados que demuestren que los posibles desenlaces o eventos clínicos relevantes (*end-points* y efectos adversos) se redujeron cuando el paciente fue aleatorizado a someterse a la nueva prueba diagnóstica (frente a un grupo control que siguió los cuidados habituales). Esto casi nunca se hace y se siguen usando más y más pruebas diagnósticas sin que haya demostración de sus beneficios (**2**).

Lo que sí suele hacerse son las fases 1 a 3 de la tabla 15.1 (**1-3**). La fase 4 requeriría un ensayo aleatorizado, algo que no se ha hecho casi nunca, salvo para valorar algunos cribados preventivos, como el de cáncer de mama (**4**).

Al considerar la validación de una prueba, se deben distinguir diferentes situaciones:

- Según haya o no *patrón de oro*:
  - se valorará *validez* solo cuando se disponga de un patrón de oro *(gold standard)* que otorgue certeza al diagnóstico de enfermedad o a descartarla,
  - en cambio, solo se podrá valorar *reproducibilidad* cuando se comparen entre sí dos pruebas con cierto grado de invalidez (ambas pueden ser equivocadas).
- Según se trate de sustituir o añadir:
  - es muy distinto valorar si una nueva prueba superará a otra antigua al *sustituirla,*
  - o si esa superación solo se da cuando se *añade* la prueba nueva a la que ya venía usándose.

***Tabla 15.1 Fases en el estudio de una prueba diagnóstica***

| FASE | OBJETIVO | TIPO DE CONTROLES |
|---|---|---|
| 1 | ¿Son superiores en general los valores de la prueba en los enfermos que en los sanos? | Sujetos claramente sanos (p. ej., donantes de sangre) |
| 2 | ¿Qué valor discrimina mejor *en condiciones ideales*? ¿Cuál es el punto de corte ideal? | Sujetos sanos |
| 3 | ¿La prueba es capaz de discriminar entre enfermos y dudosos en situaciones reales? | Sujetos sospechosos de enfermedad (se tendría la intención de diagnosticarlos) |
| 4 | ¿Mejora el pronóstico al aplicar la prueba? | Enfermos a los que no se aplicó la prueba |

- Según la validez externa y el espectro de la enfermedad donde se usará:
  - no es igual diferenciar a los que están totalmente sanos de los que están totalmente enfermos
  - que usar una prueba para resolver solo los casos sospechosos o dudosos.
- Según se valore una dicotomía o todos los posibles puntos de corte:
  - hay estudios que pretenden valorar solo la *probabilidad de acierto* (dicotomía), que incluye acertar en los enfermos (sensibilidad), acertar en los no enfermos (especificidad), acertar cuando la prueba sea positiva (valor predictivo positivo) o cuando sea negativa (valor predictivo negativo),
  - en cambio, otros estudios valoran globalmente la capacidad que tiene la prueba de discriminar entre enfermos y no enfermos para todos los posibles valores que puede presentar la prueba, sobre la base de usar cada vez un punto de corte distinto (curvas ROC, estadístico C, índice H de Harrell).

Fryback y Thornbury (5) describieron una jerarquía de seis niveles de evidencia para la valoración de una prueba diagnóstica:

1. La calidad técnica de la información que proporciona la prueba.
2. La validez diagnóstica.
3. El cambio de mentalidad y de criterios diagnósticos que introducirá la prueba en el médico que la aplicará.
4. El cambio que acarreará la introducción de la prueba en el plan de manejo clínico del paciente.
5. El cambio en el pronóstico del paciente.
6. Los costes y beneficios desde la perspectiva de toda la sociedad.

En los estudios dirigidos a valorar la validez de pruebas de diagnóstico es especialmente importante tener en cuenta un sesgo conocido como *efecto Hawthorne*, que consiste en que las personas se comportarán de manera distinta a la habitual cuando se saben observadas. Si alguien sabe que sus datos autorreferidos sobre el peso luego van a ser comprobados objetivamente pesándole en una báscula, será artificialmente más sincero que en la vida real cuando se le pida que escriba su peso en un cuestionario y asuma que nadie lo va a comprobar después.

## 15.3. CONSISTENCIA INTERNA EN ESCALAS CUANTITATIVAS: ALFA DE CRONBACH

En muchos instrumentos de medida, fundamentalmente en cuestionarios, hay una serie de preguntas o ítems que pretenden medir el mismo *constructo* o concepto teórico. Se requerirá entonces que las preguntas que componen la escala tengan consistencia interna, de modo que valoren el mismo concepto. Las respuestas obtenidas deberían estar correlacionadas entre sí y no habrían de ser independientes unas de otras. Un coeficiente útil para medir este grado de consistencia interna es el *alfa (α) de Cronbach*. Se basa en que si los ítems fuesen independientes, la varianza de su suma (varianza total) sería la suma de las varianzas de cada uno de los ítems. Lo que aquí se desea, en cambio, es que *no* sean independientes, sino que estén muy relacionados entre sí. Entonces la varianza de la suma será mayor que la suma de las varianzas de cada ítem. La fórmula del coeficiente es:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right)$$

*Tabla 15.2 Alfa de Cronbach para cuatro ítems (A, B, C y D) en cinco sujetos*

| ID | A | B | C | D | SUMA |
|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 3 | 7 |
| 2 | 4 | 1 | 2 | 1 | 8 |
| 3 | 4 | 2 | 3 | 5 | 14 |
| 4 | 6 | 3 | 4 | 8 | 21 |
| 5 | 4 | 3 | 5 | 8 | 20 |
| Var | 2,00 | 1,00 | 2,50 | 9,50 | 42,50 |
| | | | | Suma var ítems= | 15,00 |
| | | | | 1-(SUM S^2/STOT^2)= | 0,6471 |
| | | | | k/(k–1)= | 1,3333 |
| | | | | alfa= | 0,8627 |

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum s_i^2}{s_T^2}\right) = \frac{4}{4-1}\left(1-\frac{15}{42,5}\right) = 0,8627$$

donde $k$ es el número de ítems contenidos en la escala, $s_i^2$ es la varianza de cada uno de los ítems y $s_T^2$ es la varianza de la puntuación total calculada a través de la suma de las puntuaciones de cada ítem. Puede deducirse que cuando la suma de las varianzas de cada ítem sea igual a la varianza total, alfa valdrá 0. Cuanto más se aleje alfa de 0, mejor será la consistencia interna (6).

La tabla 15.2 presenta los detalles del cálculo. En STATA, el ejemplo se resolvería con la orden **alpha:**

```
. input A B C D
                A         B         C         D
  1.  2         1         1         3
  2.  4         1         2         1
  3.  4         2         3         5
  4.  6         3         4         8
  5.  4         3         5         8
  6.  end

. alpha A B C D

Test scale = mean(unstandardized items)

Average interitem covariance:     2.291667
Number of items in the scale:            4
Scale reliability coefficient:      0.8627
```

```
clear
input A B C D
2 1 1 3
4 1 2 1
4 2 3 5
6 3 4 8
4 3 5 8
end
alpha A B C D
```

Este mismo ejemplo se puede resolver con SPSS desde la opción ANALIZAR-ESCALAS-ANÁLISIS DE FIABILIDAD, o bien en la sintaxis usando la orden:

```
RELIABILITY
/VAR=A B C D
/MOD=ALPHA.
```

Los valores del coeficiente pueden oscilar entre 0 y 1. Un coeficiente de 0 se obtendría si todos los ítems fuesen independientes (cada uno mediría constructos distintos), y hubiese ausencia total de consistencia interna. En cambio, un coeficiente de 1 indica la máxima consistencia interna posible de la escala. Si todos los ítems tuviesen el mismo valor, alfa valdría 1. El valor del coeficiente

depende de diferentes factores; uno de los más importantes es el número de ítems de la escala. Cuantos más ítems tenga la escala, mayor será su coeficiente $\alpha$.

Tras ejecutar la orden **alpha** en STATA, es interesante complementar la información ejecutando la orden seguida de las opciones **detail** e **item.**

```
. alpha A B C D, detail item

Test scale = mean(unstandardized items)
```

| Item | Obs | Sign | item-test correlation | item-rest correlation | average interitem covariance | alpha |
|---|---|---|---|---|---|---|
| A | 5 | + | 0.7593 | 0.6402 | 2.916667 | 0.8607 |
| B | 5 | + | 0.9971 | 0.9959 | 2.75 | 0.8115 |
| C | 5 | + | 0.9459 | 0.9080 | 2.166667 | 0.7647 |
| D | 5 | + | 0.9580 | 0.8609 | 1.333333 | 0.8889 |
| Test scale | | | | | 2.291667 | 0.8627 |

```
Interitem covariances (obs=5 in all pairs)

          A         B         C         D
A    2.0000
B    1.0000    1.0000
C    1.5000    1.5000    2.5000
D    2.5000    3.0000    4.2500    9.5000
```

La opción **item** mostrará las correlaciones de cada uno de los ítems con la puntuación total, indicando su signo y su magnitud. Ha de tenerse en cuenta que, cuando el signo sea negativo, la respuesta en ese ítem deberá puntuar de manera inversa en la puntuación total, es decir, para un participante que responda un valor de 2 en una escala de 0 a 10, lo que aporte ese ítem a la puntuación total será 8. En la última columna, se muestra cuál sería el valor del coeficiente $\alpha$ de Cronbach si se eliminase ese ítem específico de la escala final y se volviese a recalcular. Así, en el ejemplo, si se suprimiese el ítem D, alfa aumentaría de 0,8627 a 0,8889. La opción **detail** muestra la matriz de covarianzas de los distintos ítems.

## 15.4. REPRODUCIBILIDAD: ÍNDICE KAPPA DE CONCORDANCIA EN VARIABLES CUALITATIVAS

### 15.4.1. Porcentaje de acuerdo simple

Antes de ver el coeficiente kappa de Cohen, se debe hablar del porcentaje de acuerdo simple. Para estimarlo se construirá una tabla 2 × 2 donde las pruebas para analizar se crucen (tabla 15.3). Se preguntó sobre el consumo de suplementos vitamínicos en dos entrevistas y las respuestas coincidieron en 55 + 35 ocasiones de 107 posibles. El porcentaje de acuerdo es de (55 + 35)/107 = 84,1%.

***Tabla 15.3 Porcentaje de acuerdo simple: consumo de suplementos vitamínicos (No/Sí) según dos entrevistas***

| | | ENTREVISTA 1: ¿CONSUME VD. SUPLEMENTOS VITAMÍNICOS? | | TOTAL |
|---|---|---|---|---|
| | | NO | SÍ | |
| Entrevista 2: responda si consume vitaminas sin contar sus aportes alimentarios | No | 35 | 12 | 47 |
| | Sí | 5 | 55 | 60 |
| Total | | 40 | 67 | 107 |

El acuerdo simple es (35 + 55)/107 = 84,1%.

### 15.4.2. Índice kappa de concordancia

El índice kappa de Cohen sirve para valorar la concordancia entre dos métodos de recogida de datos o entre dos observadores distintos que aplican el mismo instrumento. A diferencia del porcentaje de acuerdo simple, eliminará las coincidencias explicables por el azar.

En el ejemplo de la tabla 15.3, podría afirmarse que la concordancia simple fue del 84,1%, pero hay que tener en cuenta que algunas de las coincidencias entre ambas entrevistas podrían deberse al azar y hay que descontarlas del total. Imagínense dos personas lanzando monedas al azar: siempre habrá ocasiones en que coincidan por causalidad. Para calcular la concordancia esperada por el azar se procede como en un test de ji cuadrado: manteniendo fijos los marginales (tabla 15.4), se multiplica el total de la fila por el total de la columna y se divide por el total de los totales.

La suma de la diagonal en que concuerdan por azar ambos procedimientos es la concordancia esperada = 0,515.

Si se descuenta de la concordancia simple (0,841) esta probabilidad esperada solo por el azar (0,515), quedará una concordancia no explicada por el azar de 0,841 – 0,515 = 0,326.

Lo que sucede es que ahora hay que referirla al total posible de concordancia una vez excluido el azar (1 – 0,515 = 0,485). El coeficiente kappa es el cociente resultante de dividir la diferencia concordancia observada-esperada al azar (numerador) entre el denominador formado por la diferencia entre la unidad y la esperada al azar.

$$\text{kappa} = \frac{\text{concordancia observada} - \text{concordancia esperada}_{\text{por azar}}}{1 - \text{concordancia esperada}_{\text{por azar}}} = \frac{0{,}841 - 0{,}515}{1 - 0{,}515} = 0{,}672$$

El valor máximo para un índice kappa es 1, que indica coincidencia total. El mínimo puede ser negativo. En general, puede afirmarse que un índice kappa es *excelente* si se sitúa por encima de 0,8, *bueno o moderado* si está por encima de 0,6, y *aceptable* si supera 0,4.

Hay dos posibles problemas con el uso de kappa al comparar los diagnósticos hechos por dos profesionales distintos sobre los mismos pacientes:

a. *Paradoja del sesgo:* si uno de los observadores tiende a diagnosticar la enfermedad más que el otro, kappa tenderá a ser más bajo.

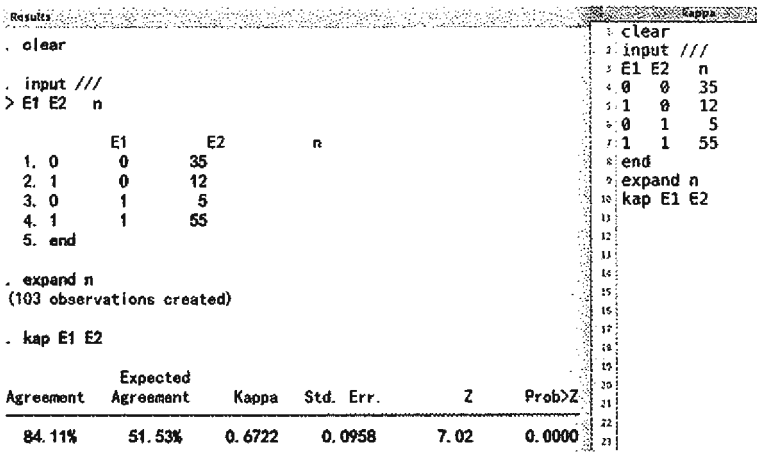
b. *Paradoja de la prevalencia:* cuanto mayor sea la prevalencia de la enfermedad, más elevado tenderá a ser kappa. Esto explica, en parte, que la reproducibilidad entre especialistas sea habitualmente superior a la que existe entre médicos generales, ya que en consultas de especialidades hay mayor prevalencia de la enfermedad.

***Tabla 15.4 Valores esperados de coincidencia por azar en la tabla 15.3***

| | | ENTREVISTA 1: ¿CONSUME VD. SUPLEMENTOS VITAMÍNICOS? | | TOTAL |
|---|---|---|---|---|
| | | NO | SÍ | |
| Entrevista 2: responda si consume vitaminas sin contar sus aportes alimentarios | No | $\frac{47 \times 40}{107} = 17{,}57$ | $\frac{47 \times 67}{107} = 29{,}43$ | 47 |
| | Sí | $\frac{60 \times 40}{107} = 22{,}43$ | $\frac{60 \times 67}{107} = 37{,}57$ | 60 |
| Total | | 40 | 67 | 107 |

La concordancia esperada por azar sería: (17,57 + 37,57)/107 = 51,5%

En STATA se usará la orden **kap,** seguida de los nombres de las dos variables, tal como muestra la siguiente imagen:



```
Results
. clear

. input ///
> E1 E2   n

            E1          E2          n
  1.  0         0          35
  2.  1         0          12
  3.  0         1           5
  4.  1         1          55
  5.  end

. expand n
(103 observations created)

. kap E1 E2

 Agreement   Expected Agreement      Kappa    Std. Err.          Z      Prob>Z
------------------------------------------------------------------------------
    84.11%               51.53%     0.6722     0.0958       7.02     0.0000
```

```
Kappa
1  clear
2  input ///
3  E1 E2   n
4  0  0   35
5  1  0   12
6  0  1    5
7  1  1   55
8  end
9  expand n
10 kap E1 E2
11
12
13
14
15
16
17
18
19
20
21
22
23
```

En SPSS se obtiene el índice kappa desde la opción ANALIZAR-ESTADÍSTICOS DESCRIPTIVOS-TABLAS DE CONTINGENCIA, o bien con la siguiente sintaxis:

```
CROSSTABS
/TAB=E1 BY E2
/STAT=KAPPA.
```

El índice kappa que se debe usar cuando existen *más de dos* categorías ordinales de clasificación es el kappa ponderado. Por ejemplo, si se trata de valorar el grado de acuerdo al clasificar pacientes en IV estadios tumorales, parece lógico otorgar más peso a los desacuerdos extremos (clasificar el estadio IV como I, o viceversa) que a los inmediatos (el I como II). Para más información se pueden consultar otras referencias (**1**) y las ayudas correspondientes de STATA.

## 15.5. COEFICIENTE DE CORRELACIÓN INTRACLASE: CONCORDANCIA EN VARIABLES CUANTITATIVAS

Cuando se desea valorar la reproducibilidad de dos pruebas que emplean una escala cuantitativa en sus mediciones, lo *peor* que puede hacerse es una *t* de Student emparejada, ya que en la *t* de Student emparejada cuanto *menor* sea el tamaño muestral, más anchura tendrá el intervalo de confianza y más difícil será distinguir en qué se diferencian las dos mediciones (paradójicamente, cuanto más ancho sea el intervalo, más compatibles serán los resultados con la igualdad [diferencia = 0] entre las dos mediciones). En definitiva, debe quedar claro que la *t* de Student emparejada *no sirve en absoluto* para medir acuerdo.

Tampoco debe usarse el índice kappa de concordancia mediante la categorización de una variable que originalmente fuese cuantitativa. No es el método de elección, ya que la categorización se traduciría en una pérdida de información.

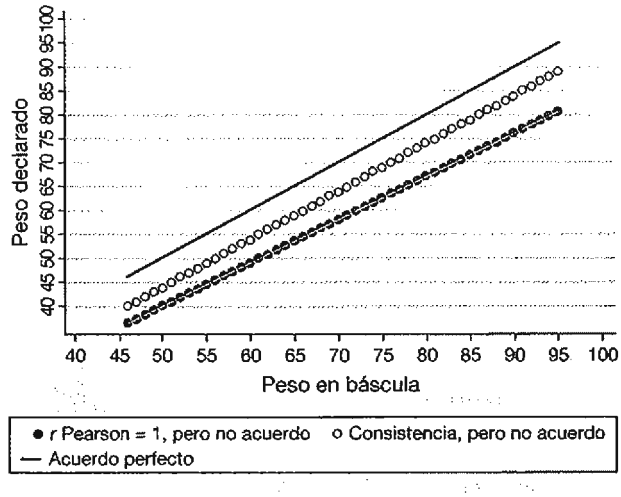



**Figura 15.1** Diferencias entre correlación (Pearson), consistencia y acuerdo. El coeficiente de correlación intraclase permite estimar acuerdo.

Una alternativa es el cálculo de un coeficiente de correlación (Pearson o Spearman, vistos en el apartado 10.2). Esta alternativa tampoco es ideal, a no ser que solo se desee valorar consistencia. Lo decisivo es que el coeficiente de correlación (*r de Pearson* o *rho de Spearman*) mide asociación entre dos variables, pero no el grado de acuerdo entre ellas; puede ser que datos que presentan pobre acuerdo presenten elevados coeficientes de correlación (fig. 15.1).

Lo indicado en esta situación es el coeficiente de correlación intraclase (CCI) (7), que permite establecer el acuerdo entre dos o más evaluaciones de una variable *cuantitativa* llevadas a cabo sobre el mismo número de personas. Permitirá, igualmente, valorar la consistencia entre dos o más evaluaciones. No debe olvidarse que el acuerdo es algo más deseable (y más difícil) que la consistencia o correlación. El CCI para estimar el acuerdo es el que debe usarse para medir concordancia, que es lo que se espera buscar en un estudio de validación. Imagínese el ejemplo representado en la figura 15.1.

El CCI tiene que ver con el ANOVA de dos factores. Imagínese la medición repetida del peso en 10 sujetos, primero basada en datos autorreferidos (declaración) y luego con medición directa (báscula). Según aparece en la tabla 15.5, el CCI se deriva del propio ANOVA de dos factores.

En la tabla se aprecia que se ha llevado a cabo una descomposición de la variabilidad total de la muestra a través de un ANOVA de dos factores: los sujetos *(id)* y las valoraciones *(rater)*. En STATA se ha adoptado el formato largo **(long)** mediante la orden **reshape,** que permite que las dos columnas del peso (declaración y báscula) se conviertan en una sola columna, y se añade una nueva variable *(rater),* que tomará distintos valores según el tipo de medición.

Una vez reestructurada la base de datos con formato *long,* bastará con usar la orden **icc** para obtener con STATA el CCI. El que STATA proporciona por omisión es el de *acuerdo total,* que resulta el más interesante y es el que se suele buscar cuando se aplica este coeficiente en estudios de validación.

*Tabla 15.5 Coeficiente de correlación intraclase: acuerdo entre peso declarado y peso medido*

| ID | DECLARACIÓN (W1) | BÁSCULA (W2) |
|---|---|---|
| 1 | 52 | 55 |
| 2 | 54 | 57 |
| 3 | 56 | 60 |
| 4 | 67 | 70 |
| 5 | 78 | 80 |
| 6 | 80 | 82 |
| 7 | 85 | 86 |
| 8 | 82 | 88 |
| 9 | 90 | 90 |
| 10 | 110 | 111 |

Se introdujeron los datos en dos columnas en STATA, por eso se debe preceder el ANOVA de la siguiente orden:

```
reshape long w, i(id) j(rater)
anova w id rater
```

```
                   Number of obs =        20     R-squared     =  0.9977
                   Root MSE      =  1.21335     Adj R-squared =  0.9952
```

| Source | Partial SS | df | MS | F | Prob > F |
|---|---|---|---|---|---|
| Model | 5819.3 | 10 | 581.93 | 395.27 | 0.0000 |
| id | 5788.05 | 9 | 643.116667 | 436.83 | 0.0000 |
| rater | 31.25 | 1 | 31.25 | 21.23 | 0.0013 |
| Residual | 13.25 | 9 | 1.47222222 | | |
| Total | 5832.55 | 19 | 306.976316 | | |

$$CCI = \frac{n(MS_{id} - MS_{resid})}{n(MS_{id}) + k(MS_{raters}) + (nk - n - k)(MS_{resid})} =$$

$$= \frac{10 \times (643.12 - 1.47)}{10 \times (643.12) + 2 \times (31.25) + (10 \times 2 - 10 - 2)(1.47)} = 0.986$$

```
. icc w id rater

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: id              Number of targets =      10
Random effects: rater           Number of raters  =       2

------------------------------------------------------------------------------
                 |      ICC   [95% Conf. Interval]
-----------------+------------------------------------------------------------
      Individual |  .9863192   .6240103    .9975845
         Average |  .9931125   .7684807    .9987908
------------------------------------------------------------------------------
F test that
  ICC=0.00: F(9.0, 9.0) = 436.83                Prob > F = 0.000

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

```
clear
input ///
id w1 w2
1   52  55
2   54  57
3   56  60
4   67  70
5   78  80
6   80  82
7   85  86
8   82  88
9   90  90
10  110 111
end

reshape long w, i(id) j(rater)


icc w id rater
```

El CCI que se buscaba es el primero que aparece, junto con el nombre «Individual». El resultado se podrá presentar así:

$$CCI = 0{,}986 (IC95\%: 0{,}624 - 0{,}998)$$

En SPSS, se pide el CCI desde la opción ANALIZAR-ESCALAS-ANÁLISIS DE FIABILIDAD. Dentro del botón *Estadísticos...* se debe seleccionar *Coeficiente de correlación intraclase*, con las opciones *Dos factores, efectos aleatorios*, y eligiendo en la ventana de la derecha la opción *Acuerdo absoluto*.

Con sintaxis, en SPSS se puede pedir como:

```
RELIABILITY
/VAR=w1 w2
/MODEL=ALPHA
/ICC=MODEL(RANDOM) TYPE(ABSOLUTE) CIN=95.
```

También existe otro CCI que mide consistencia, no acuerdo. No obstante, la recomendación de usar el CCI para valorar validez o reproducibilidad no suele hacerse para buscar consistencia, sino acuerdo total, como se explica más arriba. Por tanto, hay que adoptar la precaución de comprobar, incluso por duplicado, que lo que se pidió al ordenador es el coeficiente de *acuerdo total* (es el mostrado por omisión en STATA, pero *no* en SPSS).

La interpretación del CCI es análoga a la del coeficiente kappa. Si el CCI (acuerdo) es inferior a 0,4, se hablará de un *pobre* acuerdo; si está por encima de 0,75, será muy bueno o *excelente*. Si está entre ambos, se considerará bueno *(moderado)*.

## 15.6. GRÁFICOS DE BLAND-ALTMAN PARA ACUERDO EN VARIABLES CUANTITATIVAS

Bland y Altman (8) propusieron establecer el grado de concordancia entre dos pruebas diagnósticas medidas en escala cuantitativa (A y B) o entre dos mediciones repetidas de una misma prueba en un grupo de sujetos (A y A') mediante la construcción de *límites de tolerancia*. Estos límites estadísticos se calculan a través de la media ( $\bar{x}$ ) de las dos mediciones y de la diferencia *(d)* entre ambas. La mayoría de las diferencias, si se sigue una distribución normal, deberían situarse aproximadamente entre la media de la diferencia y ±2 desviaciones estándar de la variable diferencia (exactamente, entre $\bar{d} \pm 1{,}96s$ ). Aunque, en general, las propias mediciones no siguen la distribución normal, su diferencia sí suele hacerlo.

Para aplicar este método se construye una gráfica en la que el eje de ordenadas *(y)* representa la diferencia entre las mediciones ($d_i = A_i - B_i$) y el eje de abscisas *(x)* muestra la media de ambas mediciones ($\bar{x}_i = (A_i + B_i)/2$). De esta manera, el gráfico permite investigar cualquier posible relación entre el error de medida y el valor real, evaluar la magnitud del desacuerdo entre mediciones o identificar valores *outliers* o periféricos. Por ejemplo, una concordancia perfecta entre pruebas diagnósticas produce una línea paralela al eje de las abscisas con una ordenada igual a 0.

Supóngase el mismo ejemplo usado para el coeficiente de correlación intraclase (v. tabla 15.5). Se dispone de los valores de peso por declaración *(w1)* y peso medido por báscula *(w2)* en un grupo de 10 pacientes. La figura 15.2 recoge este método.

Se pueden hacer estos cálculos de manera sencilla en cualquier programa. Existe una instrucción en STATA que se llama **concord** y está programada para realizarlo. La instrucción **concord** no está incorporada por defecto en STATA, pero, si se está conectado a internet, se puede descargar con el sistema habitual de búsqueda:

```
findit concord
```

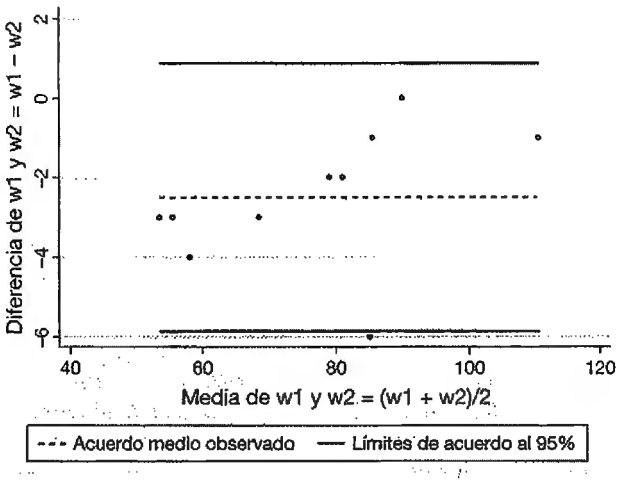



**Figura 15.2** Método de Altman y Bland para estimar acuerdo usando los mismos datos de peso declarado y medido de la tabla 15.5. Se ha usado la orden **concord** de STATA (que no viene por defecto en el programa y debe buscarse en internet con **findit concord**).

Se ofrece una explicación detallada de las características de descarga en: http://ideas.repec.org/c/boc/bocode/s404501.html.

Una vez encontrada e instalada la orden, se procedería del modo siguiente al usar los datos de la tabla 15.5:

```
concord w1 w2, loa
```

La opción **loa** pide a STATA la representación del gráfico de Bland y Altman.

## 15.7. COEFICIENTE DE CORRELACIÓN DE CONCORDANCIA DE LIN

La instrucción **concord** proporciona también el coeficiente de correlación de Lin (**9,10**). Es un coeficiente robusto de correlación de concordancia que puede tener valores desde −1 a +1, y sus valores absolutos *no* pueden superar a los del coeficiente de correlación de Pearson *(r)*.

El coeficiente de Lin mide el acuerdo absoluto entre dos valoraciones cuantitativas continuas de una misma variable. En concreto, este coeficiente estima la medida en que los puntos de las dos mediciones repetidas se aproximan o coinciden con la diagonal de 45° que representaría la coincidencia o identidad perfecta entre los dos métodos de medida si uno ocupase el eje *x* y otro el eje *y*. Esta diagonal sería la línea de perfecta concordancia y produciría un coeficiente de Lin igual a 1. Cuanto mayor sea la proximidad de los puntos a la diagonal, más cercano a 1 será el coeficiente de Lin.

En STATA, al aplicar la orden **concord** (v. apartado 15.6) a los valores de pesos obtenido por declaración *(w1)* o por báscula *(w2)* presentados en la tabla 15.5, se obtendrían los siguientes resultados.

```
. concord w1 w2

Concordance correlation coefficient (Lin, 1989, 2000):

 rho_c   SE(rho_c)    Obs     [  95% CI   ]      P          CI type
------------------------------------------------------------------
 0.985      0.009      10      0.968  1.002    0.000     asymptotic
                              0.953  0.995    0.000     z-transform

Pearson's r =  0.996   Pr(r = 0) = 0.000   C_b = rho_c/r =  0.988
Reduced major axis:    Slope =     1.044    Intercept =    -5.954

Difference = w1 - w2

          Difference                    95% Limits Of Agreement
    Average      Std Dev.                (Bland & Altman, 1986)
------------------------------------------------------------------
     -2.500        1.716                    -5.863       0.863

Correlation between difference and mean = 0.454

Bradley-Blackwood F = 12.923  (P = 0.00312)
```

El coeficiente de Lin valdría 0,985 (IC 95%: 0,953 a 0,995).

Debajo de la estimación del coeficiente aparece un resumen del método de Altman y Bland. Puede recurrirse a la ayuda de STATA para analizar otros detalles de esta salida.

## 15.8. REGRESIÓN DE PASSING-BABLOK Y REGRESIÓN DE DEMING

En el contexto de comparar dos técnicas ($x$ e $y$) de laboratorio de manera que ambas sean susceptibles de error para determinar la misma variable, la regresión de Deming y la de Passing-Bablok son métodos ideados para realizar una estimación de la recta de regresión *ortogonal* que minimice las sumas de cuadrados de las distancias perpendiculares desde cada punto hasta la recta de coincidencia perfecta (**11,12**). Téngase en cuenta que, en la regresión ordinaria de mínimos cuadrados (v. apartado 10.5.3), lo que se minimiza son las sumas de cuadrados de las distancias verticales (residuales), pero aquí no son las verticales, sino las *perpendiculares* a la recta, las que se minimizan.

La regresión de Passing-Bablok es la preferible, por ser más robusta. Usa métodos no paramétricos. Se basa en obtener todas las posibles combinaciones de puntos $(x_i, y_i)$ observados tomándolos de dos en dos. Habría, por tanto, $n(n-1)/2$ combinaciones (p. ej., con 6 puntos habría 15 parejas de puntos). Se calcula una serie de pendientes de las rectas que unirían cada una de esas posibles parejas de puntos, y se utiliza la mediana de estas pendientes como pendiente global. A partir de esta pendiente global, se calcula una ordenada en el origen para cada uno de los puntos $(a = y - bx)$, y se toma como ordenada en el origen global la mediana de todas ellas. Los intervalos de confianza se obtienen por técnicas de ordenación.

Estos métodos no están implementados en STATA ni en SPSS. Sí lo están en SAS y hay macros para ejecutarlos en R/Splus. También existen algunos programas específicos pensados para laboratorios de bioquímica clínica preparados para este tipo de regresiones:

- Analyse-it: http://analyse-it.com/
- CBStat: http://www.cbstat.com/

*Tabla 15.6 Acuerdos entre peso declarado y peso medido: preparación de los datos para hacer un gráfico de acuerdo-supervivencia*

| ID | DECLARACIÓN (W1) | BÁSCULA (W2) | DIF. ABS. | $n_i$ | $s_i$ | K-MEIER ($S_i$) |
|---|---|---|---|---|---|---|
| 9 | 90 | 90 | 0 | 10 | 10 | 1 |
| 7 | 85 | 86 | 1 | 9 | 7 | 0,778 |
| 10 | 110 | 111 | 1 | | | |
| 5 | 78 | 80 | 2 | 7 | 5 | 0,556 |
| 6 | 80 | 82 | 2 | | | |
| 1 | 52 | 55 | 3 | 5 | 2 | 0,222 |
| 2 | 54 | 57 | 3 | | | |
| 4 | 67 | 70 | 3 | | | |
| 3 | 56 | 60 | 4 | 2 | 1 | 0,111 |
| 8 | 82 | 88 | 6 | 1 | 8 | 0 |

## 15.9. GRÁFICOS DE ACUERDO-SUPERVIVENCIA

Una aproximación reciente (**13**) para la determinación del grado de concordancia entre dos mediciones cuantitativas es el uso de gráficos de supervivencia (*Kaplan-Meier*; v. apartado 11.2) para construir los denominados *survival-agreement plots (gráficos de acuerdo-supervivencia)*. Se logra así expresar el grado de concordancia como función de las diferencias absolutas entre mediciones. La idea es valorar cómo se distribuye el porcentaje de desacuerdo a medida que el desacuerdo aumenta. Se representa el porcentaje de desacuerdo entre ambas pruebas diagnósticas en el sitio que correspondería al porcentaje de supervivencia en el Kaplan-Meier tradicional. Así, en el eje de abscisas *(x)* se representan las diferencias absolutas en las mediciones $|A_i - B_i| = \left(\left|\bar{d}_i\right|\right)$ y en el de ordenadas *(y)*, la proporción de casos con valores iguales o mayores a la diferencia observada $\left(\geq \left|\bar{d}_i\right|\right)$. De esta manera, el grado de acuerdo para cada diferencia absoluta $\left(\left|\bar{d}_i\right|\right)$ se calcula como la distancia entre la curva y el límite superior del eje de ordenadas (100%).

Además, este tipo de análisis es útil para la comparación de más de dos pruebas diagnósticas o mediciones a través de la construcción de distintas curvas de supervivencia. Es posible usar el test de *log-rank* o la regresión de Cox para evaluar si la diferencia entre las dos curvas de concordancia es estadísticamente significativa.

Supóngase el ejemplo anterior de los dos pesos (*w1* y *w2*, según tabla 15.5). La tabla 15.6 incluye nuevas columnas que expresan, de izquierda a derecha, sucesivamente la diferencia absoluta entre ambas variables *(ABS(w1-w2))*, el número de sujetos *($n_i$)* en riesgo de presentar esa diferencia absoluta o una todavía mayor, los que superan esa diferencia absoluta *($s_i$)* y el estimador de Kaplan-Meier ($S_i$).

$$S_i = \prod \frac{s_i}{n_i}$$

En STATA se pueden introducir los datos como aparece en la tabla 15.7. Se obtendría una imagen como la figura 15.3, en la que puede verse cómo disminuye el porcentaje de desacuerdo a medida que crece la magnitud de este desacuerdo.

Este gráfico tiene una interpretación clara y práctica desde el punto de vista clínico. Si se considera que diferencias absolutas de peso de hasta 3 kg o menos no tienen importancia clínica, podría usarse el complementario de la supervivencia en ese punto ($1 - S_3 = 1 - 0{,}22 = 0{,}78$) para afirmar que el 78% de la muestra tuvo diferencias entre uno y otro método que solo eran inferiores o iguales a 3 kg.

***Tabla 15.7 Modo de introducir en STATA (o SPSS) los datos de la Tabla 15.6 para hacer un gráfico de acuerdo-supervivencia***

| DIF. ABS. | DESACUERDO |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 1 | 1 |
| 2 | 1 |
| 2 | 1 |
| 3 | 1 |
| 3 | 1 |
| 3 | 1 |
| 4 | 1 |
| 6 | 1 |

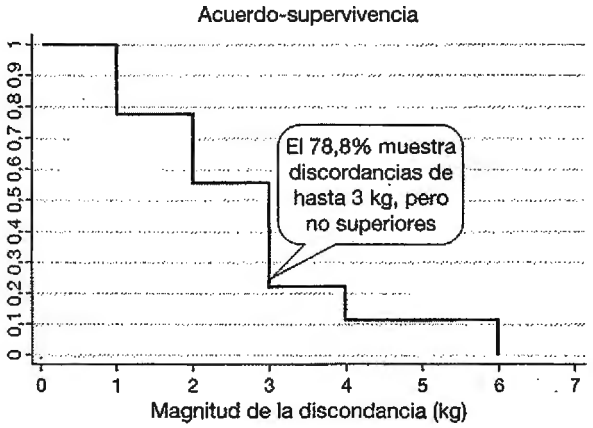



**Figura 15.3** Gráfico de acuerdo-supervivencia con los datos de la tabla 15.5 de peso declarado y medido. Se introdujeron los datos en STATA según muestra la tabla 15.7 y se dieron luego las siguientes órdenes: **stset *dif_abs*, fail(*desac*) sts graph, xlab(*0(1)7*) /// ylab(*0(.1)1*, format(*%9.1f*)) /// xtit(" " "*Magnitud de la discordancia (kg)*") /// title("*Acuerdo-supervivencia*").

Llorca y Delgado-Rodríguez (14) propusieron una modificación en la construcción de este tipo de gráfico. Su propuesta consistió en representar dos gráficos: uno para diferencias positivas y otro para diferencias negativas, en vez de representar la diferencia absoluta global. Este nuevo método de determinación de la concordancia mejoraría la caracterización del error de medida, describiendo la dirección y el sentido del mismo.

Las curvas construidas pueden ser comparadas a través de la prueba de *log-rank*. Además, puede estudiarse el efecto de otras covariables sobre el error de medida a través de un modelo multivariable, como la regresión de Cox.

## 15.10. VALIDEZ DIAGNÓSTICA: SENSIBILIDAD, ESPECIFICIDAD, VALORES PREDICTIVOS, RAZONES DE VEROSIMILITUD

### 15.10.1. Sensibilidad y especificidad

Para valorar la validez de una prueba, la condición *sine qua non* es que se compare la prueba con un criterio de referencia o patrón de oro *(gold standard)* que se utiliza como criterio de

*Tabla 15.8 Posibles resultados de una prueba diagnóstica*

| | | VERDAD *(GOLD STANDARD)* | |
|---|---|---|---|
| | | ENFERMEDAD | |
| | | Presente | Ausente |
| Prueba diagnóstica | Positiva | Verdadero positivo (VP) | Falso positivo (FP) |
| | Negativa | Falso negativo (FN) | Verdadero negativo (VN) |

**RESULTADOS ENCONTRADOS EN EL ESTUDIO DE VALIDACIÓN DE CÁNCER AUTORREFERIDO EN LA COHORTE EPIC-ESPAÑA**

| CÁNCER AUTORREFERIDO | SUJETOS CON CÁNCER SEGÚN PATRÓN DE ORO | | TOTAL |
|---|---|---|---|
| | Sí | No | |
| Sí | 184 | 76 | 260 |
| No | 136 | 40.842 | 40.978 |
| Total | 320 | 40.918 | 41.238 |

*verdad*. El esquema más simple para analizar cualquier prueba diagnóstica es el de una tabla en la que se cruzan los posibles resultados (positivo o negativo) de la prueba diagnóstica que se evalúa frente a la «verdad» valorada por el criterio de referencia (es decir, si está presente o ausente) (tabla 15.8).

Como ejemplo, en la parte inferior de la tabla se recogen los resultados obtenidos para la validación del diagnóstico autorreferido de cáncer en el estudio EPIC-España (**15**). Los investigadores validaron el diagnóstico de cáncer obtenido a través de la pregunta de un cuestionario: utilizando como *gold standard* el diagnóstico clínico obtenido a través de registros poblacionales. En la tabla 15.8 se puede calcular la sensibilidad y la especificidad.

**Sensibilidad (S):** es un parámetro que se mide en los que verdaderamente están enfermos. Es el cociente entre los verdaderos positivos (VP) y el total de personas enfermas.

$$S = \frac{VP}{\text{enfermos}} = \frac{VP}{VP+FN}$$

Su complementario es la proporción de falsos negativos (FN) entre los enfermos:

$$1 - S = \frac{VP + FN}{VP + FN} - \frac{VP}{VP + FN} = \frac{FN}{VP + FN}$$

Una prueba muy sensible detecta muchos verdaderos positivos y pocos falsos negativos. Su utilidad principal es la de no dejar de *diagnosticar* la enfermedad en ningún paciente. La sensibilidad es especialmente importante cuando una enfermedad no debe pasar desapercibida (p. ej., ante una enfermedad contagiosa) y el pronóstico mejora mucho con el tratamiento precoz (p. ej., fenilcetonuria o hipotiroidismo).

**Especificidad (E):** es un parámetro que se mide en los no enfermos. Es el cociente de los verdaderos negativos (VN) entre el total de no enfermos.

$$E = \frac{VN}{\text{no enfermos}} = \frac{VN}{VN + FP}$$

Su complementario es la proporción de falsos positivos (FP) entre los no enfermos:

$$1 - E = \frac{FP}{VN + FP}$$

Una prueba muy específica identifica muchos verdaderos negativos y pocos falsos positivos. Su utilidad principal es la de *confirmar* la enfermedad. La especificidad es especialmente importante cuando la enfermedad cambia la vida del enfermo o tiene cierto estigma (no se quiere preocupar

inútilmente o estigmatizar erróneamente a alguien con un falso positivo) y también cuando las consecuencias de un tratamiento suponen un riesgo para el enfermo (amputación, etc.).

En bastantes ocasiones se requieren pruebas que cumplan las dos exigencias de tener sensibilidad y especificidad altas y, por esta razón, se utiliza muchas veces la información de varias pruebas de manera combinada.

En el ejemplo del estudio EPIC-España, la sensibilidad y la especificidad de la pregunta sobre cáncer contenida en el cuestionario serían del 57,5 y del 99,8%, respectivamente:

$$S = \frac{184}{184+136} = 0{,}575 = 57{,}5\%$$

$$E = \frac{40.842}{40.842+76} = 0{,}998 = 99{,}8\%$$

Se ha afirmado clásicamente que la sensibilidad y la especificidad son criterios de validez interna, porque se pensaba que no estaban sometidos a la influencia de elementos ajenos a la enfermedad y a la prueba. No es así, y la prevalencia de la enfermedad se ha mostrado que influye. El aumento de la prevalencia aumenta la sensibilidad y disminuye la especificidad. La influencia es menor que en otros parámetros tratados más adelante.

### 15.10.2. Validez externa de las pruebas diagnósticas: valores predictivos

La sensibilidad indica la proporción de los realmente enfermos que resultan positivos (verdaderos positivos) en una prueba, pero en el mundo real normalmente no se sabe *a priori* quién está enfermo. Lo que más le suele interesar al médico desde el punto de vista práctico es otro aspecto de esa probabilidad: ¿cuántos de los pacientes que dieron un resultado positivo en la prueba están realmente enfermos? Este es el *valor predictivo positivo*. En términos probabilísticos, y si llamamos D al desenlace (tener la enfermedad en verdad) y T+ a tener un resultado del test positivo, la sensibilidad y el VPP se diferencian del modo siguiente:

$$\text{Sensibilidad} = p(T+ \mid D)$$
$$\text{VPP} = p(D \mid T+)$$

De manera análoga, la especificidad estima los resultados negativos (verdaderos negativos) en los no enfermos. Ahora bien, interesa más saber cuántos de los pacientes con un resultado negativo en la prueba realmente están exentos de enfermedad. Este es el *valor predictivo negativo*. La respuesta a estas dos preguntas es más útil para interpretar el significado real de un resultado positivo o negativo obtenido tras aplicar una prueba a un paciente.

**Valor predictivo positivo (VPP):** es la probabilidad de padecer la enfermedad cuando el resultado de la prueba es positivo. Se calcula mediante la siguiente expresión:

$$\text{VPP} = \frac{\text{VP}}{\text{total test (+)}} = \frac{\text{VP}}{\text{VP + FP}}$$

En la tabla 15.9 se presentan los resultados obtenidos en un estudio de validación de la información autorreferida sobre la presencia de fracturas obtenida a través de un cuestionario enviado por correo postal a mujeres perimenopáusicas pertenecientes al *Kuopio Osteoporosis Risk Factor and Prevention Study* (OSTPRE) (**16**). El estudio consideró la presencia de fractura recogida en la historia clínica de la participante como *gold standard*.

La información recogida con el cuestionario era capaz de detectar el 78,4% de las 453 fracturas que tuvieron lugar entre las participantes de este estudio. Además, el 83,5% de las fracturas detectadas por cuestionario aparecían, en realidad, en las historias clínicas de las

*Tabla 15.9 Diferencia entre valor predictivo positivo (VPP) y sensibilidad (S)*

| | | HISTORIA CLÍNICA | | TOTAL |
|---|---|---|---|---|
| | | FRACTURA+ | FRACTURA− | |
| Fractura autorreferida (cuestionario) | Fractura+ | 355 | 70 | 425[1] |
| | Fractura− | 98 | 1.577 | 1.675 |
| | Total | 453[2] | 1.647 | |

[1] $VPP = \frac{VP}{\text{total test +}} = \frac{355}{425} = 83{,}5\%$

[2] $S = \frac{VP}{\text{total enfermos}} = \frac{355}{453} = 78{,}4\%$

*VP, verdadero positivo.*

pacientes. Este es el VPP (= 83,5%). Significa que un 16,5% de los resultados aparentemente positivos eran, en realidad, fallos en la detección de fracturas (falsos positivos). Por 355 verdaderos positivos, hay 70 falsos positivos. La *odds* del valor predictivo positivo sería (355/425)/(70/425) = 355/70 = 5. La *odds* también se puede expresar como 5:1. Es decir, por cada cinco aciertos (verdaderos positivos) de las respuestas autorreferidas de las pacientes cuando eran positivas, se producía un error (falso positivo). Esta razón (*odds* del VPP) estima el número de verdaderos positivos por cada falso positivo. Equivale a expresar el VPP como *odds* en vez de como proporción:

$$Odds_{VPP} = \frac{VP}{FP} = \frac{355}{70} = \frac{5}{1}$$

o bien:

$$VPP = 0{,}835 \Rightarrow Odds_{VPP} = \frac{0{,}835}{1 - 0{,}835} = 5$$

**Valor predictivo negativo (VPN):** es la probabilidad de no padecer la enfermedad cuando el resultado de la prueba es negativo. Se calcula dividiendo los verdaderos negativos entre el total de resultados negativos (VN + FN):

$$VPN = \frac{VN}{\text{total test (–)}} = \frac{VN}{VN + FN}$$

En el mismo ejemplo anterior, de un total de 1.675 resultados negativos (las participantes no refirieron una fractura), 1.577 fueron aciertos (verdaderos negativos), pero hubo 98 casos de fracturas que *no* fueron detectadas (falsos negativos) a través del cuestionario (v. tabla 15.9). Por lo tanto, el VPN es del 94,1% (1.577/1.675).

En los valores predictivos influye notablemente la prevalencia: cuanto mayor sea la prevalencia, mayor será el valor predictivo positivo y menor el valor predictivo negativo. Por esta razón, se suele afirmar que los valores predictivos positivos y negativos miden la validez externa de una prueba.

Véase cómo se comportan los valores predictivos si se aplican las mismas pruebas anteriores a dos grupos de mujeres distintos. En el ejemplo anterior, la edad de las mujeres oscilaba entre 47 y 56 años. Supóngase que se repite este estudio en mujeres jóvenes (18-25 años), donde la prevalencia de fracturas será mucho menor y se mantienen constantes la sensibilidad y la especificidad. El cambio en los valores predictivos es importante; especialmente, disminuye el VPP, porque ha disminuido la prevalencia (tabla 15.10).

*Tabla 15.10 Cambio de los valores predictivos al cambiar la prevalencia*

| | HISTORIA CLÍNICA | | TOTAL |
|---|---|---|---|
| | FRACTURA+ | FRACTURA− | |
| **Edad = 47-56 años** | | | |
| Fractura+ | 355 | 70 | 425 |
| Fractura− | 98 | 1.577 | 1.675 |
| Total | 453 | 1.647 | 2.100 |
| S = 78,4%; VPP = 83,5% | | | |
| E = 95,7%; VPN = 94,1% | | | |
| Prevalencia: 453/2.100 = 21,6% | | | |
| **Edad = 18-25 años** | | | |
| Fractura+ | 212 | 117 | 329 |
| Fractura− | 212 | 117 | 329 |
| Total | 270 | 2.730 | 3.000 |
| S = 78,4%; VPP = 64,4% | | | |
| E = 95,7%; VPN = 95,7% | | | |
| Prevalencia: 270/3.000 = 9% | | | |

Si se conocen la sensibilidad y la especificidad de una prueba diagnóstica, así como la prevalencia de una enfermedad en una población, se pueden calcular los VPP y VPN que resultarían de aplicar dicha prueba diagnóstica a ese grupo.

El teorema de Bayes (v. apartado 3.6) ofrece la relación matemática existente entre los valores predictivos y la prevalencia (P), la sensibilidad (S) y la especificidad (E):

$$VPP = \frac{P \times S}{P \times S + [(1-P) \times (1-E)]}$$

$$VPN = \frac{(1-P) \times E}{(1-P) \times E + [(P) \times (1-S)]}$$

No obstante, lo más sencillo desde el punto de vista práctico para estimar el VPP y el VPN a partir de la prevalencia, la sensibilidad y la especificidad es construir una tabla 2 × 2 ficticia que aplique estas probabilidades a un número redondo total *(total de los totales)*, por ejemplo, a 100 o 10.000 sujetos ficticios.

Otra situación distinta (pero frecuente en estudios de validación de pruebas diagnósticas) se produce cuando la selección de sujetos se inicia al elegir un grupo que dio positivo en la prueba y otro grupo que arrojó resultados negativos. Después, se aplica a ambos grupos (idealmente de manera enmascarada con respecto a la prueba) el *gold standard* para confirmar la presencia de enfermedad o su ausencia. Así, Sánchez-Villegas et al. (17) validaron el diagnóstico autorreferido de depresión en cuestionarios del proyecto SUN (Seguimiento Universidad de Navarra), al usar como *gold standard* una entrevista diagnóstica con psiquiatra. Para ello eligieron a quienes habían contestado en el cuestionario que habían recibido un diagnóstico médico de depresión y una muestra aleatoria de los que referían no haber recibido un diagnóstico de depresión. A ambos grupos se les invitó a una entrevista con un psiquiatra. El psiquiatra no debía saber la respuesta que el participante había dado en el cuestionario. En esta situación, lo único que se puede obtener son los valores predictivos positivo y negativo. El VPP será el porcentaje de casos confirmados por el psiquiatra entre aquellos que dieron positivo en el cuestionario (dijeron que tenían depresión). El VPN será el porcentaje de personas libres de depresión según el psiquiatra entre aquellos que dijeron que no tenían

depresión. ¿Por qué no se puede conocer con este diseño la sensibilidad ni la especificidad? Porque no se puede estimar la prevalencia real, ya que no se ha examinado por psiquiatra a una muestra representativa de toda la cohorte. La solución puede ser *asumir* una prevalencia realista para la cohorte y, a partir de la prevalencia, reconstruir la tabla 2 × 2 en sentido inverso a como se ha hecho antes. Para entender mejor cómo proceder, se debe hablar de las razones de verosimilitudes.

### 15.10.3. Razón de verosimilitudes (RV)

La probabilidad *a priori* o pretest es la que existe antes de realizar ninguna prueba, coincide conceptualmente con la prevalencia de la enfermedad y es una probabilidad marginal. Suele ser la información inicial disponible en un proceso diagnóstico. En cambio, se denomina probabilidad postest a la que se obtiene tras la obtención de un resultado positivo en la prueba (VPP) y es una probabilidad condicional, como se explicó en el apartado 3.5.

Una prueba diagnóstica persigue que las probabilidades postest superen a las pretest. El cambio desde la probabilidad *a priori* (pretest) a la probabilidad *a posteriori* (postest) indica cuál es la ganancia neta de la prueba diagnóstica en el diagnóstico de la enfermedad.

Otro concepto interesante para valorar una prueba diagnóstica es la razón de probabilidades diagnósticas, comúnmente llamada razón de verosimilitudes (RV), o *likelihood ratio* en inglés, que combina en una sola expresión la sensibilidad y la especificidad. Se corresponden conceptualmente con el factor Bayes que se vio en el apartado 3.7. La razón de verosimilitudes positiva (RVP) compara la probabilidad de que un paciente enfermo presente un resultado positivo en una prueba diagnóstica (sensibilidad) frente a la probabilidad de que se obtenga un resultado positivo en un paciente no enfermo (el complementario de la especificidad, es decir, 1 –especificidad).

$$\text{RVP} = \frac{\text{test + en enfermos}}{\text{test + en sanos}} = \frac{\text{S}}{1 - \text{E}}$$

En el ejemplo de la tabla 15.9, la RVP = 0,784/(1 – 0,957) = 18,2.

La RVP relaciona la *odds* pretest de diagnosticar la enfermedad con la *odds* postest:

*Odds* probabilidad pretest × RPD+ = *Odds* probabilidad postest

La probabilidad pretest es la prevalencia (proporción de enfermos entre la población total) y la probabilidad postest es el valor predictivo positivo (proporción de enfermos entre positivos). La ecuación anterior puede expresarse como:

$$\textit{Odds}\ \text{prevalencia} \times \text{RVP} = \textit{Odds}\ \text{VPP}$$

$$\frac{\text{prevalencia}}{1 - \text{prevalencia}} \times \text{RVP} = \frac{\text{VPP}}{1 - \text{VPP}}$$

En el ejemplo anterior del estudio de validación de la depresión autorreferida en un cuestionario (17), se puede obtener el VPP y el VPN mediante entrevista por psiquiatra, pero no determinar la sensibilidad ni la especificidad. Aun así, es posible reconstruir la tabla 2 × 2, mediante la aplicación primero de la prevalencia estimada y su transformación en su *odds*. Luego se aplica el VPP también trasformado en su *odds*. Finalmente, se calcula la RVP como:

$$\text{RVP} = \frac{\textit{Odds}_{\text{VPP}}}{\textit{Odds}_{\text{Prevalencia}}} = \frac{\text{VPP}/1 - \text{VPP}}{\text{prevalencia}/1 - \text{prevalencia}}$$

Por ejemplo, si se asume una prevalencia del 26,1%, y se encuentra un VPP del 74,2% y un VPN del 81%, la RVP será:

$$RVP = \frac{\left(\frac{0,742}{1-0,742}\right)}{\left(\frac{0,261}{1-0,261}\right)} = 8,14$$

Esto permite saber que la sensibilidad es 8,14 veces superior a 1 – especificidad. Es decir:

$$RVP = S/(1-E) = 8,14$$

A partir de aquí se puede despejar la única combinación de valores de sensibilidad y especificidad que conduciría a VPN = 0,81 y VPP = 0,742. Esta única combinación corresponde a sensibilidad = 0,37 y especificidad = 0,955.

Volviendo al concepto de RVP, debe considerarse que una teórica RVP = 1 significaría que la *odds* postest sería idéntica a la *odds* pretest, es decir, la prueba no aportaría nada nuevo, porque no representaría ninguna ganancia sobre el grado de incertidumbre previo.

En cambio, cuando la RVP > 1, existe una ganancia de información, pues se incrementaría la probabilidad de que la enfermedad esté presente cuando la prueba resulta positiva, y tanto más cuanto más se aleje de 1 su valor. Una RVP < 1 reduce la probabilidad de que la enfermedad esté presente, y habría que interpretarla al revés: un resultado negativo aumentaría la probabilidad de estar enfermo, y viceversa.

La relación entre probabilidad pretest y postest medida por la RVP permite analizar algunos aspectos de la rentabilidad o ganancia diagnóstica de una prueba. Así, cuando la probabilidad de padecer la enfermedad sea muy baja (baja probabilidad pretest), y se pide una prueba, aunque la prueba sea muy buena, con un alto cociente, por ejemplo 10, entre sensibilidad y 1 – especificidad (RVP > 10), la capacidad de diagnóstico no dejará de ser pobre. Supóngase una población de embarazadas en la que, *a priori*, se sepa que solo 1 de cada 3.000 de ellas presentará un hijo con síndrome de Down. Un valor RVP de 30 (excelente) producirá una probabilidad postest muy baja, de solo 1/100, es decir, menos del 1% de los test que den positivos serán verdad. Los demás serán falsos positivos, solo habrá 1 verdadero positivo por cada 100 falsos positivos:

$$\frac{VPP}{1-VPP} = \left(\frac{1}{3.000}\right)(30) = \frac{1}{100}$$

En su concepto clásico, la razón de verosimilitudes negativa (RVN) es el cociente del complementario de la sensibilidad entre la especificidad. Estima el valor por el cual se multiplica la *odds* de estar enfermo cuando el resultado de la prueba sea negativo.

$$RVN = \frac{prob(test-enfermos)}{prob(test-sanos)} = \frac{1-sensibilidad}{especificidad}$$

La RVN valora la contribución que realiza un resultado negativo en la NO confirmación de la enfermedad. Es un concepto que resulta difícil de entender, ya que incluye dos negaciones (negatividad y no confirmación). Por otra parte, se mueve en una escala inversa a la de la RVP, porque es más importante cuanto más se acerca a 0, dado que busca que haya pocos resultados negativos en enfermos. Por esta razón, no es directamente comparable con la RVP.

Para finalizar este apartado, hay que indicar que la prevalencia influye de algún modo en las RV. Las simulaciones indican que su grado de afectación es bastante mayor que sobre la

sensibilidad y la especificidad. Una mayor prevalencia motiva que la RVP descienda y que aumente la RVN.

## 15.11. DISCRIMINACIÓN DIAGNÓSTICA Y PRONÓSTICA: CURVAS ROC

Cuando la variable que se utiliza para clasificar a una persona como «enferma» o «no enferma» es cuantitativa y continua, como la puntuación global en una prueba psicológica, el nivel de colesterol o la glucemia basal, es posible utilizar diferentes puntos de corte para caracterizar al sujeto como enfermo. La figura 15.4 representa la puntuación en un cuestionario (test de actitudes alimentarias, *eating attitudes test* o EAT) para detectar trastorno del comportamiento alimentario (TCA).

Se aprecia que tanto las personas de la población sin este trastorno como las que sí tienen esta patología pueden presentar un rango variable de valores en el test EAT. Los valores en las personas sanas y en las enfermas siguen distribuciones diferentes, donde las puntuaciones del test tienden a ser mayores en enfermos (curva de la derecha) que en personas libres de esta enfermedad (curva de la izquierda). La distribución de personas con TCA es menor (curva más plana) que la de personas sanas, porque hay menos enfermos que no enfermos en la población.

Sin embargo, las dos distribuciones se solapan, es decir, hay personas con puntuaciones de EAT por encima del punto de corte que no presentan TCA, mientras que, por el contrario, también hay personas con puntuaciones por debajo del mismo punto de corte que padecen este trastorno psiquiátrico. Los primeros serán falsos positivos, y los segundos, falsos negativos. Obsérvese que cuando se desplaza el punto de corte hacia la derecha, es decir, se mueve la definición de TCA hacia puntuaciones altas del test, disminuye la probabilidad de hallar un falso positivo y aumenta así la especificidad de la prueba (a costa de reducir su sensibilidad). Por el contrario, cuando se desplaza el punto de corte hacia puntuaciones bajas en el EAT, desciende la probabilidad de hallar falsos negativos y, con ello, se incrementa la sensibilidad de la prueba (pero disminuye su especificidad). En cualquier prueba diagnóstica basada en una prueba que dé un resultado cuantitativo,

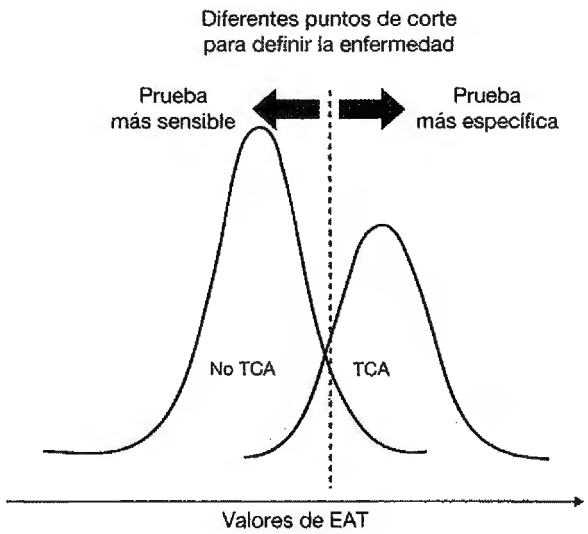



**Figura 15.4** Cambios de la sensibilidad y de la especificidad con diferentes puntos de corte para definir la enfermedad.

el aumento de sensibilidad obtenido al bajar el punto de corte siempre se hace en detrimento de la especificidad, y viceversa.

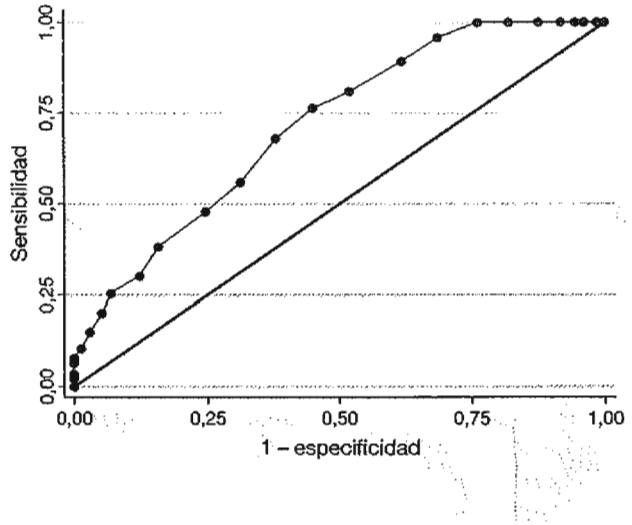
Las relaciones observadas en la figura 15.4 permiten ayudar a comprender mejor la utilidad de la sensibilidad y la especificidad. Cuando se utiliza como criterio diagnóstico un punto de corte más alto de TCA, la prueba será muy *específica,* hay pocos falsos positivos frente a los verdaderos positivos. De ahí que en los libros de texto se afirme que las pruebas específicas sirven para *confirmar* la enfermedad (aunque, en realidad, la especificidad es un criterio que se mide en los no enfermos exclusivamente). Por el contrario, cuando se usa una puntuación baja de TCA como punto de corte para establecer un diagnóstico, el criterio sería poco específico (muchos falsos positivos) y la prueba resultaría muy sensible (la mayoría de los enfermos serán verdaderos positivos); no obstante, por esa abundancia de falsos positivos no se está seguro de si el individuo está enfermo o no, y bajaría el valor predictivo positivo. Se obtiene así una idea de por qué, con frecuencia, se utilizan de manera combinada una prueba sensible al inicio del proceso diagnóstico y, como confirmación, una prueba específica en los que han dado positivo en la primera. Es la ventaja de usar las pruebas en serie (no se hace la segunda prueba más específica hasta que no se sabe que la primera, más sensible, dio positiva), en vez de usarlas en paralelo, pidiéndolas todas a la vez.

La relación entre sensibilidad y especificidad mostrada en la figura 15.4 se puede representar gráficamente mediante una curva ROC (*Receiver Operating Characteristic,* en inglés). Esta curva resume los valores de sensibilidad y especificidad que se obtienen al establecer todos los posibles puntos de corte cuando se evalúa una prueba diagnóstica que depende del punto de corte que se elija en una escala continua. Es una manera de resumir la información que se obtendría con todos los posibles puntos de corte de una prueba diagnóstica con resultados cuantitativos. En la curva ROC, la sensibilidad se representa en el eje de ordenadas y el complementario de la especificidad (1 – E), en el de abscisas.

Si se varía la puntuación del test EAT que se usa como criterio para el diagnóstico de trastorno del comportamiento alimentario, se puede calcular la especificidad y la sensibilidad para cada uno de los posibles puntos de corte (5, 10, 15, 20, 30, etc.). Con estos valores de S y 1 – E se construye la curva ROC, que se forma al unir los diferentes valores de cada punto de corte (fig. 15.5).

La curva ROC permite ver la combinación de sensibilidad y especificidad para los distintos puntos de corte de la prueba diagnóstica, y transmite una idea global de la capacidad de discriminación del test. El área bajo la curva ROC equivale a la probabilidad de que, si se eligiesen al azar a dos individuos, uno con la enfermedad y otro sin ella, la prueba los clasifique a ambos correctamente. Por lo tanto, el área bajo la curva ROC (*area under the curve,* AUC) estima la capacidad de distinguir o de *discriminar* entre enfermos y no enfermos que tiene una prueba diagnóstica. La prueba diagnóstica tendrá mayor capacidad de discriminación cuanto más próxima a 1 sea el área bajo la curva. Si esta área tuviese su valor máximo posible, que es 1 (100%), la prueba sería perfecta, ya que clasificaría al 100% de los enfermos como enfermos y al 100% de los exentos de enfermedad como libres de la dolencia.

En cambio, si el área bajo la curva valiese 0,5 (50%, área debajo de la diagonal representada en la figura 15.5), existiría la misma probabilidad de clasificar a un enfermo como exento de enfermedad que como enfermo, y la prueba sería absolutamente inútil. Un área de 0,5 bajo la curva equivale a no discriminar, se interpreta como una prueba *no informativa.* Una de la posibles áreas de 0,5 (debajo de la línea diagonal mencionada) sería, de hecho, aquella que tuviera unos valores de sensibilidad exactamente iguales a 1 – especificidad en todos los posibles puntos de corte. Esto significaría que la probabilidad de que la prueba diese positiva sería la misma en enfermos que en no enfermos. En estas circunstancias, se dice que la prueba es *«inútil»,* ya que no reduce en nada




| EAT_40 | TCA 0 | TCA 1 | Total |
|---|---|---|---|
| 0 | 6 | 0 | 6 |
| 1 | 3 | 0 | 3 |
| 2 | 12 | 0 | 12 |
| 3 | 8 | 0 | 8 |
| 4 | 15 | 0 | 15 |
| 5 | 23 | 0 | 23 |
| 6 | 30 | 0 | 30 |
| 7 | 31 | 0 | 31 |
| 8 | 40 | 11 | 51 |
| 9 | 37 | 17 | 54 |
| 10 | 53 | 22 | 75 |
| 11 | 37 | 12 | 49 |
| 12 | 38 | 22 | 60 |
| 13 | 36 | 32 | 68 |
| 14 | 35 | 21 | 56 |
| 15 | 48 | 25 | 73 |
| 16 | 20 | 21 | 41 |
| 17 | 28 | 12 | 40 |
| 18 | 10 | 15 | 25 |
| 19 | 12 | 13 | 25 |
| 20 | 9 | 12 | 21 |
| 21 | 7 | 7 | 14 |
| 22 | 0 | 3 | 3 |
| 23 | 0 | 8 | 8 |
| 24 | 0 | 3 | 3 |
| 25 | 0 | 4 | 4 |
| 26 | 0 | 1 | 1 |
| 31 | 0 | 1 | 1 |
| Total | 538 | 262 | 800 |

**Figura 15.5** Curva ROC, junto con la tabulación de los datos (ficticios) que la originaron. La variable de clasificación es la puntuación en el test EAT_40; la variable de resultado es el diagnóstico clínico de trastornos de la conducta alimentaria (1 = diagnóstico; 0 = exento).

el grado de incertidumbre previo acerca de la presencia de enfermedad. Es costumbre representar la diagonal, porque así se aprecia más claramente en cuánto supera la prueba que se valora a lo que sería una prueba no informativa.

Un área menor de 0,5 requiere invertir los criterios de diagnóstico, es decir, considerar los negativos como positivos, y viceversa. Al invertirlos, se comprueba que la prueba discrimina más del 50%.

En modelos multivariables de regresión logística se puede obtener una curva ROC en la que la variable predictora corresponde a los valores de probabilidad predichos por el modelo para cada sujeto en función de sus variables independientes. La variable criterio será el desenlace dicotómico observado. Lo ideal es que, en los sujetos que tuvieron el desenlace, la probabilidad predicha sea superior a los sujetos que no presentaron el desenlace. Mediante el cálculo del área bajo la curva, estas curvas ROC informan del grado de discriminación del modelo.

La construcción de curvas ROC permite, además, la comparación de varias pruebas diagnósticas o de varios modelos multivariables. Dado un conjunto de pruebas, la que produzca una curva ROC con mayor área bajo la curva será la de elección, debido a su superior capacidad de discriminación. A modo de ejemplo, la tabla 15.11 presenta la comparación de áreas bajo la curva ROC de distintos índices para detectar el síndrome metabólico en niños (**18**). De las tres pruebas evaluadas, el perímetro de cintura discriminará mejor para detectar el síndrome metabólico en

*Tabla 15.11 Capacidad predictiva de diferentes pruebas antropométricas en la detección de síndrome metabólico en niños*

| PRUEBA | ÁREA BAJO LA CURVA ROC | IC 95% |
|---|---|---|
| Índice de masa corporal | 0,849 | 0,780-0,919 |
| Perímetro de cintura | 0,868 | 0,801-0,934 |
| Razón pliegue tríceps/subescapular | 0,834 | 0,757-0,910 |

niños, con una capacidad de discriminación del 86,8%; sin embargo, habría que valorar si las diferencias son estadísticamente significativas (v. más adelante).

## 15.12. COMPARACIÓN DE CURVAS ROC

Como se ha indicado en el apartado anterior, es posible comparar dos curvas ROC valorando la hipótesis nula de que las áreas bajo ambas curvas ROC sean idénticas (**19**). Imagínese que se desea predecir el riesgo de que los participantes incluidos en un estudio desarrollen enfermedad cardiovascular. Para realizar dicha predicción, se emplean dos ecuaciones diferentes. En una de ellas se incluyen solo factores de riesgo cardiovascular tradicionales y en la otra se introduce también el resultado de un nuevo biomarcador que se está evaluando. Se obtendrán modelos de regresión logística con la variable que indica quiénes desarrollan enfermedad cardiovascular como variable dependiente (p. ej., *eventocv*) y los factores de riesgo tradicionales como variables independientes en un primer modelo y con los factores de riesgo tradicionales y el nuevo biomarcador como variables independientes en un segundo modelo. Tras definir cada uno de estos modelos, se solicitará al *software* estadístico que guarde los valores predichos para todos los sujetos en función del modelo que se acaba de definir. Si se emplea STATA, esto se podrá obtener ejecutando la orden **predict** inmediatamente después de ejecutar el modelo mediante la orden **logistic** e indicando el nombre que se desea asignar a la nueva variable que contendrá la información con los valores predichos. Si se denominan las variables con los valores predichos como *pre_tradic* y *pre_biomarc*, se realizará esta comparación con la orden de STATA:

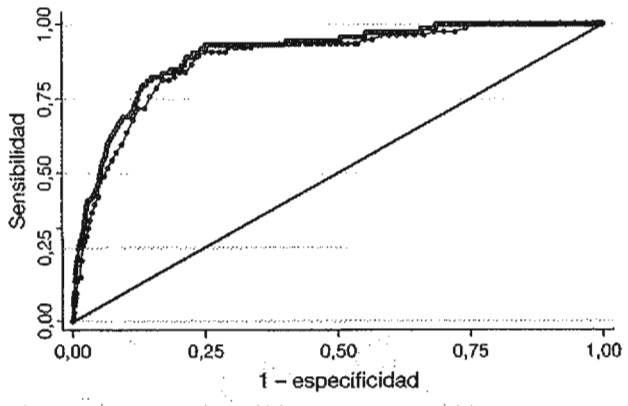
```
roccomp eventocv pre_biomarc pre_tradic, graph summary
```

El resultado que se obtendrá es el que se muestra en la figura 15.6. El valor *p* para la comparación del área bajo ambas curvas es estadísticamente significativo, por lo que se concluirá que ambas áreas bajo la curva son distintas y el biomarcador aporta una información predictiva que va más allá de la aportada por los factores de riesgo clásicos.

## 15.13. ÍNDICE *C* DE HARRELL PARA PREDICCIONES EN ANÁLISIS DE SUPERVIVENCIA

El índice *C* de Harrell es una medida de discriminación en los modelos de supervivencia (**20,21**). Se trata de una extensión para los modelos de supervivencia del área bajo la curva ROC que se puede calcular en un modelo de regresión logística.

En la tabla 15.12 se presentan los datos ficticios de un estudio en el que se ha recabado, entre 40 pacientes, información sobre si fumaban o no (*fum* = 0 para no fumadores; *fum* = 1 para fumadores), el tiempo durante el que se les ha seguido *(tiempo)* y si habían fallecido o no al final de ese tiempo (*d* = 0 si estaba vivo; *d* = 1 si había fallecido). Si se calculase un modelo de regresión de Cox con estos datos, se obtendría una *hazard ratio* de mortalidad de 3,10 (0,63-15,38) para fumadores comparados con no fumadores. Si bien las diferencias no resultan estadísticamente significativas, el hecho de fumar sugiere un mayor riesgo de mortalidad. Por




| | Obs | ROC Area | Std. Err. | Asymptotic Normal [95% Conf. | Interval] |
|---|---|---|---|---|---|
| pred_biomarc | 15550 | 0.8995 | 0.0167 | 0.86664 | 0.93228 |
| pred_tradic | 15550 | 0.8801 | 0.0192 | 0.84243 | 0.91772 |

Ho: area(pred_biomarc) = area(pred_tradic)
chi2(1) = 5.04 Prob>chi2 = 0.0248

**Figura 15.6** Comparación de dos curvas ROC.

ello, el modelo predice un mayor riesgo de mortalidad para los fumadores que para los no fumadores (riesgo de mortalidad *predicho* en fumadores > riesgo de mortalidad *predicho* en no fumadores).

Se aprecia que se han producido eventos de interés en los tiempos 1 y 5, por lo que en cada uno de estos tiempos se podría formar una tabla de contingencia con los datos *observados* de hábito tabáquico y de supervivencia entre los 40 pacientes incluidos en el análisis (tabla 15.13).

Para calcular el índice *C* de Harrell, la comparación se ha de realizar para cada tiempo y para cada posible pareja compuesta, al comparar cada uno que fallece con cada sujeto que sigue vivo si el valor de la función de Cox $(EXP\,(\lambda_{0t} + b_1x_1 + \ldots\, b_px_p))$ del fallecido supera a la del no fallecido en ese tiempo, del mismo modo que se procedía con las comparaciones pareja a pareja para la *U* de Mann-Whitney (v. apartado 6.7).

El índice *C* de Harrell se define como la proporción de parejas que surgen de la comparación entre sujetos que fallecen y quienes siguen vivos que concuerdan en sus valores *predichos* y *observados*. Así, en el Tiempo = 1 (v. tabla 15.13) hay tres sujetos que fallecen y 37 que continúan vivos, por lo que hay $3 \times 37 = 111$ posibles comparaciones por parejas. Hay 19 sujetos que no fuman (con un menor riesgo *predicho*) y que sobreviven al Tiempo 1 y 2 que fuman (con un mayor riesgo *predicho*) y que no sobreviven al Tiempo 1; a partir de estos datos, se puede calcular que hay $19 \times 2 = 38$ comparaciones en las que los valores *observado* y *predicho* coinciden, pues el fallecido riene más riesgo que el no fallecido (parejas *concordantes*). Se observa también que hay $18 \times 1 = 18$ parejas *discordantes* para las cuales el riesgo *observado* es mayor para los no fumadores, si bien tienen un menor riesgo *predicho*. En el resto de las parejas $(1 \times 19 + 2 \times 18 = 55)$ se daría un empate en el riesgo predicho.

*Tabla 15.12 Ejemplo de datos de supervivencia*

| ID | TIEMPO | D | FUM |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 |
| 4 | 5 | 0 | 0 |
| 5 | 5 | 0 | 0 |
| 6 | 5 | 0 | 0 |
| 7 | 5 | 0 | 0 |
| 8 | 5 | 0 | 0 |
| 9 | 5 | 0 | 0 |
| 10 | 5 | 0 | 0 |
| 11 | 5 | 0 | 0 |
| 12 | 5 | 0 | 0 |
| 13 | 5 | 0 | 0 |
| 14 | 5 | 0 | 0 |
| 15 | 5 | 0 | 0 |
| 16 | 5 | 0 | 0 |
| 17 | 5 | 0 | 0 |
| 18 | 5 | 0 | 0 |
| 19 | 5 | 0 | 0 |
| 20 | 5 | 0 | 0 |
| 21 | 5 | 0 | 0 |
| 22 | 5 | 0 | 1 |
| 23 | 5 | 0 | 1 |
| 24 | 5 | 0 | 1 |
| 25 | 5 | 0 | 1 |
| 26 | 5 | 0 | 1 |
| 27 | 5 | 0 | 1 |
| 28 | 5 | 0 | 1 |
| 29 | 5 | 0 | 1 |
| 30 | 5 | 0 | 1 |
| 31 | 5 | 0 | 1 |
| 32 | 5 | 0 | 1 |
| 33 | 5 | 0 | 1 |
| 34 | 5 | 0 | 1 |
| 35 | 5 | 0 | 1 |
| 36 | 5 | 1 | 0 |
| 37 | 5 | 1 | 1 |
| 38 | 5 | 1 | 1 |
| 39 | 5 | 1 | 1 |
| 40 | 5 | 1 | 1 |

*Tabla 15.13 Comparación del riesgo de mortalidad a lo largo del seguimiento entre fumadores y no fumadores*

| | FALLECEN | NO FALLECEN |
|---|---|---|
| **Tiempo = 1** | | |
| No fumadores | 1 | 19 |
| Fumadores | 2 | 18 |
| Total | 3 | 37 |
| **Tiempo = 5** | | |
| No fumadores | 1 | 18 |
| Fumadores | 4 | 14 |
| Total | 5 | 32 |

De manera análoga, en el Tiempo 5 se pueden contabilizar 5 × 32 = 160 parejas. De ellas, 18 × 4 = 72 serían concordantes, ya que los fumadores presentarían un mayor riesgo observado y predicho que los no fumadores. Se observan también 1 × 14 parejas discordantes, que serían parejas en las que los no fumadores tendrían un mayor riesgo observado, pero un menor riesgo predicho que los fumadores. Finalmente, hay 18 × 1 + 14 × 4 = 74 parejas en las que el riesgo predicho coincidiría.

Si se considerase ahora el total de comparaciones que se pueden realizar a lo largo del tiempo (uniendo los datos para el Tiempo 1 y para el Tiempo 5):

- Número de parejas totales: 111 + 160 = 271.
- Número de parejas que concuerdan: 38 + 72 = 110.
- Número de parejas discordantes: 18 + 14 = 32.
- Número de parejas que empatan: 55 + 74 = 129.

Al igual que ocurre en otros estimadores estadísticos como la *U* de Mann-Whitney, los empates se han de asignar por igual entre los grupos que se comparan. Por ello, en este caso se asignarán por igual las parejas empatadas al grupo de parejas que concuerdan y al grupo de parejas discordantes. Así, el índice *C* de Harrell se calculará como:

$$C = (110 + 129 / 2) / 271 = 0{,}6439$$

Este índice se puede determinar con STATA tras pedir el modelo de Cox correspondiente. Para obtener el índice *C*, se usará la siguiente orden postestimación:

```
estat concordance, harrell
```

que producirá el siguiente resultado:

```
. estat concordance, harrell

        failure _d:  d
  analysis time _t:  tiempo

Harrell's C concordance statistic

Number of subjects (N)                =        40
Number of comparison pairs (P)        =       271
Number of orderings as expected (E)   =       110
Number of tied predictions (T)        =       129

         Harrell's C = (E + T/2) / P  =    .6439
                             Somers' D =    .2878
```

## 15.14. ÍNDICE NETO DE RECLASIFICACIÓN, CAPACIDAD DE ESTRATIFICACIÓN Y OTROS ÍNDICES DE DISCRIMINACIÓN

Estas técnicas se suelen emplear a la hora de valorar si la introducción de una nueva prueba diagnóstica mejora la información para predecir un evento de interés con respecto a la información de la que ya se dispone. Así, se podría plantear si resulta beneficioso añadir la información sobre un nuevo biomarcador a la ecuación de riesgo de Framingham para predecir el riesgo cardiovascular[1].

1 La ecuación de Framingham es el modelo matemático más conocido de predicción de riesgo. Tiene en cuenta diversos factores cardiovasculares para producir una estimación de cuál será el riesgo absoluto de que esa persona padezca un evento coronario en los próximos 10 años según su sexo, edad, hábito tabáquico, presión arterial y lípidos. Ver por ejemplo: http://www.globalrph.com/atp_calc.htm y http://cvdrisk.nhlbi.nih.gov/calculator.asp.

En la evaluación de modelos para predecir el riesgo de desarrollar un evento de interés, es importante distinguir entre los conceptos de calibración y de discriminación:

- La *calibración* de un modelo hace referencia a la capacidad de una prueba de predecir el porcentaje de personas que desarrollarán el evento dentro de cada subgrupo de riesgo predicho por el modelo. Se evalúa presentando el porcentaje de sujetos que desarrollan el evento dentro de las categorías de riesgo predichas por el modelo, como hace el test de Hosmer-Lemeshow (v. apartado 13.12) al categorizar a la muestra en deciles predichos de riesgo.
- La *discriminación* hace referencia a la capacidad de un modelo de distinguir entre dos personas, una que desarrollará la enfermedad y otra que no la desarrollará. Si la discriminación es buena, el riesgo predicho para la persona que desarrolla el evento será mayor que el predicho para la persona que no lo desarrollará (área bajo la curva, índice *C*).

Algunos autores afirman que es preferible una buena *discriminación* (**20**). Además, en la práctica clínica, un médico no se suele encontrar ante la situación de distinguir entre dos pacientes, uno con la enfermedad y otro sin la enfermedad, sino ante la situación de tener que valorar el riesgo de desarrollar una enfermedad que tiene un paciente determinado (**22**). A pesar de esto, una buena calibración es un primer paso importante para evaluar un modelo predictivo.

Para valorar de manera integrada la discriminación de un modelo, se ha propuesto la utilización de la mejora neta de la reclasificación y la mejora integrada de la discriminación (**23**).

Para calcular la *mejora neta de la reclasificación*, en primer lugar hay que construir una tabla de contingencia, como la tabla 15.12, en la cual se clasifica a los sujetos en categorías de riesgo predichas por el modelo con los factores de riesgo conocidos y por el modelo con los factores de riesgo conocidos y el biomarcador que se desea evaluar (tabla 15.14).

Se debe tener en cuenta la proporción de sujetos con evento y sin evento cuya probabilidad aumenta (↑) o disminuye (↓) al introducir el nuevo predictor. Así, el biomarcador se considerará que aporta información valiosa si al añadir ese nuevo biomarcador al modelo se consigue:

- Elevar la probabilidad predicha para los sujetos que desarrollan el evento (E = 1). Es decir, $p(\uparrow|E = 1)$ = proporción de personas con evento para las que aumenta el riesgo.
- Reducir la probabilidad predicha para los sujetos sin evento (E = 0). Es decir, $p(\downarrow|E = 0)$ = proporción de personas sin evento para las que disminuye el riesgo.

***Tabla 15.14 Tabla de contingencia en la cual se clasifica a los sujetos en función de categorías de riesgo predichas por el modelo con factores de riesgo conocidos y por el modelo con los factores de riesgo conocidos y el biomarcador que se desea evaluar***

| MODELO DE RIESGO CON FACTORES CONOCIDOS | MODELO DE RIESGO CON FACTORES CONOCIDOS Y NUEVO BIOMARCADOR | | | | TOTAL |
|---|---|---|---|---|---|
| | 0% A <0,1% | 0,1% A <1% | 1% A <2,5% | ≥2,5% | |
| 0% a <0,1% | 6,718 | 258 | 0 | 0 | 6,976 |
| | 0,04 | 0,39 | — | — | 0,06 |
| 0,1% a <1% | 1.681 | 1.569 | 316 | 37 | 6.603 |
| | 0,06 | 0,20 | 1,90 | 2,70 | 0,26 |
| 1% a <2,5% | 0 | 520 | 643 | 240 | 1,403 |
| | — | 1,54 | 1,71 | 2,92 | 1,85 |
| ≥2,5% | 0 | 10 | 150 | 408 | 568 |
| | — | 0,00 | 2,00 | 5,88 | 4,75 |
| Total | 8.399 | 5.357 | 1.109 | 685 | 15.500 |
| | 0,05 | 0,34 | 1,80 | 4,67 | 0,48 |

Por el contrario, se considerará contraproducente si:

- Disminuye la probabilidad predicha para los sujetos con evento (E = 1). Es decir, $p(\downarrow|E=1)$ = proporción de personas con evento para las que disminuye el riesgo.
- Aumenta la probabilidad predicha para los sujetos sin evento (E = 0). Es decir, $p(\uparrow|E=0)$ = proporción de personas sin evento para las que aumenta el riesgo.

Así, la mejora neta de la reclasificación (MNR) se puede definir como:

$$MNR = \left[p(\uparrow|E=1) + p(\downarrow|E=0)\right] - \left[p(\downarrow|E=1) + p(\uparrow|E=0)\right]$$

En el ejemplo de la tabla 15.2, la mejora neta de la reclasificación valdrá:

$$MNR = [(15/74) + (2.349/15.476)] - [(12/74) + (836/15.476)] = 0{,}1383$$

Se ha definido también un contraste asintótico que valora el grado de compatibilidad de los datos con la hipótesis nula de que la mejora neta de la reclasificación sea 0:

$$z = \frac{\text{mejora neta de la reclasificación}}{\sqrt{\dfrac{p(\uparrow|E=1)+p(\downarrow|E=1)}{\text{número de eventos}} + \dfrac{p(\uparrow|E=0)+p(\downarrow|E=0)}{\text{número de no eventos}}}}$$

que, en el ejemplo de la tabla 15.12, valdría 0,1383/0,0703 = 1,967. El valor $p$ que corresponde a una $z$ = 1,967 es 0,049. Por tanto, se rechazaría la hipótesis nula que mantiene que la mejora neta de la reclasificación valiese 0. Una desventaja de la mejora neta de la reclasificación es que depende de los puntos de corte que se hayan escogido para definir las categorías de riesgo.

La *mejora integrada de la discriminación* ofrece la ventaja de no necesitar categorías de riesgo y se basa en los cambios en sensibilidad y en 1 – especificidad en los modelos con y sin el nuevo biomarcador. Equivaldría a considerar una categorización tan detallada que cada persona perteneciese a su propia categoría. La mejora integrada de la discriminación (MID) se puede definir como:

$$MID = \left(\bar{p}_{\text{nuevo, eventos}} - \bar{p}_{\text{tradicional, eventos}}\right) - \left(\bar{p}_{\text{nuevo, NO eventos}} - \bar{p}_{\text{tradicional, NO eventos}}\right)$$

donde:

- $\bar{p}_{\text{nuevo, eventos}}$ : media de las probabilidades predichas de presentar un evento por el modelo que incluye el biomarcador nuevo para los sujetos que sufren un evento.
- $\bar{p}_{\text{tradicional, eventos}}$ : media de las probabilidades predichas de presentar un evento por el modelo que incluye solo los factores de riesgo tradicionales para los sujetos que sufren un evento.
- $\bar{p}_{\text{nuevo, no eventos}}$ : media de las probabilidades predichas de presentar un evento por el modelo que incluye el biomarcador nuevo para los sujetos que no sufren un evento.
- $\bar{p}_{\text{tradicional, no eventos}}$ : media de las probabilidades predichas de presentar un evento por el modelo que incluye solo los factores de riesgo tradicionales para los sujetos que no sufren un evento.

También se ha definido un contraste de hipótesis asintótico para valorar si los datos son compatibles con la hipótesis nula de que la mejora integrada de la discriminación sea 0 (22).

## 15.15. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| Alfa de Cronbach | **alpha** ***lista_de_items*****, detail items** | **RELIABILITY /VAR=*A B C D* /MOD=ALPHA.** |
| Índice kappa | **kap** ***observador1 observador2*** | **CROSSTABS /TAB=*observador1* BY *observador2* /STAT=KAPPA.** |
| Coeficiente de correlación intraclase | **icc** ***variable sujeto observador*** | **RELIABILITY /VAR=*observacion1 observacion2* /MODEL=ALPHA /ICC=MODEL (RANDOM) TYPE (ABSOLUTE) CIN=95.** |
| Gráfico de Bland-Altman | **concord** ***observacion1 observacion2*****, loa** | |
| Coeficiente de correlación de Lin | **concord** ***observacion1 observacion2*** | |
| Comparación de curvas ROC | **roccomp** ***gold_estandar medicion1 medicion2*****, graph summary** | |
| Índice *C* de Harrell | **estat concordance, harrell** | |

## REFERENCIAS


1. Delgado Rodríguez M, Llorca Díaz J, Doménech Massons JM. Estudios para pruebas diagnósticas y factores pronósticos. Barcelona: Signo; 2005. p. 1-187.
2. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144(11):850-5.
3. Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology: how to do clinical practice research. 3rd ed. Philadelphia: Lippincott, Williams and Wilkins; 2006.
4. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Ann Intern Med 2009;151(10):727-37.
5. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11(2):88-94.
6. Bland JM, Altman DG. Cronbach's alpha. BMJ 1997;314(7080):572.
7. Prieto L, Lamarca R, Casado A. La evaluación de la fiabilidad en las observaciones clínicas: el coeficiente de correlación intraclase. Med Clin (Barc) 1998;110(4):142-5.

8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307-10.

9. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45(1):255-68.

10. Lin LIK. A note on the concordance correlation coefficient. Biometrics 2000;56(1):324-5.

11. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. J Clin Chem Clin Biochem 1983;21(11):709-20.

12. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. J Clin Chem Clin Biochem 1984;22(6):431-45.

13. Luiz RR, Costa AJL, Kale PL, Werneck GL. Assessment of agreement of a quantitative variable: a new graphical approach. J Clin Epidemiol 2003;56(10):963-7.

14. Llorca J, Delgado-Rodríguez M. Survival analytical techniques were used to assess agreement of a quantitative variable. J Clin Epidemiol 2005;58(3):314-5.

15. Navarro C, Chirlaque MD, Tormo MJ, Pérez-Flores D, Rodríguez-Barranco M, Sánchez-Villegas A, et al. Validity of self reported diagnoses of cancer in a major Spanish prospective cohort study. J Epidemiol Community Health 2006;60(7):593-9.

16. Honkanen K, Honkanen R, Heikkinen L, Kroger H, Saarikoski S. Validity of self-reports of fractures in perimenopausal women. Am J Epidemiol 1999;150(5):511-6.

17. Sánchez-Villegas A, Schlatter J, Ortuno F, Lahortiga F, Pla J, Benito S, Martínez-González MA. Validity of a self-reported diagnosis of depression among participants in a cohort study using the Structured Clinical Interview for DSM-IV (SCID-I). BMC Psychiatry 2008;8:43.

18. Moreno LA, Pineda I, Rodríguez G, Fleta J, Sarría A, Bueno M. Waist circumference for the screening of the metabolic syndrome in children. Acta Paediatr 2002;91(12):1307-12.

19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1998;44(3):837-45.

20. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.

21. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med 2004;23(13):2109-23.

22. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med 2008;149(10):751-60.

23. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27(2):157-72.

# ANÁLISIS FACTORIAL 16

*A. Sánchez-Villegas, M. Bes-Rastrollo, M. Á. Martínez-González*

## 16.1. INTRODUCCIÓN AL ANÁLISIS FACTORIAL

Imagine el ejemplo de la tabla 16.1 en el que se recogieron datos referentes a 10 variables de riesgo[1], edad *(age)* y sexo *(gender)*. Una puntuación más alta implica una mayor exposición a cada una de las 10 variables de riesgo (alcohol, tabaco, etc.). El objetivo es buscar patrones de asociación de estas variables de riesgo. Esto permitirá definir que hay tipos de personas en los que se daría simultáneamente, por ejemplo, una mayor exposición al alcohol, el cannabis y la promiscuidad sexual, mientras que otros tienden a exponerse más al sedentarismo, el tabaco y la dieta insana. Se busca reducir esta información de las 10 variables de riesgo a la que pueden proporcionar solo dos o tres factores que sean combinaciones lineales de ellas.

### 16.1.1. Características y objetivos del análisis factorial

El análisis factorial no establece dependencias entre variables ni hace contrastes de hipótesis. Su propósito es identificar una serie de ***factores subyacentes*** (muy pocos) en esta matriz de datos. En el ejemplo presentado, se podrían someter las 10 variables de riesgo (las columnas desde *alcohol* a *partner*) a un análisis factorial.

El análisis comenzaría con un análisis de la estructura de las correlaciones entre estas 10 variables y, a través de la reducción de datos, determinaría un número pequeño de nuevos factores («componentes» principales) que resumen las 10 variables originales. Cada uno de los nuevos factores o componentes será una combinación lineal de las 10 variables iniciales, entre las cuales unas reciben más peso que otras, y puede representar una dimensión específica subyacente (1). Existen varios enfoques. Los más importantes son el análisis factorial de componentes principales (AFCP, *principal component factor method*) y el análisis factorial común (AFC, *common factor analysis*), que serán tratados con una finalidad *exploratoria*[2]. Dentro del análisis factorial común existen diferentes aproximaciones, como el análisis factorial principal *(principal factor method)*, el método factorial de máxima verosimilitud *(maximum-likelihood factor method)* o los análisis alfa y los análisis imagen *(alpha factoring, image factoring)*.

Este capítulo se centrará, principalmente, en el manejo e interpretación de los resultados obtenidos para AFCP a través del programa STATA. Se empleará tanto la opción aportada desde el menú del programa como el uso de instrucciones a través de sintaxis. Se hará especial hincapié en el AFCP, al ser un análisis más simple y, por tanto, más sencillo de entender que otras aproximaciones.

---

1 *smoke* = tabaco, *junkfood* = frecuencia con que se come en restaurantes tipo *fast-food*, *soda_dr* = *soda drinks*, es decir, refrescos azucarados, *sedent_w* = *sedentary work*, índice de sedentarismo en el trabajo, *sedent_l* = *sedentary leisure-time*, índice de sedentarismo en el tiempo libre, *riskdriv* = conducción arriesgada, *celldriv* = uso de teléfono móvil mientras se conduce, *partner* = número de parejas sexuales en el último año. La *edad* está en años. La variable *gender* es el sexo (véase la nota al respecto en el capítulo 2, donde se sugiere que es mejor usar *sex* que *gender*), 1 = varón; 2 = mujer. Puede descargarse esta base de datos desde la página web del Departamento de Medicina Preventiva (www.unav.es/preventiva), dentro de Docencia, en la sección Bioestadística.

2 Otra finalidad sería la de usarlos como método *confirmatorio* (v. apartado 16.10).

*Tabla 16.1 Base de datos empleada como ejemplo en el capítulo*

| ID | ALCOHOL | SMOKE | CANNABIS | JUNKFOOD | SODA_DR | SEDENT_W | SEDENT_L | RISKDRIV | CELLDRIV | PARTNER | AGE | GENDER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 0 | 2 | 1 | 3 | 2 | 1 | 1 | 1 | 52 | 1 |
| 2 | 5 | 2 | 2 | 3 | 5 | 4 | 3 | 7 | 10 | 1 | 39 | 2 |
| 3 | 4 | 1 | 0 | 2 | 1 | 0 | 6 | 4 | 0 | 1 | 34 | 1 |
| 4 | 4 | 1 | 0 | 5 | 5 | 0 | 2 | 5 | 10 | 1 | 36 | 1 |
| 5 | 0 | 0 | 0 | 3 | 0 | 1 | 4 | 0 | 0 | 0 | 32 | 1 |
| 6 | 0 | 0 | 0 | 2 | 1 | 7 | 2 | 0 | 3 | 1 | 43 | 1 |
| 7 | 4 | 4 | 0 | 3 | 2 | 8 | 9 | 2 | 2 | 0 | 38 | 1 |
| 8 | 10 | 8 | 4 | 9 | 10 | 7 | 3 | 8 | 7 | 2 | 20 | 1 |
| 9 | 4 | 3 | 1 | 2 | 0 | 9 | 2 | 2 | 1 | 0 | 44 | 1 |
| 10 | 4 | 3 | 1 | 3 | 2 | 5 | 0 | 5 | 1 | 3 | 23 | 1 |
| 11 | 4 | 5 | 1 | 4 | 5 | 5 | 2 | 7 | 2 | 1 | 50 | 1 |
| 12 | 2 | 0 | 0 | 3 | 2 | 2 | 4 | 2 | 3 | 1 | 32 | 1 |
| 13 | 9 | 10 | 6 | 10 | 10 | 10 | 3 | 6 | 10 | 2 | 27 | 1 |
| 14 | 2 | 0 | 0 | 5 | 1 | 2 | 5 | 2 | 3 | 1 | 52 | 1 |
| 15 | 4 | 5 | 1 | 4 | 4 | 3 | 2 | 2 | 9 | 4 | 25 | 1 |
| 16 | 0 | 0 | 0 | 1 | 2 | 0 | 3 | 0 | 0 | 1 | 38 | 1 |
| 17 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 2 | 1 | 47 | 1 |
| 18 | 8 | 9 | 7 | 2 | 0 | 1 | 1 | 9 | 7 | 2 | 36 | 1 |
| 19 | 7 | 9 | 7 | 6 | 10 | 0 | 4 | 8 | 10 | 2 | 23 | 1 |
| 20 | 7 | 10 | 9 | 9 | 10 | 0 | 10 | 6 | 5 | 8 | 39 | 1 |
| 21 | 5 | 2 | 2 | 2 | 0 | 0 | 0 | 9 | 7 | 2 | 24 | 2 |
| 22 | 2 | 0 | 0 | 2 | 1 | 2 | 3 | 1 | 0 | 1 | 45 | 2 |
| 23 | 10 | 8 | 4 | 6 | 8 | 8 | 6 | 6 | 5 | 1 | 31 | 2 |
| 24 | 4 | 6 | 1 | 9 | 10 | 3 | 7 | 8 | 8 | 3 | 20 | 2 |
| 25 | 5 | 4 | 4 | 4 | 4 | 1 | 8 | 7 | 1 | 1 | 46 | 2 |
| 26 | 9 | 9 | 10 | 5 | 7 | 10 | 9 | 10 | 10 | 2 | 24 | 2 |
| 27 | 4 | 5 | 1 | 1 | 3 | 2 | 3 | 3 | 0 | 1 | 43 | 2 |
| 28 | 4 | 5 | 1 | 5 | 1 | 10 | 8 | 4 | 0 | 1 | 45 | 2 |
| 29 | 7 | 0 | 0 | 3 | 1 | 9 | 2 | 8 | 10 | 1 | 30 | 2 |
| 30 | 2 | 0 | 0 | 3 | 2 | 2 | 7 | 3 | 2 | 1 | 33 | 2 |
| 31 | 7 | 8 | 8 | 7 | 8 | 8 | 9 | 9 | 10 | 3 | 40 | 2 |
| 32 | 8 | 9 | 7 | 1 | 0 | 0 | 0 | 7 | 6 | 2 | 26 | 2 |
| 33 | 1 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | 1 | 1 | 47 | 2 |
| 34 | 4 | 0 | 0 | 3 | 2 | 2 | 3 | 3 | 5 | 1 | 23 | 2 |
| 35 | 3 | 4 | 1 | 4 | 5 | 9 | 6 | 3 | 0 | 1 | 44 | 2 |
| 36 | 4 | 2 | 2 | 3 | 2 | 7 | 3 | 2 | 3 | 1 | 46 | 2 |
| 37 | 2 | 2 | 0 | 4 | 5 | 7 | 4 | 2 | 4 | 1 | 58 | 2 |
| 38 | 8 | 7 | 5 | 4 | 4 | 7 | 5 | 7 | 4 | 2 | 35 | 2 |
| 39 | 9 | 9 | 6 | 9 | 10 | 0 | 7 | 6 | 5 | 2 | 22 | 2 |
| 40 | 1 | 0 | 0 | 3 | 1 | 7 | 4 | 0 | 1 | 1 | 52 | 2 |

### 16.1.2. Estandarización y obtención de valores *z*

El primer paso del análisis factorial consiste en estandarizar o tipificar las variables originales, restándole a cada dato su media y dividiéndolo por la desviación estándar (valores *z*). Esta estandarización de cada variable conduce a la obtención de una nueva variable *z* con media 0 y varianza 1. Esto presenta muchas ventajas, como se verá después.

$$z_{alcohol} = \frac{alcohol - \bar{x}}{s_{alcohol}}, \text{para el sujeto con id} = 1, z_{alcohol} = \frac{2 - 4,47}{2,92} = -0,85$$

El ordenador sustituye el original (*alcohol* = 2) por su valor *z* ($z_{alcohol} = -0,85$). Este nuevo valor indica a cuántas desviaciones estándar se encuentra ese sujeto de la media de los 40 sujetos. Esta operación se repite para los 400 datos ($n = 40 \times 10$ variables = 400 datos). A partir de este

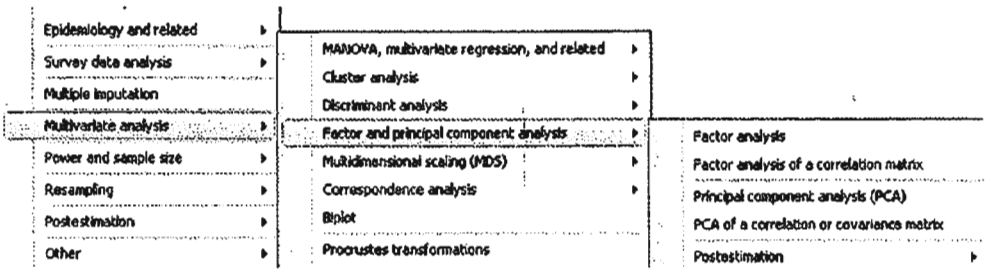
momento, para el análisis factorial ya no existen las 10 variables originales, sino sus correspondientes valores *z*. La ventaja es que ahora cada una de las 10 variables tiene una varianza de 1. Por tanto, la *varianza total* ($h^2$) del conjunto de las 10 variables valdrá 10 (2):

$$h^2 = var(z_{alcohol}) + var(z_{smoke}) + var(z_{cannabis}) + var(z_{junkfood}) + var(z_{soda\_dr}) + \\ + var(z_{sedent\_w}) + var(z_{sedent\_l}) + var(z_{riskdriv}) + var(z_{celldriv}) + var(z_{partner}) =$$
$$h^2 = 1+1+1+1+1+1+1+1+1+1 = 10$$

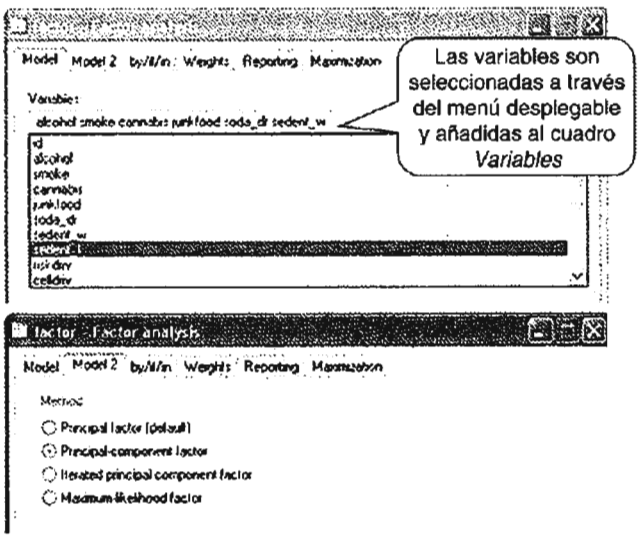
### 16.1.3. Extracción de factores

La extracción de factores es un proceso iterativo que consiste en refinar poco a poco la matriz de correlaciones entre variables de forma repetida para obtener los diferentes factores subyacentes a los datos. En STATA, puede llevarse a cabo a través del menú:

**Statistics → Multivariate analysis → Factor and Principal components analysis → Factor analysis**




El programa permite realizar diferentes tipos de análisis factorial, entre los que se incluyen el AFCP y algunas modalidades de AFC, como el análisis factorial principal (el que STATA realiza por defecto) o el método de máxima verosimilitud. En el ejemplo anterior se realizará un AFCP. El programa presenta un submenú con varias lengüetas. La primera de ellas, *Model*, permite añadir las variables que se desea incluir en el análisis. Su selección se lleva a cabo a través de un menú despegable. La segunda lengüeta, *Model 2*, permite elegir el método de extracción de factores.



El análisis puede ser realizado de igual forma mediante la instrucción:

```
factor
```

Por defecto, STATA realiza un análisis factorial principal, con la subinstrucción:

```
pf
```

Para efectuar un AFCP, debe escribirse la subinstrucción:

```
pcf
```

Así, habría que escribir:

```
factor v1 v2 v3 vp, pcf
```

siendo $v_1$-$v_p$ las variables implicadas en el análisis.

En el ejemplo:

```
fac alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner, pcf
```

STATA presenta dos tablas. A continuación se interpretará el significado de ambas. La primera presenta la existencia de 10 factores (tantos como variables) que serán capaces de explicar el 100% de la variabilidad ($h^2$) de los 400 datos (40 sujetos × 10 variables).

```
Factor analysis/correlation                    Number of obs    =         40
    Method: principal-component factors        Retained factors =          3
    Rotation: (unrotated)                      Number of params =         27
```

| Factor | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 5.43802 | 4.12459 | 0.5438 | 0.5438 |
| Factor2 | 1.31343 | 0.17829 | 0.1313 | 0.6751 |
| Factor3 | 1.13514 | 0.42531 | 0.1135 | 0.7887 |
| Factor4 | 0.70982 | 0.20214 | 0.0710 | 0.8596 |
| Factor5 | 0.50768 | 0.11342 | 0.0508 | 0.9104 |
| Factor6 | 0.39427 | 0.16585 | 0.0394 | 0.9498 |
| Factor7 | 0.22842 | 0.11002 | 0.0228 | 0.9727 |
| Factor8 | 0.11840 | 0.03176 | 0.0118 | 0.9845 |
| Factor9 | 0.08663 | 0.01844 | 0.0087 | 0.9932 |
| Factor10 | 0.06820 | . | 0.0068 | 1.0000 |

El primer factor (o componente subyacente) es capaz de explicar una cantidad de varianza igual a 5,438. Como la varianza total ($h^2$) es 10, esto supone el 54,38% de la varianza total. Entre los 3 primeros factores explican el 78,866% de $h^2$

El primer factor explica una cantidad de la varianza total que vale 5,438. El segundo explica 1,313, y el tercero, 1,135. Estas cantidades se denominan *autovalores* ($\lambda_j$) (*eigenvalues*, en inglés). El criterio por omisión que aplica STATA en el AFCP realizado desde la instrucción *factor* es extraer solo aquellos factores que sean capaces de explicar una cantidad de varianza total superior o igual a 1. Por lo tanto, el autovalor de un factor ($\lambda_j$) se define como la cantidad de varianza total de la muestra que puede ser explicada por un determinado factor. En el ejemplo, la cantidad total de varianza es 10, porque hay 10 variables y el procedimiento las estandariza para que, en cada variable, se obtengan valores $z$ con varianza igual a 1. La suma de los autovalores de todos los factores posibles será igual a $h^2$, es decir, 10 en el ejemplo ($h^2 = 10$).

En el AFCP, el máximo valor de este parámetro corresponderá a la suma de las varianzas de cada variable, aunque este valor se corresponderá también con el número total de variables, ya que se trata de variables estandarizadas con media = 0 y varianza = 1. Si se extrajesen todos

los factores (y no solo aquellos con autovalores ≥ 1), se explicaría el 100% de la variabilidad de la muestra. Para calcular el porcentaje de varianza de una sola variable explicada por el factor ($\lambda_j$) solo será necesario dividir el autovalor por el número de variables en la muestra ($\lambda_j/i$ variables). Intuitivamente se entenderá que es lógico seleccionar solo los factores con un autovalor superior o igual a 1, ya que son los únicos capaces de explicar más varianza que la que explica cada una de las variables originales, y la finalidad es reducir el número de variables. En este ejemplo, los factores con autovalor superior a la unidad son tres. Por ello, STATA solo presenta en la segunda tabla del análisis el resultado obtenido para la extracción de tres factores.

La segunda tabla presentada por STATA corresponde a las correlaciones (coeficientes de Pearson) entre cada uno de los nuevos factores obtenidos y las variables originales ($a_{ij}$). A estos coeficientes de correlación ($a_{ij}$) se les denomina *factor loading* en inglés.

El cuadrado de este coeficiente de correlación o *factor loading* equivale a la *comunalidad* de la variable analizada: $c_i = a_{ij}^2$ (es un $R^2$ múltiple o coeficiente de determinación múltiple). La comunalidad se define como el porcentaje de variabilidad de cada variable que puede ser explicada por los factores extraídos. La colectividad o comunalidad ($c_i$) de una variable estandarizada puede oscilar entre 0 y 1. Un valor de 0 para la colectividad ($c_i$) de la variable $z_i$ se traduciría en que ninguna cantidad de la varianza de dicha variable sería explicada por los factores extraídos. Un valor 1 en $c_i$ indicaría que toda la varianza de $z_i$ sería explicada por el conjunto de factores extraídos en el análisis. En el AFCP, la comunalidad inicial para cada variable es 1, pues, si se obtuviesen todos los factores posibles, se podría explicar el 100% de la varianza de las 10 variables originales. STATA presenta en la última columna de esta segunda tabla los valores de *uniqueness* (singularidad), que corresponde a la diferencia 1 – comunalidad. Es decir, el valor de singularidad de cada variable representa el porcentaje de variabilidad de la misma que *no* es explicada por los factores.

El primer factor subyacente extraído (o componente principal) se relaciona positivamente con todas las variables originales de riesgo. Con la que más se asocia es con el tabaco *(smoke)*, con la que menos se asocia es con el sedentarismo en el trabajo *(sedent_w)*

Factor loadings (pattern matrix) and unique variances

| Variable | Factor1 | Factor2 | Factor3 | Uniqueness |
|---|---|---|---|---|
| alcohol | 0.8791 | -0.1723 | 0.2710 | 0.1241 |
| smoke | 0.9037 | -0.0279 | 0.0270 | 0.1818 |
| cannabis | 0.8667 | -0.1393 | -0.0443 | 0.2275 |
| junkfood | 0.8101 | 0.3442 | -0.1100 | 0.2132 |
| soda_dr | 0.8405 | 0.2582 | -0.1313 | 0.2097 |
| sedent_w | 0.1533 | 0.5328 | 0.7472 | 0.1343 |
| sedent_l | 0.3986 | 0.7438 | -0.2711 | 0.2143 |
| riskdriv | 0.8176 | -0.3114 | 0.2007 | 0.1943 |
| celldriv | 0.6998 | -0.3646 | 0.2198 | 0.3290 |
| partner | 0.6278 | -0.1064 | -0.5561 | 0.2852 |

La última columna determina cuánta varianza de cada variable NO se puede explicar a partir de los factores obtenidos o extraídos. La interpretación sería que entre los 3 factores extraídos se explica el 87,6% de la variabilidad en la exposición al alcohol. La variable original *alcohol* es la que mejor queda explicada por los 3 factores. En cambio, la variable original que peor se explica es la de usar el teléfono móvil al conducir *(celldriv)*

El segundo factor (o componente) extraído se relaciona sobre todo con el sedentarismo en el tiempo libre *(sedent_l)* y también con el sedentarismo en el trabajo *(sedent_w)*. En menor medida se asocia con hábitos de comida rápida. Contrariamente al primero, este segundo factor se asocia negativamente con comportamientos de riesgo en la conducción

El tercer factor se relaciona positivamente con el sedentarismo en el trabajo *(sedent_w)* y negativamente con comportamientos sexuales de riesgo *(partner)*

STATA permite obtener también las correlaciones *(factor loadings)* entre los factores y las variables a través de la instrucción:

```
estat structure
```

Structure matrix: correlations between variables and unrotated common factors

| Variable | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| alcohol | 0.8791 | -0.1723 | 0.2710 |
| smoke | 0.9037 | -0.0279 | 0.0270 |
| cannabis | 0.8667 | -0.1393 | -0.0443 |
| junkfood | 0.8101 | 0.3442 | -0.1100 |
| soda_dr | 0.8405 | 0.2582 | -0.1313 |
| sedent_w | 0.1533 | 0.5328 | 0.7472 |
| sedent_l | 0.3986 | 0.7438 | -0.2711 |
| riskdriv | 0.8176 | -0.3114 | 0.2007 |
| celldriv | 0.6998 | -0.3646 | 0.2198 |
| partner | 0.6278 | -0.1064 | -0.5561 |

En algunas ocasiones, cuando las correlaciones *(factor loadings)* son débiles, al investigador no le interesa presentar todas ellas, sino que especifica al ordenador que omita mostrar aquellos coeficientes con un valor bajo (arbitrariamente, suelen ser *factor loadings* menores que 0,30). Esta instrucción puede ser solicitada a STATA mediante la subinstrucción:

```
blanks ()
```

En el ejemplo:

```
fac alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner, pcf bl(.30)
```

El principal reto del AFCP consiste en interpretar ahora estos hallazgos, en ponerles «nombre». Se han identificado tres patrones de comportamientos asociados, que entre los tres explican cerca del 80% (78,87%) de la variabilidad total observada en las 10 variables originales. Con esto se ha conseguido una reducción de variables. ¿Cómo denominar a estos tres factores? En general, en la interpretación se tiende a prescindir de aquellas variables cuyo coeficiente de correlación con el factor $(a_{ij})$ es inferior a 0,3.

1. El primero *carga* con signo positivo (asociación directa) todas las variables de comportamientos de riesgo y podría corresponder al sujeto que es poco consciente y responsable de su salud. Su retrato robot podría ser así. Su principal característica es ser fumador intenso, seguido muy de cerca de tener un alto consumo de alcohol y frecuente exposición a cannabis; además, tiene hábitos alimentarios insanos *(soda_dr y junkfood)*, conduce arriesgadamente y usa el teléfono móvil mientras conduce. En menor medida tiende a ser sexualmente promiscuo y, todavía con menor asociación, a tener un estilo de vida sedentario, más en el tiempo libre que en el trabajo. Se le podría denominar *estilo de vida globalmente insano*.
2. El segundo carga positivamente de manera intensa solo las dos variables del sedentarismo. Después del sedentarismo, la siguiente correlación más alta en magnitud es de signo negativo y corresponde a usar el móvil al conducir. La siguiente es de signo positivo y se asocia a la comida rápida. Con una magnitud ligeramente inferior, pero del mismo signo que sus variables de contenido similar, se encuentran la conducción arriesgada (asociación inversa) y los refrescos azucarados (signo positivo). El retrato robot de este segundo personaje es el de un individuo cuya principal característica es el sedentarismo, que, además, se acompaña de hábitos de consumo alimentario que le inducen a la obesidad, pero es cuidadoso al conducir. Se le podría denominar *sedentario y amigo de McDonald, pero buen conductor*.

3. El tercer patrón es el de una persona sedentaria en su trabajo (pero no en el tiempo libre) y que, además, evita exposiciones sexuales de riesgo.

### 16.1.4. Construcción de los factores (coeficientes o pesos de cada factor o componente)

Hasta ahora no se han visto los factores, sino tan solo sus correlaciones con las variables originales y el porcentaje de la varianza total que son capaces de explicar. El AFCP opera desde la perspectiva de que los factores que se extraen en el análisis son *ortogonales* entre sí, es decir, no se encuentran correlacionados y son combinaciones lineales (esto es, sumas ponderadas) de las variables estandarizadas incluidas en el análisis. En cada factor se asigna un peso a cada variable; estos pesos o ponderaciones se denominan coeficientes (*scoring coefficients*, según la nomenclatura de STATA).

Existen varios métodos para la creación de los factores y la obtención de los coeficientes (**2-4**). La regresión, la más utilizada, es aplicada por STATA por defecto. Otro método disponible en STATA es el de Bartlett.

Esta aproximación (AFCP) asume que los factores extraídos podrían calcularse perfectamente a través de las variables originales incluidas en el análisis.

$$\text{Factor}_j = \sum_{i=1} w_{ij} \times z_i$$

donde:

$i$ = cada variable.

$z_i$ = valor estandarizado de cada variable = $(x_i - \mu)/\sigma$.

$j$ = cada factor.

$w_{ij}$ = peso de la variable $i$ sobre el factor $j$. El peso de la variable $i$ sobre el factor $j$ representa la correlación de la variable $i$ con el factor $j$. STATA los denomina *scoring coefficients*.

Es decir:

$$\text{Factor 1} = w_1z_1 + w_2z_2 + w_3z_3 + w_4z_4 + w_5z_5 + w_6z_6 + w_7z_7 + w_8z_8 + w_9z_9 + w_{10}z_{10}$$

Esto se repetiría con unos $w_i$ distintos para el Factor 2, el Factor 3, etc. En el ejemplo:

Scoring coefficients (method = regression)

| Variable | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| alcohol | 0.16166 | -0.13119 | 0.23871 |
| smoke | 0.16619 | -0.02121 | 0.02377 |
| cannabis | 0.15938 | -0.10604 | -0.03906 |
| junkfood | 0.14896 | 0.26208 | -0.09690 |
| soda_dr | 0.15455 | 0.19655 | -0.11563 |
| sedent_w | 0.02818 | 0.40566 | 0.65828 |
| sedent_l | 0.07330 | 0.56634 | -0.23881 |
| riskdriv | 0.15034 | -0.23711 | 0.17685 |
| celldriv | 0.12869 | -0.27760 | 0.19362 |
| partner | 0.11545 | -0.08103 | -0.48989 |

Estos son los tres factores reales que se han extraído. El primero se calcularía del modo siguiente:

$$\text{Factor 1} = 0{,}162 * \text{zalcohol} + 0{,}166 * \text{zsmoke} + \ldots + 0{,}115 * \text{zpartner}$$

El segundo sería:

$$\text{Factor 2} = -0{,}131 * \text{zalcohol} - 0{,}021 * \text{zsmoke} + \ldots - 0{,}081 * \text{zpartner}$$

El tercero se determina como:

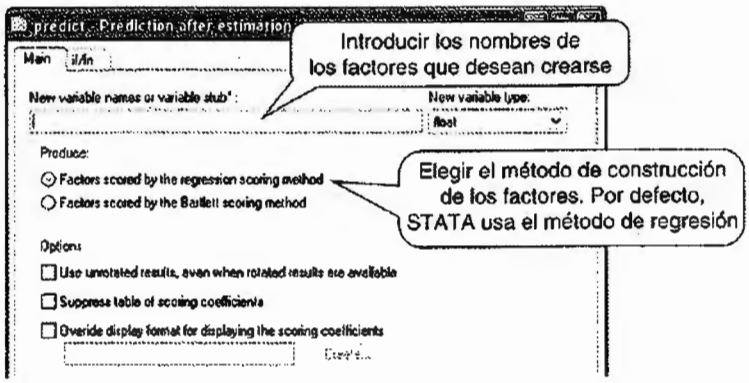
$$\text{Factor 3} = 0{,}239 * \text{zalcohol} + 0{,}024 * \text{zsmoke} + \ldots - 0{,}490 * \text{zpartner}$$

Además, STATA, al recibir la orden anterior, habrá calculado y guardado ya estos tres factores como tres nuevas columnas, que son las que aparecen ahora al final de nuestra base de datos. Esto permite que puedan usarse para análisis posteriores.

Ahora se verá cómo pedir a STATA los pesos o ponderaciones $(w_j)$ que tienen que aplicarse a cada variable estandarizada $(z_j)$ para obtener el valor real del factor y cómo se crean estas tres variables (factores finales). Desde el menú:

**Statistics → Postestimation → Prediction, residuals, etc.**

se abre el siguiente cuadro de diálogo:




Utilizando la sintaxis del programa, la instrucción es[3]:

```
predict f1 f2 ... fj, regression
```

siendo $f_1$-$f_j$, los nombres que desea darse a cada factor.

En el ejemplo, se decide llamar a cada factor «factor»:

```
predict factor1 factor2 factor3
```

Si desease obtener los coeficientes a través del método de Bartlett, debería utilizarse la subinstrucción:

```
bartlett
predict factor1 factor2 factor3, b
```

Puede comprobarse que la correlación entre cada par de estos tres factores es 0. Por eso se dice que son *ortogonales*. Desde la instrucción:

```
estat common
```

puede obtenerse la matriz de correlación entre los factores extraídos.

3 No es necesario escribir la subinstrucción **regression**, ya que STATA aplica el método de regresión por defecto.

```
. estat common
Correlation matrix of the common factors
```

| Factors | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| Factor1 | 1 | | |
| Factor2 | 0 | 1 | |
| Factor3 | 0 | 0 | 1 |

## 16.2. NÚMERO DE FACTORES PARA EXTRAER

El objetivo del análisis factorial es conseguir un modelo con parsimonia. Es decir, maximizar la cantidad de varianza de las variables que puede ser explicada por el mínimo número de factores o componentes subyacentes. Para alcanzar dicho objetivo podemos seguir diferentes líneas de actuación, que se exponen a continuación.

### 16.2.1. Autovalores superiores a 1 o criterio de Kaiser

Una de las aproximaciones para determinar el número inicial de factores consiste en seleccionar solo los factores con autovalores superiores a la unidad. Esta aproximación es conocida como norma de Kaiser-Guttman (3,5,6).

En general, se establece que este criterio tiene mayor validez cuando:

- Existen menos de 40 variables.
- El tamaño muestral es elevado.
- El número de factores esperados oscila entre (n.° de variables/5) y (n.° de variables/3) (7).

### 16.2.2. Porcentaje de varianza extraída

Un segundo criterio de selección del número de factores es el de determinar el porcentaje acumulado de varianza extraída por diferentes factores. El investigador termina con la extracción de factores cuando alcanza un porcentaje acumulado de varianza explicada del 75-80%. Sin embargo, no existe unanimidad a la hora de establecer el porcentaje de varianza acumulada suficiente para determinar el número de factores.

```
Factor analysis/correlation                    Number of obs    =         40
    Method: principal-component factors        Retained factors =          3
    Rotation: (unrotated)                      Number of params =         27
```

| Factor | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 5.43802 | 4.12459 | 0.5438 | 0.5438 |
| Factor2 | 1.31343 | 0.17829 | 0.1313 | 0.6751 |
| Factor3 | 1.13514 | 0.42531 | 0.1135 | 0.7887 |
| Factor4 | 0.70982 | 0.20214 | 0.0710 | 0.8596 |
| Factor5 | 0.50768 | 0.11342 | 0.0508 | 0.9104 |
| Factor6 | 0.39427 | 0.16585 | 0.0394 | 0.9498 |
| Factor7 | 0.22842 | 0.11002 | 0.0228 | 0.9727 |
| Factor8 | 0.11840 | 0.03176 | 0.0118 | 0.9845 |
| Factor9 | 0.08663 | 0.01844 | 0.0087 | 0.9932 |
| Factor10 | 0.06820 | . | 0.0068 | 1.0000 |

***Interpretación:*** *en este ejemplo se han extraído todos los posibles factores (10) que deben coincidir con el número de variables incluidas en el análisis. Los tres primeros factores explicarían el 78,87% de la varianza total. La elección final sería la extracción de tres factores (según el criterio de Kaiser, igualmente se extraerían los tres primeros factores que poseen autovalores superiores a la unidad).*

### 16.2.3. Test de la pendiente de Cattell o gráfico de sedimentación *(scree plot)*

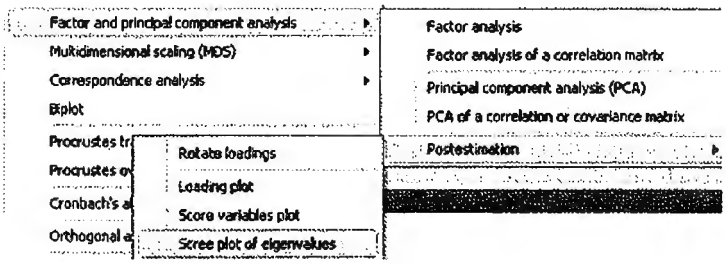
El tercer criterio consiste en representar gráficamente los factores extraídos (colocados en el eje de abscisas) frente a sus autovalores (dispuestos en el de ordenadas) para establecer un punto de

inflexión en la gráfica (8). Para determinar el punto de inflexión, basta con dibujar una línea recta a lo largo de los valores de autovalor más bajos. El punto en el que los factores se curvan (punto de inflexión) sobre la línea recta identifica el número de factores que deben ser extraídos (7,9). Sin embargo, algunas representaciones son difíciles de interpretar. Pueden tener diferentes puntos de inflexión o podría no existir un punto de inflexión claro en la representación.

El test de la pendiente de *Cattell* es más preciso cuando el tamaño de muestra es elevado, los valores de colectividad son altos y la razón variables: factores es al menos de 3:1.

El gráfico de sedimentación es realizado por STATA desde la opción:

**Statistics → Multivariate analysis → Factor and Principal components analysis → Postestimation-Scree plot of eigenvalues**




Puede obtenerse también a través de la instrucción:

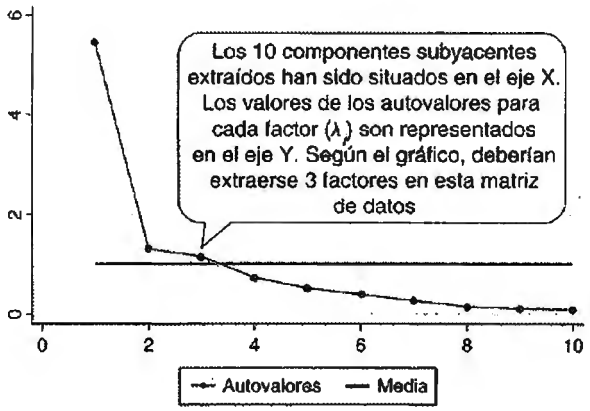
```
screeplot
```

Además, STATA, a través de la subinstrucción:

```
mean
```

permite trazar una línea horizontal en la gráfica que represente el autovalor medio que corresponde a un valor de 1. Es decir, la gráfica expresará el número de factores que no cumplen el criterio de Kaiser.

```
screeplot, mean
```

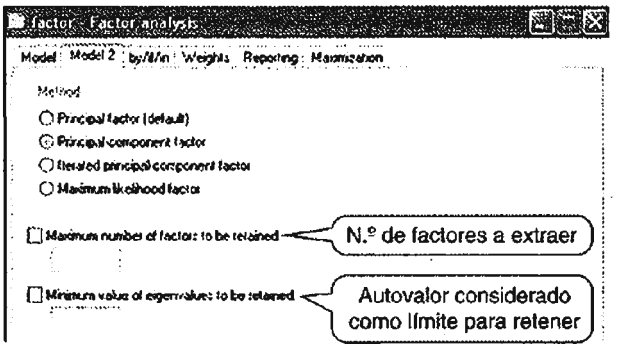



Varios autores indican que, si el número de factores para extraer es difícil de establecer a través de los criterios anteriormente expuestos, el investigador debería llevar a cabo diferentes análisis

factoriales con extracción de diferente número de factores en cada ocasión. Observando los diferentes resultados, el investigador podrá elegir el modelo con mayor parsimonia que dé mayor sentido intuitivo al problema expresado.

A continuación se mostrará cómo se extrae un número determinado de factores con STATA.

Al solicitar a STATA un AFCP, recuérdese que se abre el siguiente cuadro de diálogo con diferentes lengüetas, una de ellas con el nombre *Model 2*. La lengüeta denominada *Model 2* permite especificar, además del método de extracción, el número de factores que se desea extraer o el valor mínimo del autovalor que quiere mantenerse.




Si el análisis se realiza a través de instrucciones, puede seleccionarse el número de factores a través de la subinstrucción:

```
factors (n.º de factores que se desean)
```

o bien todos aquellos factores con un autovalor superior a un determinado valor, mediante la subinstrucción:

```
mineigen (valor considerado)
```

Por omisión, en el AFCP, STATA mantiene los factores con *eigenvalue* superior a 1.

En el ejemplo, si se quisieran extraer solo dos factores:

```
fac alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner, pcf fa (2)
```

## 16.3. CÁLCULOS NUMÉRICOS

### 16.3.1. Cálculo del valor de una variable a través de los factores extraídos

El AFCP opera desde la perspectiva de que los factores que se extraen en el análisis son *ortogonales* entre ellos, es decir, no se encuentran correlacionados y son combinaciones lineales de las variables estandarizadas incluidas en el análisis.

Esta aproximación asume que las variables incluidas en el análisis podrían calcularse perfectamente a través de los factores extraídos de la siguiente forma:

$$z_i = \sum_{j=1} a_{ij} \times \text{factor}_j$$

$$z_{\text{variable1}} = a_{11} \times \text{factor1} + a_{12} \times \text{factor2} + \ldots + a_{1j} \times \text{factor}_j$$

donde:

$i$ = cada variable.

$z_i$ = valor estandarizado de cada variable.

$j$ = cada factor.

$a_{ij}$ = coeficiente de correlación de la variable $i$ con el factor $j$ *(factor loading)*. Es representado en la matriz factorial.

$a_{11} \ldots a_{1j}$ = coeficientes de correlación de la variable 1 con los j factores de la solución.

### 16.3.2. Correlación entre dos variables *(r)*

La correlación entre dos variables, además de observarse en la matriz de correlación, puede obtenerse a través de los coeficientes de correlación de dichas variables con los factores $(a_{ij})$. La correlación para dos variables ($i$ e $i'$) es la suma de los productos de los pesos para estas dos variables $(a_i) \times (a_{i'})$ a lo largo de los j factores. Esta ecuación queda expresada a través de la siguiente fórmula:

$$r_{ii'} = \sum_{j=1} \left(a_i \times a_{i'}\right)_j$$

$$r_{\text{variable1,2}} = \left(a_{11} \times a_{21}\right) + \left(a_{12} \times a_{22}\right) + \left(a_{13} \times a_{23}\right) + \ldots\ldots + \left(a_{1j} \times a_{2j}\right)$$

donde:

$r_{ii'}$ = coeficiente de correlación de *Pearson* entre las variables $i$ e $i'$.

$a_{11}, a_{21} \ldots a_{1j}, a_{2j}$ = coeficientes de correlación de las variables *1* y *2* con los $j$ factores.

### 16.3.3. Colectividad de una variable *(c)*

Una vez llevado a cabo el análisis factorial, si calculamos la suma de los coeficientes de una variable $i$ con los $j$ factores al cuadrado, obtendremos el valor de su colectividad $(c_i)$.

$$c_i = \sum_{j=1} \left(a_i\right)_j^2$$

$$c_1 = \left(a_{11}\right)^2 + \left(a_{12}\right)^2 + \left(a_{13}\right)^2 + \ldots\ldots + \left(a_{1j}\right)^2 = 1$$

donde:

$i$ = cada variable.

$j$ = cada factor.

$a_{11} \ldots a_{1j}$ = coeficiente de correlación de la variable *1* sobre los $j$ factores.

### 16.3.4. Autovalor de un factor *(λ)*

El autovalor asociado a un factor $j$ puede calcularse a través de la suma de los cuadrados de los pesos de dicho factor $j$ en cada variable $i$:

$$\lambda_j = \sum_{i=1} \left(a_j\right)_i^2$$

$$\lambda_1 = \left(a_{11}\right)^2 + \left(a_{21}\right)^2 + \left(a_{31}\right)^2 + \ldots\ldots + \left(a_{i1}\right)^2$$

donde:

$i$ = cada variable.

$j$ = cada factor.

$a_{11} \ldots a_{i1}$ = pesos de las $i$ variables sobre el factor *1*.

*Tabla 16.2 Sinónimos y equivalencias en el análisis factorial*

| | | | | |
|---|---|---|---|---|
| Variable | Característica individual | Ítem | Columna original | $i$ |
| Factor | Componente (principal) | Patrón | Puntuación | $j$ |
| Comunalidad | % de cada variable explicado por los factores | Colectividad | $R^2$ múltiple | $c$ |
| Coeficiente | Peso | Ponderación | *Factor score* | $w$ |
| Varianza total | Suma de las varianzas de todas las $z$ originales | | n.º de variables | $h^2$ |
| Autovalor | % de la muestra explicado por un factor | | *Eigenvalue* | $\lambda$ |
| Coeficiente de correlación variable-factor | | | *Factor load* | $a$ |

## 16.4. SINONIMIAS Y EQUIVALENCIAS

Para mejorar la interpretación de los términos que aparecen en un AFCP, se presenta la tabla 16.2 de sinónimos y equivalencias.

## 16.5. CONDICIONES DE APLICACIÓN DEL ANÁLISIS FACTORIAL DE COMPONENTES PRINCIPALES (AFCP)

Antes de iniciar la extracción de factores, se recomienda examinar la matriz de correlación de las variables implicadas en el análisis.

### 16.5.1. Coeficientes de correlación entre variables (*r*) inferiores a 0,30

Si una primera inspección visual de la matriz de correlación de las variables revela que no existe un suficiente número de correlaciones significativas superiores a 0,30, la aplicación de un análisis factorial es probablemente inapropiada, ya que no será posible obtener un grupo reducido de factores que representen las variables iniciales, y podría ocurrir que obtuviéramos tantos factores como variables iniciales poseíamos.

### 16.5.2. Coeficientes de correlación entre variables (*r*) superiores a 0,80

Tampoco sería apropiada su aplicación si las variables están muy correlacionadas ($r > 0{,}80$), ya que podrían dar problemas de multicolinealidad y conducir a un sistema inestable.

En STATA pueden solicitarse los coeficientes de correlación entre variables a través de la instrucción del menú:

**Statistics → Summaries, tables and tests → Summary and descriptive statistics → correlations and covariances**

o de la instrucción:

```
correlate
cor alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner
```

| | alcohol | smoke | cannabis | junkfood | soda_dr | sedent_w | sedent_l | riskdriv | celldriv | partner |
|---|---|---|---|---|---|---|---|---|---|---|
| alcohol | 1.0000 | | | | | | | | | |
| smoke | 0.8457 | 1.0000 | | | | | | | | |
| cannabis | 0.7815 | 0.8784 | 1.0000 | | | | | | | |
| junkfood | 0.5959 | 0.6437 | 0.5139 | 1.0000 | | | | | | |
| soda_dr | 0.6134 | 0.6999 | 0.5826 | 0.8884 | 1.0000 | | | | | |
| sedent_w | 0.2018 | 0.1516 | 0.0446 | 0.1961 | 0.1270 | 1.0000 | | | | |
| sedent_l | 0.1800 | 0.3068 | 0.3274 | 0.4811 | 0.4476 | 0.1876 | 1.0000 | | | |
| riskdriv | 0.8297 | 0.7015 | 0.7115 | 0.5012 | 0.5364 | 0.0582 | 0.1527 | 1.0000 | | |
| celldriv | 0.6212 | 0.4823 | 0.5356 | 0.4988 | 0.5444 | 0.0807 | -0.0276 | 0.6879 | 1.0000 | |
| partner | 0.3778 | 0.5410 | 0.3687 | 0.5047 | 0.5038 | -0.1868 | 0.2379 | 0.3827 | 0.3730 | 1.0000 |

El coeficiente de correlación de Pearson (*r*) para el sedentarismo en el tiempo libre (*sedent_l*) y el uso de cannabis es de 0,327

### 16.5.3. Test de esferidad de Bartlett

La hipótesis nula de este test es que no existe correlación ninguna entre las variables. El estadístico utilizado en el test es una *ji cuadrado* con $i\,(i-1)/2$ grados de libertad, donde $i$ es el número de variables incluidas en el análisis (**10**). Se rechazará la hipótesis nula cuando la *ji cuadrado* observada sea superior a la *ji cuadrado* contenida en las tablas y correspondiente a un valor de significación estadística del 5%.

Sin embargo, este test es muy sensible al tamaño muestral. Grandes muestras se traducen en elevados valores del test. De esta forma, la hipótesis nula es, a menudo, rechazada. Por ello, se ha argumentado que este test debe ser considerado únicamente cuando los resultados aportados no sean estadísticamente significativos (no puede rechazarse la hipótesis nula de matriz de identidad), y la matriz de datos no debería estudiarse a través de un análisis factorial.

STATA no aplica este test.

### 16.5.4. Coeficientes de correlación parcial entre variables elevados

Otro método de evaluación de la adecuación del análisis factorial es la observación de las correlaciones parciales entre las variables. El *coeficiente de correlación parcial* de un par de variables se define como la correlación entre dicho par una vez eliminado el efecto de las variables restantes. Si realmente existen factores subyacentes a las variables iniciales, estos coeficientes deben ser bajos. Cuando los coeficientes de correlación parciales son elevados, el análisis factorial no es apropiado.

STATA calcula la *anti-imagen* de la matriz de correlación. Los valores de esta matriz expresan los valores negativos de los coeficientes de correlación parcial de cada par de variables. Estos valores, en valor absoluto, deben ser lo más pequeños posible.

El coeficiente de correlación parcial para las variables alcohol-tabaco *(smoke)* vale 0,393 (en la tabla aparece el coeficiente negativo, –0,393)

```
Anti-image correlation coeffic ents —— partialing out all other variables
```

| variable | alcohol | smoke | cannabis | junkfood | soda_dr | sedent_w | sedent_l |
|---|---|---|---|---|---|---|---|
| alcohol | 1.0000 | | | | | | |
| smoke | -0.3932 | 1.0000 | | | | | |
| cannabis | -0.1676 | -0.6098 | 1.0000 | | | | |
| junkfood | -0.2163 | -0.0440 | 0.2571 | 1.0000 | | | |
| soda_dr | 0.1063 | -0.3094 | 0.0759 | -0.6823 | 1.0000 | | |
| sedent_w | -0.1539 | -0.1736 | 0.1587 | -0.1288 | 0.1557 | 1.0000 | |
| sedent_l | 0.1865 | 0.2166 | -0.3981 | -0.2821 | -0.1636 | -0.2209 | 1.0000 |
| riskdriv | -0.4983 | -0.0213 | -0.0268 | 0.0515 | 0.0442 | 0.1855 | -0.1445 |
| celldriv | -0.0855 | 0.3891 | -0.2992 | -0.0935 | -0.2796 | -0.1686 | 0.4314 |
| partner | 0.2525 | -0.1539 | -0.2508 | -0.2702 | 0.0716 | 0.2923 | 0.0438 |

| variable | riskdriv | celldriv | partner |
|---|---|---|---|
| riskdriv | 1.0000 | | |
| celldriv | -0.3855 | 1.0000 | |
| partner | 0.0073 | -0.1066 | 1.0000 |

### 16.5.5. Medidas individuales de adecuación y medida de adecuación de la muestra de Kaiser-Meyer-Olkin

El test de Kaiser-Meyer-Olkin (KMO) es una medida de adecuación de la muestra que compara los coeficientes de correlación de *Pearson* entre cada par de variables con sus respectivos coeficientes de correlación parciales. Este parámetro indica el grado de correlación de una variable concreta con el resto de variables de la matriz o base de datos. Su valor puede oscilar entre 0 y 1. Los criterios empleados para determinar la adecuación de la muestra son:

Si KMO es >0,90: excelente.

Si KMO es ≥0,80 y <0,90: buena.

Si KMO es ≥0,70 y <0,80: normal.

Si KMO es ≥0,60 y <0,70: mediocre.

Si KMO es <0,60: inaceptable.

Idealmente, para la ejecución de un análisis factorial, este parámetro debe ser superior a 0,70 en cada una de las variables del estudio (11).

STATA calcula este test para cada variable y para el conjunto de las mismas.

```
Kaiser-Meyer-Olkin measure of sampling adequacy
```

| Variable | kmo |
|---|---|
| alcohol | 0.8447 |
| smoke | 0.8009 |
| cannabis | 0.7997 |
| junkfood | 0.7891 |
| soda_dr | 0.8101 |
| sedent_w | 0.3873 |
| sedent_l | 0.5640 |
| riskdriv | 0.8625 |
| celldriv | 0.7447 |
| partner | 0.8340 |
| overall | 0.7919 |

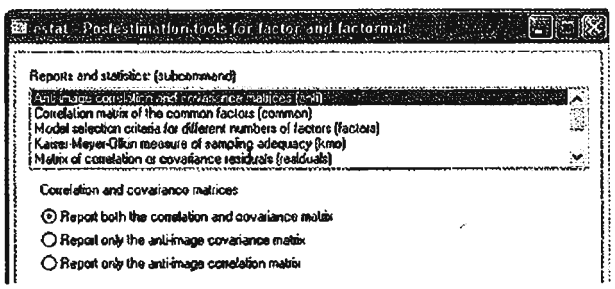
La medida de adecuación muestral para el *alcohol* vale 0,845 y para el tabaco *(smoke)*, 0,801. La medida de adecuación total es 0,792, por tanto, normal

Si los coeficientes de adecuación de la muestra son mediocres (0,60) para algunas variables y los coeficientes de correlación parcial entre pares resultan demasiado elevados, el investigador debe identificar la variable con menor coeficiente de adecuación de la muestra y eliminarla del análisis final.

Si esta medida no conduce a una mejora en los valores del KMO o de los coeficientes de adecuación de la muestra individuales, el investigador debería aumentar el tamaño muestral o replantearse la adecuación de llevar a cabo un análisis factorial.

Para determinar si el AFCP cumple con los criterios de aplicación, calculando la anti-imagen de la matriz de correlación o llevando a cabo pruebas como la medida de adecuación de la muestra de Kaiser-Meyer-Olkin, el investigador deberá acudir al menú:

**Statistics ≥ Postestimation ≥ Reports and statistics**




Con instrucciones:

```
estat anti
```

Esta instrucción aporta la anti-imagen de la matriz de correlación y de la matriz de covarianzas. Si no quisiéramos visualizar alguna de ellas, bastaría con dar la instrucción al programa a través de las subinstrucciones:

```
nocorr
```

y

```
nocov
```

Para obtener la medida de adecuación de la muestra de Kaiser-Meyer-Olkin, la instrucción será:

```
estat kmo
```

## 16.6. CONSIDERACIONES SOBRE EL TAMAÑO MUESTRAL

El número de sujetos necesarios para llevar a cabo un análisis factorial depende del número de variables que se vayan a analizar. Sin embargo, entre los diferentes expertos en el manejo de esta técnica no existe un criterio único de estimación de tamaño muestral.

Se ha sugerido la inclusión de 10 sujetos por variable analizada (**12**). De manera general, la adecuación del tamaño muestral se establece de la siguiente manera (**3**):

- 50 sujetos: muy pobre.
- 100 sujetos: pobre.
- 200 sujetos: normal.
- 300 sujetos: buena.
- 500 sujetos: muy buena.
- o más sujetos: excelente.

Algunos autores indican la necesidad de al menos 300 sujetos para realizar correctamente un análisis factorial, aunque este número puede ser reducido a 150 si existen varias variables con coeficientes de correlación con factores $(a_{ij})$ superiores a 0,80 (**13**).

## 16.7. ROTACIÓN DE LOS FACTORES

A veces, los factores obtenidos en los diferentes tipos de análisis factorial no se traducen en una agrupación fácilmente interpretable de variables con significado. La rotación mejora el significado y la interpretación de los factores obtenidos originariamente (**14**).

Existen dos tipos de rotaciones. La rotación ortogonal asume que los factores generados son independientes entre sí (no están correlacionados). En la rotación oblicua se supone que los factores no son independientes y que existe correlación entre dos o más de ellos.

### 16.7.1. Rotación ortogonal

Debido a que los factores no están correlacionados, los coeficientes de correlación entre las diferentes variables y sus correspondientes factores $(a_{ij})$ coinciden con coeficientes de regresión estandarizados $(\beta_{ij})$ en un modelo de regresión en el que la variable dependiente es la inicial y las variables independientes son los factores extraídos tras la rotación.

$$z_i = \sum_{j=1} \beta_{ij} \times \text{factor}_j$$

$$z_{\text{variable1}} = \beta_{11} \times \text{factor}1 + \beta_{12} \times \text{factor}2 + \ldots\ldots + \beta_{1j} \times \text{factor}_j$$

donde:

$i$ = cada variable.

$z_i$ = valor estandarizado de cada variable.

$j$ = cada factor.

$\beta_{ij}$ = coeficiente de regresión estandarizado del factor *j* para la variable *i* (coincide con el coeficiente de correlación entre la variable *i* y el factor *j*).

$\beta_{i1} \ldots \beta_{ij}$ = coeficientes de regresión del factor *j* para la variable *I* de la solución.

Independientemente de la aplicación de una rotación ortogonal, las correlaciones entre las variables *(r)* no se modifican. Tampoco se produce cambio alguno sobre las colectividades de las mismas.

Existen varias aproximaciones a las rotaciones ortogonales. Los ejemplos más conocidos son las rotaciones *Varimax, Quartimax, Equamax* u *Oblimin.*

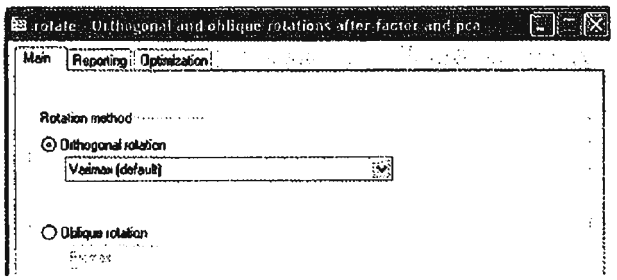
### 16.7.2. Rotación oblicua

En este tipo de rotación, los factores no son independientes, sino que se encuentran correlacionados entre sí en mayor o menor grado. Cuando se lleva a cabo una rotación oblicua, la contribución de cada factor a la variabilidad de la variable $(\beta_{ij})$ no equivale a su coeficiente de correlación entre la variable y el factor $(a_{ij})$.

Existen diferentes aproximaciones para alcanzar la estructura simple en una rotación oblicua. Algunos de los más comunes son: *Oblimin* oblicua y *Promax.*

La instrucción del menú de STATA que permite rotar los factores es:

**Statistics → Multivariate analysis → Factor and principal component analysis → Postestimation → Rotate loadings**




De igual forma, STATA realiza rotaciones a través de la instrucción:

```
rotate
```

El tipo de rotaciones debe ser especificado a través de subinstrucciones[4]:

```
rot, varimax
rot, equamax
rot, quartimax
rot, oblimin
rot, oblimin oblique
rot, promax
```

4 En el caso de la rotación ortogonal *Varimax*, STATA la aplica por defecto, así que en ese caso no es necesario especificar dicha subinstrucción.

```
Factor analysis/correlation                    Number of obs    =        40
    Method: principal-component factors        Retained factors =         3
    Rotation: orthogonal varimax (Kaiser off)  Number of params =        27
```

| Factor | Variance | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 4.40807 | 2.11325 | 0.4408 | 0.4408 |
| Factor2 | 2.29483 | 1.11115 | 0.2295 | 0.6703 |
| Factor3 | 1.18368 | . | 0.1184 | 0.7887 |

LR test: independent vs. saturated: chi2(45) = 308.18 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

| variable | Factor1 | Factor2 | Factor3 | Uniqueness |
|---|---|---|---|---|
| alcohol | 0.9082 | 0.1902 | 0.1225 | 0.1 |
| smoke | 0.8034 | 0.4154 | -0.0164 | 0.1 |
| cannabis | 0.7983 | 0.3429 | -0.1330 | 0.2275 |
| junkfood | 0.5301 | 0.7091 | 0.0550 | 0.2132 |
| soda_dr | 0.5866 | 0.6679 | -0.0075 | 0.2097 |
| sedent_w | 0.1136 | 0.1624 | 0.9091 | 0.1343 |
| sedent_l | -0.0370 | 0.8762 | 0.1285 | 0.2143 |
| riskdriv | 0.8936 | 0.0841 | -0.0064 | 0.1943 |
| celldriv | 0.8187 | -0.0226 | -0.0133 | 0.3290 |
| partner | 0.4387 | 0.4667 | -0.5518 | 0.2852 |

Coeficientes de correlación entre cada variable y los factores rotados *(factor loadings)*

Con la instrucción:

```
estat rotatecompare
```

pueden compararse los coeficientes *(factor loadings)* de los factores rotados y sin rotar.

Factor loadings

| variable | Rotated Factor1 | Rotated Factor2 | Rotated Factor3 | Unrotated Factor1 | Unrotated Factor2 | Unrotated Factor3 |
|---|---|---|---|---|---|---|
| alcohol | 0.9082 | 0.1902 | 0.1225 | 0.8791 | -0.1723 | 0.2710 |
| smoke | 0.8034 | 0.4154 | -0.0164 | 0.9037 | -0.0279 | 0.0270 |
| cannabis | 0.7983 | 0.3429 | -0.1330 | 0.8667 | -0.1393 | -0.0443 |
| junkfood | 0.5301 | 0.7091 | 0.0550 | 0.8101 | 0.3442 | -0.1100 |
| soda_dr | 0.5866 | 0.6679 | -0.0075 | 0.8405 | 0.2582 | -0.1313 |
| sedent_w | 0.1136 | 0.1624 | 0.9091 | 0.1533 | 0.5328 | 0.7472 |
| sedent_l | -0.0370 | 0.8762 | 0.1285 | 0.3986 | 0.7438 | -0.2711 |
| riskdriv | 0.8936 | 0.0841 | -0.0064 | 0.8176 | -0.3114 | 0.2007 |
| celldriv | 0.8187 | -0.0226 | -0.0133 | 0.6998 | -0.3646 | 0.2198 |
| partner | 0.4387 | 0.4667 | -0.5518 | 0.6278 | -0.1064 | -0.5561 |

## 16.8. REFINAMIENTO DEL ANÁLISIS: ELIMINACIÓN DE VARIABLES

### 16.8.1. Evaluación de la correlación variable-factor ($a_{ij}$)

En teoría, cada variable tendrá una alta correlación con un único factor tras la rotación de los factores. En la práctica, después de la rotación suele ocurrir que algunas variables tienen una baja correlación con todos los factores o una correlación alta con múltiples factores. ¿Qué se hace entonces?

#### 16.8.1.1. Variables con bajos coeficientes de correlación con todos los factores

Como ya se indicó, en los primeros pasos del análisis factorial debe observarse la matriz de correlación de las variables y eliminar del análisis aquellas variables que presenten correlaciones *(r)* inferiores a 0,30. De esta manera se evitará la creación de estructuras inestables.

Pese a esta medida, es posible que la solución final presente algunas variables con baja correlación con todos los factores. En este caso, se recomiendan dos medidas (14):

1. Eliminar las variables que no posean correlaciones lo suficientemente altas con alguno de los factores.
2. Evaluar la colectividad de la o las variables con bajas correlaciones con los factores y su contribución en la muestra objeto de valoración. Si la colectividad de esta variable es pequeña y su importancia relativa en la base de datos es mínima, la variable será eliminada.

#### 16.8.1.2. Variables con altos coeficientes de correlación con múltiples factores

Es frecuente encontrar variables con importantes correlaciones con diferentes factores ($a_{ij} > 0{,}30$), especialmente en soluciones en las que se aplicó una rotación oblicua.

Existen diferentes soluciones a este fenómeno:

1. Eliminar aquella variable con altos valores $a_{ij}$, ya que su inclusión dificulta la interpretación de los diferentes factores (**15**).
2. Decidir sobre en cuál de los posibles factores debe incluirse dicha variable. Generalmente, la variable suele introducirse dentro del factor con el que se encuentra más relacionada desde el punto de vista conceptual (**16**).

#### 16.8.1.3. Variables importantes en el análisis con baja correlación con el factor o los factores

Pueden existir variables que presentan bajas correlaciones con los factores identificados, pero que, sin embargo, son de importancia vital en la base de datos. En este caso no existe consenso entre los diferentes expertos, aunque se recomienda el mantenimiento de las mismas en la solución final.

### 16.8.2. Interpretación de los factores

No existe una prueba estadística para indicar si una variable es o no significativa para su inclusión o no en el o los factores identificados tras la realización del análisis factorial. Algunos autores sugieren que las variables con correlaciones ($a_{ij}$) inferiores a 0,30 tras una rotación ortogonal no sean incluidas en la definición del factor, debido a que menos del 9% de la variabilidad de dichas variables (colectividad = $(0{,}30)^2$) es explicada por dicho factor (**3**).

Se especifican los siguientes criterios:

| | | |
|---|---|---|
| $a_{ij} = 0{,}45$ | 20% de variabilidad | Pobre |
| $a_{ij} = 0{,}55$ | 30% de variabilidad | Buena |
| $a_{ij} = 0{,}63$ | 40% de variabilidad | Muy buena |
| $a_{ij} = 0{,}71$ | 50% de variabilidad | Excelente |

Se han identificado tres condiciones que facilitan el proceso de interpretación de los factores (**3**):

1. Cuanto mayor es el coeficiente de correlación entre la variable y el factor, mayor es el grado de solapamiento entre la variabilidad de la variable y el factor, y más parecido es el factor a la variable en cuestión.
2. Cuanto mayor es el número de variables con altos coeficientes de correlación con un factor, más sencillo resulta aislar lo que el factor probablemente representa.
3. Cuanto más pura es una variable que define a un factor, más sencillo resulta realizar inferencias sobre la naturaleza del factor.

## 16.9. ANÁLISIS FACTORIAL COMÚN FRENTE A ANÁLISIS FACTORIAL DE COMPONENTES PRINCIPALES

El AFC difiere del AFCP en el tratamiento que recibe la varianza total del conjunto de variables:

- *Análisis de componentes principales (AFCP):*
  - Esta primera aproximación establece que la varianza total de las variables es explicada por los factores extraídos.
  - Estima que la colectividad inicial para cada variable es 1.
  - Pueden obtenerse tantos factores como variables se incluyen en el análisis.
- *Análisis factorial común (AFC):*
  - Esta segunda aproximación separa la varianza total de las variables en dos componentes: uno de los componentes, denominado ***varianza común,*** es la cantidad de varianza que las variables comparten en común y que es reflejada (explicada) por uno o más factores; el otro componente, denominado ***varianza única,*** es exclusivo de cada variable y no participa en la identificación del factor o factores.

- Estima que la colectividad inicial para cada variable es siempre inferior a la unidad. Establece diferentes métodos para calcularla que darán lugar a diferentes aproximaciones; destaca el uso del coeficiente de correlación múltiple $R^2$ empleado en el análisis factorial principal.
- Se obtiene un número de factores inferior al número de variables incluidas en el análisis.

En el análisis de componentes principales, la varianza total de la muestra coincide con la varianza común. Sin embargo, en el AFC, se supone que la varianza de una variable puede ser explicada no solo por un número pequeño de componentes subyacentes, sino también por su varianza única.

Es decir, esta segunda aproximación se basa en la varianza común compartida por las variables y no en la varianza total de la muestra para la extracción de los factores.

Los factores extraídos no son, por tanto, meras combinaciones lineales de las variables estandarizadas incluidas en el análisis, sino hipotéticos componentes que son estimados a través de las variables originales.

Debido a esto, la colectividad de cada variable estandarizada $c_i$ es inferior a la unidad (a diferencia del análisis de componentes principales, donde el valor era 1). Por tanto, mediante este método de extracción siempre se generará un número inferior de factores al número original de variables incluidas en el análisis.

A diferencia del AFCP, el autovalor en el AFC no estima la cantidad de varianza total que es explicada por un factor, sino la cantidad de varianza común explicada por dicho factor. Tampoco al sumar los autovalores se obtendrá un valor igual al número de variables incluidas en el análisis, sino inferior.

STATA realiza por defecto el método factorial principal (un tipo de aproximación de AFC) cuando se aplica la instrucción:

```
factor
```

En el ejemplo:

```
fac alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner
```

STATA obtiene el siguiente listado:

```
Factor analysis/correlation                       Number of obs    =         40
    Method: principal factors                     Retained factors =          6
    Rotation: (unrotated)                         Number of params =         45
```

| Factor | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Factor1 | 5.22560 | 4.29427 | 0.7569 | 0.7569 |
| Factor2 | 0.93133 | 0.38324 | 0.1349 | 0.8918 |
| Factor3 | 0.54809 | 0.07749 | 0.0794 | 0.9712 |
| Factor4 | 0.47060 | 0.37088 | 0.0682 | 1.0393 |
| Factor5 | 0.09972 | 0.07502 | 0.0144 | 1.0538 |
| Factor6 | 0.02470 | 0.05559 | 0.0036 | 1.0574 |
| Factor7 | -0.03089 | 0.03446 | -0.0045 | 1.0529 |
| Factor8 | -0.06535 | 0.05285 | -0.0095 | 1.0434 |
| Factor9 | -0.11820 | 0.06332 | -0.0171 | 1.0263 |
| Factor10 | -0.18152 | . | -0.0263 | 1.0000 |

```
LR test: independent vs. saturated:  chi2(45) =  308.18 Prob>chi2 = 0.0000
```

```
Factor loadings (pattern matrix) and unique variances
```

| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Uniqueness |
|---|---|---|---|---|---|---|---|
| alcohol | 0.8747 | -0.2634 | 0.2140 | 0.1078 | -0.0408 | -0.0571 | 0.1032 |
| smoke | 0.9038 | -0.0541 | -0.0342 | 0.2663 | -0.1735 | -0.0013 | 0.0781 |
| cannabis | 0.8602 | -0.1614 | -0.1720 | 0.2866 | 0.0610 | 0.0709 | 0.1136 |
| junkfood | 0.7993 | 0.4244 | 0.0676 | -0.2239 | -0.0361 | -0.0182 | 0.1246 |
| soda_dr | 0.8308 | 0.3523 | 0.0068 | -0.2209 | -0.0574 | -0.0136 | 0.1333 |
| sedent_w | 0.1356 | 0.1536 | 0.4894 | 0.1128 | -0.0200 | 0.0674 | 0.7008 |
| sedent_l | 0.3664 | 0.5423 | 0.0142 | 0.2265 | 0.1805 | -0.0037 | 0.4876 |
| riskdriv | 0.7975 | -0.3397 | 0.0900 | -0.0417 | 0.1532 | -0.0687 | 0.2105 |
| celldriv | 0.6688 | -0.3018 | 0.0845 | -0.3706 | 0.0555 | 0.0795 | 0.3077 |
| partner | 0.5819 | 0.0676 | -0.4604 | -0.0622 | -0.0106 | 0.0169 | 0.4406 |

Aunque hipotéticamente existan 10 factores (tantos como variables), solo se presentan aquellos que poseen un autovalor superior a 0, en este caso seis factores. En esta aproximación, STATA usa por defecto un criterio de autovalor >0 (**mineigen (0)**), a diferencia del análisis de componentes principales, en el que por defecto STATA extraía solo los factores con autovalor >1 (**mineigen (1)**).

Puede observarse que ni los autovalores, ni los coeficientes *(factor loadings)* ni las comunalidades (en realidad, los valores de 1 – comunalidad) de las variables coinciden con los obtenidos tras realizar la aproximación AFCP. En este caso, la comunalidad inicial de las variables no es 1, sino $R^2$.

Puede solicitarse a STATA el valor del coeficiente de determinación para cada variable a través de la instrucción:

```
estat smc
```

```
Squared multiple correlations of variables with all other variables
```

| variable | smc |
|---|---|
| alcohol | 0.8568 |
| smoke | 0.8868 |
| cannabis | 0.8503 |
| junkfood | 0.8308 |
| soda_dr | 0.8365 |
| sedent_w | 0.2515 |
| sedent_l | 0.4770 |
| riskdriv | 0.7558 |
| celldriv | 0.6540 |
| partner | 0.5047 |

Obsérvese que el valor de $R^2$ para el alcohol (0,8568) coincide con el valor 1-0,1032 *(uniqueness)* para esta variable

## 16.10. ANÁLISIS FACTORIAL CONFIRMATORIO FRENTE AL EXPLORATORIO

En este capítulo se ha analizado en profundidad el análisis factorial desde la perspectiva del análisis exploratorio. Este tipo de análisis factorial es utilizado cuando el investigador desconoce cuántos factores serán necesarios para explicar las interrelaciones entre un grupo de características, indicadores o variables (**13**). Este tipo de análisis factorial es el más común en la investigación del área biosanitaria. Este tipo de análisis se encuentra disponible en los programas de *software* estadísticos tradicionales (p. ej., STATA).

Por otro lado, el análisis factorial confirmatorio se utiliza para determinar el grado en el que un grupo de factores identificados *a priori* son capaces de representar los datos de una matriz. Este método es empleado cuando el investigador tiene cierto conocimiento sobre la estructura subyacente a una serie de datos. Este tipo de análisis requiere un conocimiento exhaustivo de la estructura de covarianzas, por lo que es necesaria la utilización de programas de *software* estadísticos más complejos.

## 16.11. DIFERENTE APROXIMACIÓN EN STATA PARA REALIZAR UN ANÁLISIS FACTORIAL DE COMPONENTES PRINCIPALES

Si se desea realizar un AFCP, existe otra alternativa en STATA. Debe acudirse al menú:

**Statistics → Multivariate analysis → Factor and Principal components analysis → Principal component analysis (PCA)**

o a la instrucción:

```
pca  v1 v2 v3 vp
```

siendo $v_1$-$v_p$ las variables implicadas en el análisis.

En este caso, STATA construirá tantos factores como variables se incluyan en el análisis. Por tanto, se debe especificar el número de factores que se desea extraer o el valor límite de autovalor que se va a considerar a través de las subinstrucciones:

```
components()
```

y/o

```
mineigen()
```

En el ejemplo, si solo se desearan aquellos factores con autovalor superior a 1:

```
pca alcohol smoke cannabis junkfood soda_dr sedent_w
sedent_l riskdriv celldriv partner, mine (1)
```

| Component | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| Comp1 | 5.43802 | 4.12459 | 0.5438 | 0.5438 |
| Comp2 | 1.31343 | .178291 | 0.1313 | 0.6751 |
| Comp3 | 1.13514 | .425311 | 0.1135 | 0.7887 |
| Comp4 | .709825 | .202143 | 0.0710 | 0.8596 |
| Comp5 | .507681 | .113416 | 0.0508 | 0.9104 |
| Comp6 | .394266 | .165846 | 0.0394 | 0.9498 |
| Comp7 | .228419 | .110023 | 0.0228 | 0.9727 |
| Comp8 | .118397 | .0317622 | 0.0118 | 0.9845 |
| Comp9 | .0866346 | .0184378 | 0.0087 | 0.9932 |
| Comp10 | .0681968 | . | 0.0068 | 1.0000 |

Principal components (eigenvectors)

| Variable | Comp1 | Comp2 | Comp3 | Unexplained |
|---|---|---|---|---|
| alcohol | 0.3770 | -0.1504 | 0.2543 | .1241 |
| smoke | 0.3875 | -0.0243 | 0.0253 | .1818 |
| cannabis | 0.3717 | -0.1215 | -0.0416 | .2275 |
| junkfood | 0.3474 | 0.3004 | -0.1032 | .2132 |
| soda_dr | 0.3604 | 0.2253 | -0.1232 | .2097 |
| sedent_w | 0.0657 | 0.4649 | 0.7013 | .1343 |
| sedent_l | 0.1709 | 0.6491 | -0.2544 | .2143 |
| riskdriv | 0.3506 | -0.2717 | 0.1884 | .1943 |
| celldriv | 0.3001 | -0.3181 | 0.2063 | .329 |
| partner | 0.2692 | -0.0929 | -0.5219 | .2852 |

Puede observarse que los valores de autovalor y de singularidad (1 – comunalidad) coinciden con los obtenidos al emplear la instrucción **factor, pcf.**

Sin embargo, con la instrucción **pca,** STATA no presenta los *factor loadings* (coeficientes de correlación variable-factor), sino los denominados *eigenvectors* (vectores de autovalor). Estos vectores se definen como el cociente entre el coeficiente de correlación de cada variable y cada factor *(factor loading)* y la raíz cuadrada del autovalor.

El eigenvector del *factor 1* y el *alcohol* es, por ejemplo, 0,377. Este valor corresponde al cociente entre 0,8791 (correlación factor 1-alcohol) y la raíz cuadrada de 5,438 (autovalor del factor 1):

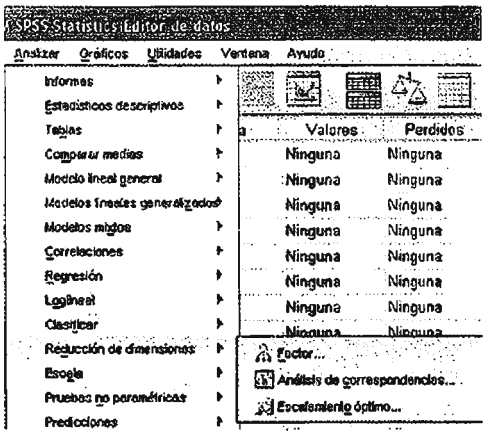
$$0{,}377 = 0{,}8791 / \sqrt{5{,}438}$$

El resto de las instrucciones y la interpretación de resultados son similares a lo expuesto a lo largo del presente capítulo.

## 16.12. ANÁLISIS FACTORIAL DE COMPONENTES PRINCIPALES CON SPSS

En SPSS, el AFCP puede realizarse a través del menú con la instrucción:

**Analizar → Reducción de dimensiones → Factor**




SPSS presentará las comunalidades, los autovalores *(eigenvalues)* y el porcentaje de varianza explicada, así como la matriz de componentes que muestra los correspondientes coeficientes de correlación *(factor loadings)* entre cada uno de los nuevos factores obtenidos y las variables originales. Los resultados coinciden con los obtenidos con el programa STATA.

**Varianza total explicada**

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 5,438 | 54,380 | 54,380 | 5,438 | 54,380 | 54,380 |
| 2 | 1,313 | 13,134 | 67,514 | 1,313 | 13,134 | 67,514 |
| 3 | 1,135 | 11,351 | 78,866 | 1,135 | 11,351 | 78,866 |
| 4 | ,710 | 7,098 | 85,964 | | | |
| 5 | ,508 | 5,077 | 91,041 | | | |
| 6 | ,394 | 3,943 | 94,984 | | | |
| 7 | ,228 | 2,284 | 97,268 | | | |
| 8 | ,118 | 1,184 | 98,452 | | | |
| 9 | ,087 | ,866 | 99,318 | | | |
| 10 | ,068 | ,682 | 100,000 | | | |

Método de extracción: Análisis de Componentes principales.

**Comunalidades**

| | Inicial | Extracción |
|---|---|---|
| alcohol | 1,000 | ,876 |
| smoke | 1,000 | ,818 |
| cannabis | 1,000 | ,773 |
| junkfood | 1,000 | ,787 |
| soda_dr | 1,000 | ,790 |
| sedent_w | 1,000 | ,866 |
| sedent_l | 1,000 | ,786 |
| riskdriv | 1,000 | ,806 |
| celldriv | 1,000 | ,671 |
| partner | 1,000 | ,715 |

Método de extracción: Análisis de Componentes principales.

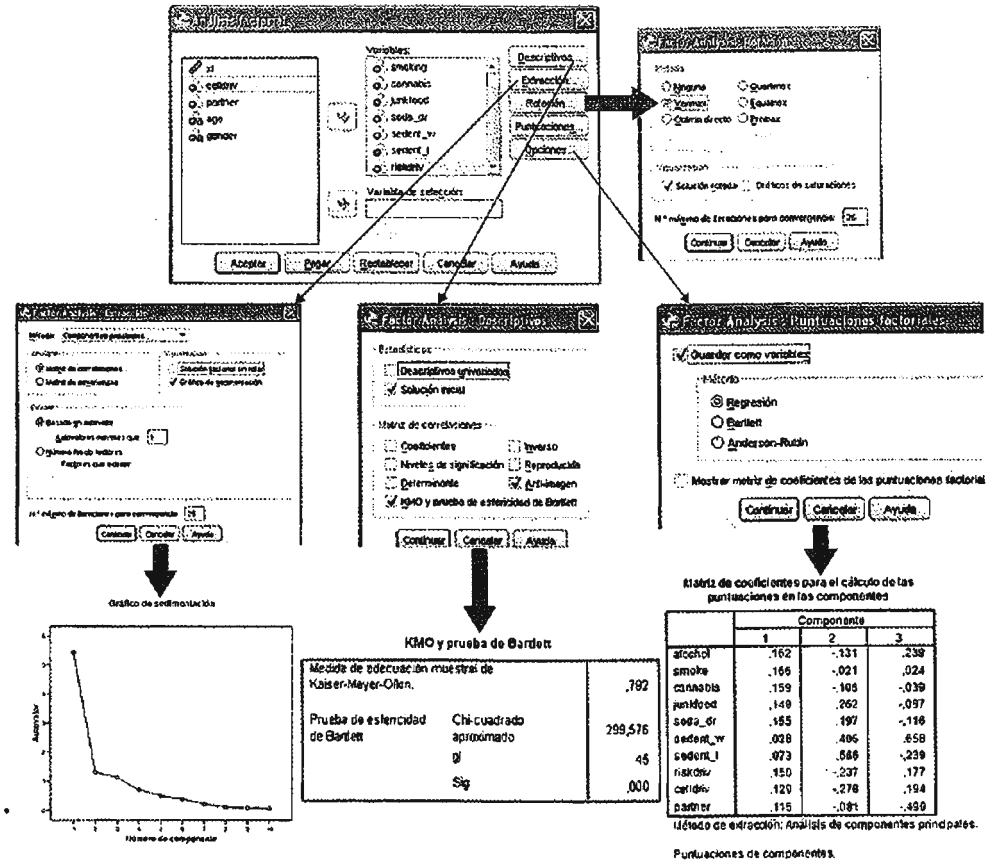
**Matriz de componentes[a]**

| | Componente | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| alcohol | ,879 | -,172 | ,271 |
| smoke | ,904 | -,028 | ,027 |
| cannabis | ,867 | -,139 | -,044 |
| junkfood | ,810 | ,344 | -,110 |
| soda_dr | ,840 | ,258 | -,131 |
| sedent_w | ,153 | ,533 | ,747 |
| sedent_l | ,399 | ,744 | -,271 |
| riskdriv | ,818 | -,311 | ,201 |
| celldriv | ,700 | -,365 | ,220 |
| partner | ,628 | -,106 | -,556 |

Método de extracción: Análisis de componentes principales.

a. 3 componentes extraídos

A través de los botones colocados en la parte superior derecha pueden verificarse las condiciones de aplicación del AFCP (DESCRIPTIVOS); es posible determinar el número de factores para extraer (EXTRACCIÓN), pueden rotarse los factores para una mejor interpretación de los valores (ROTACIÓN) o construir los factores (PUNTUACIONES).




## 16.13. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| **AFCP** | **factor** $v_1\ v_2\ v_3\ v_p$, **pcf** | **FACTOR** |
| | | **/VARIABLES** $v_1\ v_2\ v_3\ v_p$ |
| | **pca** $v_1\ v_2\ v_3\ v_p$ | **/EXTRACTION PC** |
| **AFC** | | |
| AFP | **factor** $v_1\ v_2\ v_3\ v_p$, **pf** | **/EXTRACTION PAF** |
| MV | **factor** $v_1\ v_2\ v_3\ v_p$, **ml** | **/EXTRACTION ML** |
| Alfa | — | **/EXTRACTION ALPHA** |
| Imagen | — | **/EXTRACTION IMAGEN** |
| **Gráfico de sedimentación** | **screeplot** | **/PLOT EIGEN** |
| **Rotación** | | |
| Sin rotar | p.d. | **/ROTATION NOROTATE** |
| Varimax | **rotate, varimax** | **/ROTATION VARIMAX** |
| Quartimax | **rotate, quartimax** | **/ROTATION QUARTIMAX** |
| Equamax | **rotate, equamax** | **/ROTATION EQUAMAX** |

| Procedimiento | STATA | SPSS |
|---|---|---|
| **N.° de factores** | | |
| Autovalor >1 | **mineigen** | **/CRITERIA MINEIGEN(1)** |
| N.° determinado | **factors (#)** | **/CRITERIA FACTORS(#)** |
| **Medidas de adecuación** | **estat kmo**<br>**estat anti** | **/PRINT DET KMO AIC** |
| **Guardar factores** | **predict** $f_1 f_2 f_3$<br>**regression**<br>**predict** $f_1 f_2 f_3$<br>**bartlett** | **/SAVE REG(ALL)**<br>**/SAVE BART(ALL)** |
| Ejemplo del capítulo con análisis de componentes principales, con rotación *Varimax*, extracción de dos factores y creación de los mismos según un método de regresión | **fac *alcohol smoke cannabis junkfood soda_dr sedent_w sedent_l riskdriv celldriv partner*,**<br>**pcf fa (*2*)**<br>**rot**<br>**predict *factor1* *factor2*, r** | **FACTOR**<br>**/VARIABLES *alcohol smoke cannabis junkfood soda_dr sedent_w sedent_l riskdriv celldriv partner***<br>**/CRITERIA FACTORS(*2*)**<br>**/EXTRACTION PC**<br>**/ROTATION VARIMAX**<br>**/SAVE REG(ALL).** |

#, n.° de factores; AFCP, análisis factorial de componentes principales; AFC, análisis factorial común, AFP, análisis factorial principal; f, nombre que desea darse a cada factor; MV, máxima verosimilitud; p.d., por defecto (el programa calcula el parámetro sin añadir instrucciones); *v*, variables.

## REFERENCIAS


1. Goddard J, Kirby A. An introduction to factor analysis. Norwich: Geo Abstracts; 1976.
2. Harman HH. Modern factor analysis. 3rd ed. Chicago: Universidad de Chicago; 1976.
3. Comrey AL, Lee HB. A first course in factor analysis. Hillsdale: Lawrence Erlbaum; 1992.
4. Kim J, Mueller CW. Factor analysis: Statistical methods and practical issues. Beverly Hills: Sage; 1978.
5. Guttman L. Some necessary conditions for common-factor analysis. Psychometrika 1954;19:149-61.
6. Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas 1960;20:141-51.
7. Gorsuch RL. Factor analysis. 2nd ed. Hillsdale: Lawrence Erlbaum; 1983.
8. Cattell RB. The scree test for the number of factors. Multivariate Behav Res 1966;1:245-76.
9. Cattell RB, Jaspars J. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. Multivariate Behav Res Monogr 1967;67:3.
10. Bartlett MS. Test of significance in factor analysis. Br J Psychology 1950;3:77-85.
11. Kaiser HF. An index of factorial simplicity. Psychometrika 1974;39:32-6.
12. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.

13. Tabachnick BG, Fidell LS. Using multivariate statistics. 4th ed. Boston: Allyn & Bacon; 2001.

14. Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate data analysis with readings. 4th ed. Englewood Cliffs: Prentice Hall; 1995.

15. Kline P. The handbook of psychological testing. 2nd ed. London: Routledge; 2000.

16. Pett MA, Lackey NR, Sullivan JJ. Making sense of factor analysis. Thousand Oaks: Sage Publications; 2003.

# ANÁLISIS DE CLÚSTERES O CONGLOMERADOS

17

*A. Sánchez-Villegas, F. Guillén-Grima, M. Á. Martínez-González*

## 17.1. INTRODUCCIÓN Y CONCEPTO

En algunas ocasiones se dispone de una serie de observaciones (personas, animales, etc.) y se desea clasificarlas en grupos. Así sucede, por ejemplo, con las plantas. Los primeros sistemas de clasificación de plantas se deben a Linneo, quien, en el siglo XVIII, creó una sistematización basada en las semejanzas y diferencias entre las especies vegetales. El mismo enfoque se aplicó a la medicina y, durante cientos de años, los criterios para la clasificación de enfermedades se han basado en semejanzas y diferencias entre distintos síndromes o entidades clínicas. El problema de este tipo de métodos puede residir en la subjetividad en la creación de los grupos. Para prevenir posibles arbitrariedades, existen métodos estadísticos de clasificación, fundamentalmente las técnicas de *clúster*. Aunque el término *cluster analysis* fue utilizado por primera vez por Tryion en 1939, solo a finales de la década de los sesenta y principios de los setenta del siglo XX comenzaron a desarrollarse técnicas de clasificación automática que se han dado en llamar análisis de clúster o análisis de conglomerados.

Diferentes programas estadísticos permiten realizar un análisis de conglomerados. Uno de los habituales es STATA. Este análisis puede realizarse desde el menú o con instrucciones a través de la sintaxis del programa. Desde el menú:

**Statistics → Multivariate analysis → Cluster analysis**

Con instrucción:

```
cluster
```

El análisis de conglomerados intenta determinar la agrupación natural de diferentes observaciones, estableciendo el grado de similitud o diferencia entre ellas (semejanzas o diferencias). Existen numerosas técnicas para calcular las distancias (medidas de semejanza o de diferencia) entre datos, tanto de variables cuantitativas como de variables dicotómicas.

Se manejan diferentes tipos de análisis de conglomerados, cada uno de los cuales posee métodos específicos. Por ejemplo, algunos métodos buscan reducir la heterogeneidad dentro del grupo, es decir, tratan de que los conglomerados sean homogéneos (es decir, con sujetos similares dentro de un grupo) y diferentes de otros clústeres (los sujetos de un grupo serán diferentes de los de otro); después, subdividen los grupos hasta que se alcanza un número óptimo de clústeres. En otros métodos, el número de conglomerados es fijado *a priori* por el investigador y puede ser reducido o aumentado dependiendo de las características de los datos. Los resultados obtenidos difieren notablemente según la técnica utilizada; por ello, es muy importante detallar qué técnicas se han empleado y por qué (1).

## 17.2. TIPOS DE ANÁLISIS DE CLÚSTER

Básicamente se realizan dos grupos de análisis: el clúster de K medias o K medianas, conocido en inglés como *partition method*, y el clúster jerárquico, denominado en inglés *hierarchical method*.

### 17.2.1. Clúster de K medias y clúster de K medianas

En este tipo de análisis, el investigador decide de antemano el número de grupos (*k* grupos) que van a formarse. Para entender correctamente este análisis, resulta prioritario definir primero el concepto de centroide. Se llama centroide a la medida de tendencia central que se usa para describir el clúster y cuyo valor es comparado con el valor que toma cada dato. Si se emplea la media como centroide, se realizará un análisis de K medias; si se elige la mediana, el análisis será de K medianas (2).

Cada observación es asignada al clúster o grupo que posee un centroide con un valor más cercano (o similar). Al introducirse un nuevo dato en el clúster, el valor del centroide de dicho clúster cambia automáticamente, con lo que algunos individuos podrían cambiar de centroide. Cada vez que se cambia a un individuo de grupo, hay que recalcular los centroides. Los cálculos se repiten hasta que ningún dato cambie de grupo.

Este método es más rápido y permite trabajar con bases de datos de mayor número de individuos que el análisis jerárquico, ya que no necesita especificar una matriz de distancias (v. apartado 17.3.1).

### 17.2.2. Clúster jerárquico

Dentro del método jerárquico hay dos variedades: el método jerárquico *aglomerativo* o el jerárquico *divisivo*.

En el método jerárquico *aglomerativo* se considera que cada caso es un clúster. Es decir, el análisis comienza con *N* clústeres de tamaño 1, donde *N* es el tamaño muestral. A continuación, los casos se agrupan según su similitud hasta que todos forman un único clúster.

El procedimiento *divisivo* empieza al revés. Todos los casos forman un único clúster y los casos se van separando. Este procedimiento, menos utilizado, puede resultar menos eficiente. Los resultados obtenidos por estos dos métodos pueden ser diferentes.

## 17.3. MÉTODO PARA LA FORMACIÓN DE CONGLOMERADOS

### 17.3.1. Determinar las medidas de similitud o disimilaridad (distancia entre observaciones)

El primer paso para llevar a cabo un análisis de clúster consiste en determinar la distancia existente entre las observaciones que se van a agrupar. En las variables cuantitativas, las medidas de similitud o disimilaridad más utilizadas son la distancia euclidiana (también denominada euclídea) y la distancia euclidiana al cuadrado[1].

A continuación se recurrirá a un ejemplo para mostrar cómo se calcula la distancia euclidiana. Supóngase que se desea diseñar un clúster para agrupar las comunidades autónomas (CC. AA.) en función de su producto interior bruto (PIB). El objetivo es formar grupos de CC. AA., de modo que, dentro de cada grupo, los PIB de las CC. AA. sean homogéneos (homogeneidad intraclúster en el PIB). Para simplificar el ejemplo, se parte de una lista de siete CC. AA., junto con su PIB *per capita*, ajustado con el nivel de la Unión Europea (100%). Se podría comenzar, por ejemplo, con el PIB *per cápita* de Murcia y restarle el valor de cada una de las regiones. El problema sería que se obtendrían unas diferencias con valor positivo y otras con valor negativo, cuando lo que interesa es la distancia absoluta, sin signo. Para evitar la aparición de diferencias negativas, se calculará la

1 Otras medidas de similitud para variables cuantitativas son la distancia de Minkowski o la de Mahalanobis. Cuando la variable es categórica con dos grupos (p. ej., que un tumor presente o no una determinada mutación), se estiman las distancias entre observaciones mediante el coeficiente de emparejamiento simple para datos binarios u otros índices, como el de Jaccard o el de Dice-Sorensen. Estos métodos empleados para variables dicotómicas exceden los objetivos del presente capítulo.

*Tabla 17.1 Distancias euclidianas entre Murcia y otras comunidades autónomas*

| COMUNIDAD AUTÓNOMA | PIB PER CÁPITA | DIFERENCIA | DISTANCIA EUCLIDIANA |
|---|---|---|---|
| Andalucía | 72 | 11 | 11 |
| Aragón | 103 | −20 | 20 |
| Asturias | 81 | 2 | 2 |
| Cantabria | 93 | −10 | 10 |
| Castilla-La Mancha | 77 | 6 | 6 |
| Castilla y León | 89 | −6 | 6 |
| Murcia | 83 | 0 | 0 |

*Tabla 17.2 Matriz de distancias (o matriz de disimilitudes) para el ejemplo de las comunidades autónomas*

| | ANDALUCÍA | ARAGÓN | ASTURIAS | CANTABRIA | CASTILLA-LA MANCHA | CASTILLA Y LEÓN |
|---|---|---|---|---|---|---|
| Aragón | 31 | | | | | |
| Asturias | 9 | 22 | | | | |
| Cantabria | 21 | 10 | 12 | | | |
| Castilla-La Mancha | 5 | 26 | 4 | 16 | | |
| Castilla y León | 17 | 14 | 8 | 4 | 12 | |
| Murcia | 11 | 20 | 2 | 10 | 6 | 6 |

distancia euclídea, que consiste en elevar al cuadrado el valor de la diferencia, y a continuación extraer la raíz cuadrada.

$$\text{Distancia euclidiana}\,(x, y) = \sqrt{\sum (x - y)^2}$$

Obsérvese el símbolo de sumatorio en la fórmula: de igual forma que se actúa con una variable *PIB*, se podría aplicar con otras variables para las que también se desea una homogeneidad intraclúster (tabla 17.1).

Una vez calculada la distancia entre Murcia y el resto de CC. AA., el siguiente paso es elaborar una matriz de distancias. Se trata de construir una tabla en la que se reflejen las distancias entre cada par de casos (no solo Murcia-resto). En la tabla 17.2 se presenta la matriz de distancias euclidianas entre las siete CC. AA.

Por defecto, si no se especifica nada más, STATA calculará la distancia euclidiana, denominada *L2*. Si se desea calcular otra distancia diferente, esta deberá ser especificada mediante la subinstrucción:

```
measure ()
```

Algunos ejemplos son:

```
mea(L2squared)  #distancia euclidiana al cuadrado

mea (L1)    #Minkowski

mea (matc)        #coeficiente de emparejamiento simple

mea (Jac)  #Jaccard
```

### 17.3.2. Elegir el método de unión de conglomerados

Existen diferentes formas para vincular las observaciones y crear conglomerados con homogeneidad interna en cuanto a la o las variables que se usan para construir los clústeres. Una vez definida la distancia euclidiana para cada par de observaciones, la unión de observaciones y la creación de clústeres pueden establecerse según diferentes aproximaciones.

#### 17.3.2.1. En el análisis de clúster jerárquico

**Mínima distancia o vecino más próximo *(single linkage)*.** El criterio para unir observaciones (y/o clústeres) es la distancia mínima entre los dos puntos más cercanos de dos clústeres. Véase el ejemplo específico de las CC. AA.

En un análisis de clúster aglomerativo, al principio cada observación forma un clúster, por lo que la mínima distancia se encontraría entre Murcia y Asturias, con una distancia euclidiana de 2 (v. tabla 17.2, matriz de distancias). Por tanto, Asturias y Murcia formarán un clúster. Una vez creado este clúster, la matriz de distancias debe ser actualizada. Se debe buscar ahora la distancia mínima entre el clúster formado por Asturias y Murcia y cada uno de los otros clústeres (en este caso, formados por un elemento, la propia comunidad). Esta distancia será establecida eligiendo la distancia euclidiana de Asturias o de Murcia, según cuál sea menor (tabla 17.3, datos en negrita). Las restantes distancias no se han modificado con respecto a la matriz original.

La tabla 17.3 informa de que los clústeres más cercanos son ahora Cantabria y Castilla y León ($d = 4$), que formarán otro clúster, o el conglomerado Asturias-Murcia con Castilla-La Mancha ($d = 4$, igualmente). Se elegirá el primer caso porque resulta más sencillo. Se repite de nuevo la matriz, ahora con cinco conglomerados: dos con dos observaciones (Asturias-Murcia; Cantabria-Castilla y León) y tres con una sola observación (Andalucía, Aragón y Castilla-La Mancha). Nuevamente, debe elegirse la menor distancia observada entre el nuevo clúster creado (Cantabria y Castilla y León) y cada uno de los clústeres restantes. Dicha distancia será la menor dentro del clúster con el resto de grupos, ya sea la de Cantabria o la de Castilla y León. Para la comparación entre los clústeres Asturias-Murcia y Cantabria-Castilla y León existen cuatro distancias: 12, Cantabria-Asturias; 10, Murcia-Cantabria; 8, Asturias-Castilla y León; 6, Murcia-Castilla y León. Se elegirá la distancia 6, que es la menor observada (tabla 17.4).

Castilla-La Mancha debe ser incluida en el conglomerado Asturias-Murcia por presentar la mínima distancia entre conglomerados ($d = 4$), y así sucesivamente. Cuando termina el algoritmo, todos los objetos se encuentran en un único clúster.

**Máxima distancia o vecino más lejano *(complete linkage)*.** En este caso, como criterio de unión se utiliza la distancia mínima entre los dos puntos más separados de dos clústeres. Véase el ejemplo específico de las CC. AA.

***Tabla 17.3 Matriz de distancias mediante el uso de la mínima distancia o vecino más próximo para el ejemplo de las comunidades autónomas (I)***

| | DISTANCIA EUCLIDIANA | | | | | |
|---|---|---|---|---|---|---|
| | ANDALUCÍA | ARAGÓN | ASTURIAS-MURCIA | CANTABRIA | CASTILLA-LA MANCHA | CASTILLA Y LEÓN |
| Andalucía | 0 | 31 | **9** | 21 | 5 | 17 |
| Aragón | 31 | 0 | 20 | 10 | 26 | 14 |
| **Asturias-Murcia** | **9** | 20 | **0** | **10** | 4 | 6 |
| Cantabria | 21 | 10 | **10** | 0 | 16 | 4 |
| Castilla-La Mancha | 5 | 26 | 4 | 16 | 0 | 12 |
| Castilla y León | 17 | 14 | 6 | 4 | 12 | 0 |

*Tabla 17.4 Matriz de distancias mediante el uso de la mínima distancia o vecino más próximo para el ejemplo de las comunidades autónomas (II)*

| | DISTANCIA EUCLIDIANA | | | |
|---|---|---|---|---|
| | ANDALUCÍA | ARAGÓN | ASTURIAS-MURCIA | CANTABRIA-CASTILLA Y LEÓN |
| Aragón | 31 | | | |
| Asturias-Murcia | 9 | 20 | | |
| **Cantabria-Castilla y León** | 17 | 10 | 6 | |
| Castilla-La Mancha | 5 | 26 | 4 | 12 |

*Tabla 17.5 Matriz de distancias mediante el uso de la máxima distancia o vecino más lejano para el ejemplo de las comunidades autónomas (I)*

| | DISTANCIA EUCLIDIANA | | | | |
|---|---|---|---|---|---|
| | ANDALUCÍA | ARAGÓN | ASTURIAS-MURCIA | CANTABRIA | CASTILLA-LA MANCHA |
| Aragón | 31 | | | | |
| **Asturias-Murcia** | 11 | 22 | | | |
| Cantabria | 21 | 10 | 12 | | |
| Castilla-La Mancha | 5 | 26 | 6 | 16 | |
| Castilla y León | 17 | 14 | 8 | 4 | 12 |

Una vez establecido el clúster Asturias-Murcia, se debe buscar la distancia máxima entre el clúster formado por Asturias y Murcia y cada uno de los otros clústeres. Esta distancia será establecida eligiendo la distancia euclidiana de Asturias o de Murcia, según cuál sea mayor con respecto al resto de comunidades. Por ejemplo, la distancia entre Asturias y Andalucía es 9 y entre Murcia y Andalucía es 11 (v. tabla 17.2). Según esta aproximación, por tanto, se elegirá la distancia 11.

La tabla 17.5 presenta los datos para todas las posibles comparaciones entre CC. AA. siguiendo la aproximación de la máxima distancia.

De nuevo, la distancia euclidiana menor es la correspondiente a la comparación de los datos de Cantabria y de Castilla y León ($d = 4$). Tras crear el clúster Cantabria-Castilla y León, la matriz de distancias tomará los siguientes valores siguiendo el criterio de la máxima distancia. Para la comparación entre los conglomerados Asturias-Murcia y Cantabria-Castilla y León existen cuatro distancias (12, Cantabria-Asturias; 10, Murcia-Cantabria; 8, Asturias-Castilla y León; 6, Murcia-Castilla y León). En este caso se elige la distancia 12, la máxima observada entre conglomerados (tabla 17.6).

El siguiente clúster que se creará será el formado por Castilla-La Mancha-Andalucía ($d = 5$), y así sucesivamente.

**Distancia media o vinculación entre grupos *(average linkage)*.** Consiste en calcular la distancia entre cada uno de los puntos de un clúster con cada punto de otro clúster, y obtener la media de dichas distancias.

*Tabla 17.6 Matriz de distancias mediante el uso de la máxima distancia o vecino más lejano para el ejemplo de las comunidades autónomas (II)*

| | DISTANCIA EUCLIDIANA | | | |
|---|---|---|---|---|
| | ANDALUCÍA | ARAGÓN | ASTURIAS-MURCIA | CANTABRIA-CASTILLA Y LEÓN |
| Aragón | 31 | | | |
| Asturias-Murcia | 11 | 22 | | |
| **Cantabria-Castilla y León** | 21 | 14 | 12 | |
| Castilla-La Mancha | 5 | 26 | 6 | 16 |

*Tabla 17.7 Distancia media entre el clúster Asturias-Murcia y el clúster Cantabria-Castilla y León*

| | | DISTANCIAS |
|---|---|---|
| Asturias | Cantabria | 12 |
| Asturias | Castilla y León | 8 |
| Murcia | Cantabria | 10 |
| Murcia | Castilla y León | 6 |
| Distancia media | | 9 |

Por ejemplo, si se desease calcular la distancia entre los clústeres Asturias-Murcia y Cantabria-Castilla y León, se calcularían las distancias entre cada uno de los puntos de un clúster con las de cada uno de los puntos del otro. Se obtendrían así las distancias presentadas en la tabla 17.7. La media de estas distancias es 9.

Este mismo procedimiento se repetiría para hallar la distancia media entre el clúster Asturias-Murcia y cada uno de los demás clústeres, y para determinar la distancia media entre el clúster Cantabria-Castilla y León y el resto de conglomerados. En resumen, de esta manera puede establecerse la distancia (medidas de similitud) entre, por ejemplo, los clústeres de Asturias-Murcia y de Cantabria-Castilla y León a través de tres aproximaciones distintas:

Aproximación de distancia más corta: 6.

Aproximación de distancia más larga: 12.

Aproximación de distancia media: 9.

Existen otros métodos de unión, como el de Ward, el del centroide *(centroid)*, el de la mediana *(median)* o el de la media ponderada *(weighted average)*, que no se abordarán en el presente capítulo.

- Desde STATA pueden llevarse a cabo todos los procedimientos de unión descritos en este apartado.

Desde el menú:

**Statistics → Multivariate analysis → Cluster analysis → Cluster data-Single linkage**

**Statistics → Multivariate analysis → Cluster analysis → Cluster data → Complete linkage**

**Statistics → Multivariate analysis → Cluster analysis → Cluster data → Average linkage**

Desde las instrucciones:

```
cluster singlelinkage v1 v2 v3 vp, name
(nombre del análisis)

cluster completelinkage v1 v2 v3 vp, name
(nombre del análisis)

cluster averagelinkage v1 v2 v3 vp, name
(nombre del análisis)
```

$v_1$-$v_p$ son las variables implicadas en el análisis.

*Tabla 17.8 Distancia euclidiana de cada observación (comunidades autónomas) a los dos centroides (A y B)*

| DATO | CLÚSTER INICIAL | DISTANCIA EUCLIDIANA CON CENTROIDE DEL CLÚSTER A | DISTANCIA EUCLIDIANA CON CENTROIDE DEL CLÚSTER B |
|---|---|---|---|
| Andalucía | A | 13,3 | 13,5 |
| Aragón | A | 17,7 | 17,5 |
| Asturias | A | 4,3 | 4,5 |
| Cantabria | B | 7,6 | 7,5 |
| Castilla-La Mancha | B | 8,3 | 8,5 |
| Castilla y León | B | 3,6 | 3,5 |
| Murcia | B | 2,3 | 2,5 |

Por defecto, STATA aplica el método de la mínima distancia. En el ejemplo de las CC. AA., se desea denominar al análisis «riqueza» y crear clústeres a partir de una sola variable *PIB*.

```
cluster singlelinkage    PIB, name(riqueza)
```

o su versión reducida:

```
cluster s PIB, n(riqueza)
```

### 17.3.2.2. En el análisis de clúster de K medias o K medianas

Después de establecer el número de clústeres que se desean, se eligen al azar los elementos integrantes de los posibles clústeres, asignando a cada uno un número similar de elementos. Véase el ejemplo de las CC. AA. y su PIB.

En primer lugar, debe establecerse el número de conglomerados que se crearán, por ejemplo dos, el A y el B. Estos dos primeros clústeres se eligen al azar. Por ejemplo, uno de ellos, el clúster A, integraría a Andalucía, Aragón y Asturias, y el B, a Cantabria, Castilla-La Mancha, Castilla y León y Murcia. A continuación se calculan los centroides de cada clúster.

Si se elige el método de K medias, el centroide será la media.

$$\text{Centroide clúster A} = (72 + 103 + 81) / 3 = 85,3$$
$$\text{Centroide clúster B} = (93 + 77 + 89 + 83) / 4 = 85,5$$

A continuación, se establece la distancia euclidiana de cada observación a los dos centroides:

Para Andalucía, con respecto al centroide del clúster A, será, por ejemplo:
$(72\text{-}85,3)^{2*0,5} = 13,3$.

Para el centroide del clúster B se obtendrá: $(72\text{-}85,5)^{2*0,5} = 13,5$.

Para el resto de datos, la distancia a cada clúster puede verse en la tabla 17.8.

Cada observación es asignada al clúster que posee un centroide con un valor más cercano (menor distancia euclidiana al valor de dicho dato). En consecuencia, algunos datos colocados de manera arbitraria en el clúster A o en el B pueden cambiar de centroide. Así sucede con Aragón (cambia de A a B) y Castilla-La Mancha y Murcia (de B a A). Cada vez que un individuo cambia de grupo, hay que recalcular los centroides. Los cálculos se repiten hasta que ningún dato cambie de grupo.

Desde STATA, el análisis se realizará con el menú:

**Statistics → Multivariate analysis → Cluster analysis → Cluster data-kmeans**

**Statistics → Multivariate analysis → Cluster analysis → Cluster data-kmedians**

o con las instrucciones:

```
cluster kmeans v1 v2 v3 vp, k(n.º de clústeres
que se desean) name(nombre del análisis)

cluster kmedians v1 v2 v3 vp, k(n.º de clústeres
que se desean) name (nombre del análisis)
```

En el ejemplo:

```
cluster kmeans PIB, k(2) name(riqueza)
```

o

```
cluster k PIB, k(2) n(riqueza)
```

## 17.4. GRÁFICOS DEL ANÁLISIS DE CLÚSTER: DENDROGRAMAS

El gráfico obtenido tras un análisis de conglomerados se denomina dendrograma. Los dendrogramas pueden representarse de forma horizontal o vertical.

En el ejemplo de las CC. AA., el dendrograma se iniciaría uniendo Asturias y Murcia (fig. 17.1).

En el caso de un dendrograma vertical, en la parte baja del mismo se colocan las observaciones individuales, que son consideradas sus propios clústeres. Conforme los valores se agrupan en conglomerados, se conectan mediante líneas verticales, a su vez unidas con líneas verticales procedentes de otros clústeres a través de líneas horizontales según se incrementan las distancias representadas en el eje vertical.

En un dendrograma horizontal, las observaciones se disponen a la izquierda de la gráfica y, conforme se agrupan en clústeres, son conectadas con líneas horizontales que, a su vez, se unen con líneas horizontales procedentes de otros clústeres a través de líneas verticales según diferentes valores de distancia representados en el eje horizontal.

A continuación se muestra un ejemplo de dendrograma vertical y horizontal con las 19 comunidades autónomas (fig. 17.2).

STATA permite la construcción de dendrogramas. Desde el menú:

**Statistics → Multivariate analysis → Cluster analysis → Postclustering → Dendrograms**

Desde las instrucciones:

```
cluster dendrogram
```

o

```
cluster tree
```

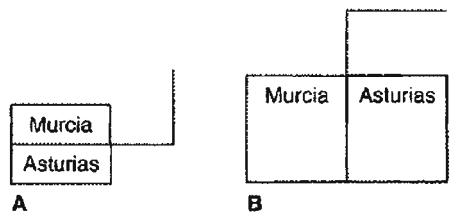



**Figura 17.1** Ejemplo de dendrograma horizontal (A) y vertical (B).

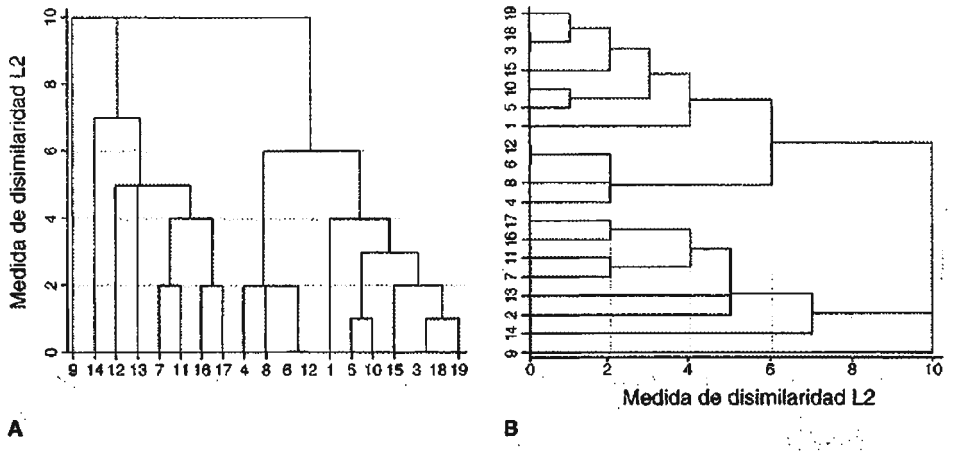



**Figura 17.2** Dendrograma vertical (A) y horizontal (B) para el ejemplo del producto interior bruto.

Además, STATA permite opciones como incluir etiquetas en las gráficas (subinstrucción **labels** ()), modificar la orientación del dendrograma (subinstrucciones **vertical** u **horizontal**) o mostrar solo algunos datos (el subcomando **cutnumber** () permite limitar el número de ramas de la gráfica y el subcomando **cutvalue** () hace posible limitar los valores de distancia representados).

En el ejemplo anterior:

```
cluster dendrogram riqueza
```

o en su forma reducida:

```
cluster dend riqueza
```

## 17.5. ESTANDARIZACIÓN Y TRANSFORMACIÓN DE VARIABLES

Cuando el clúster se forma con dos o más variables, las de mayor magnitud pueden dominar a las otras. Por ejemplo, si se utilizan como variables el PIB en millones de euros *(PIB)*, el número de habitantes *(hab)* y la tasa de mortalidad infantil *(morti)*, el PIB tendrá más influencia. Para resolver este problema, antes de comenzar los cálculos debe procederse a la estandarización de todas las variables. La estandarización de variables es una transformación matemática que conduce a la creación de una nueva variable estandarizada con una media aritmética de 0 y una desviación típica de 1. La fórmula que se aplica para estandarizar una variable es la siguiente:

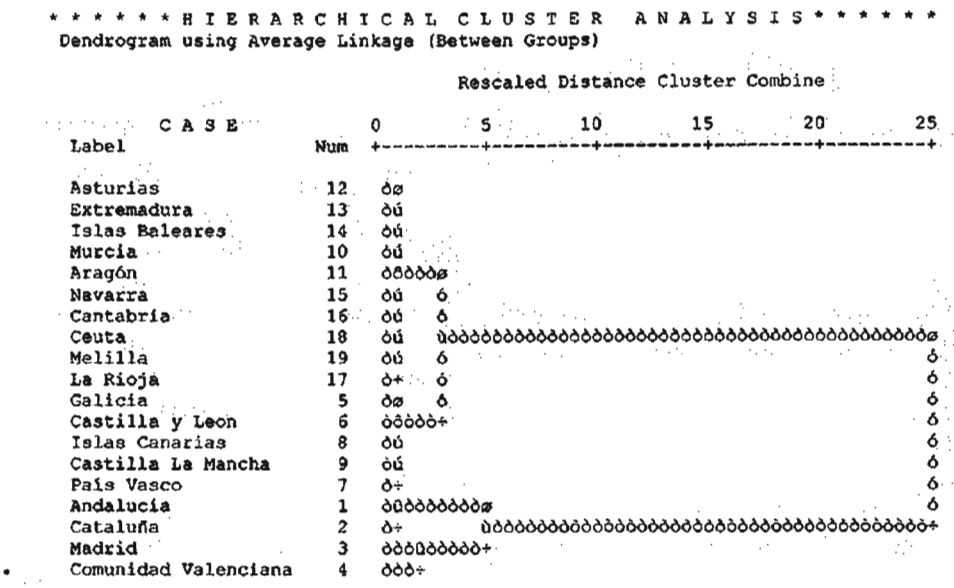
$$\text{Variable estandarizada} = \frac{\text{variable} - \text{media}}{\text{desviación típica}}$$

Así, a cada valor de la variable hay que restarle un valor de la media de la variable y dividir esta diferencia por la desviación típica o estándar de la misma.

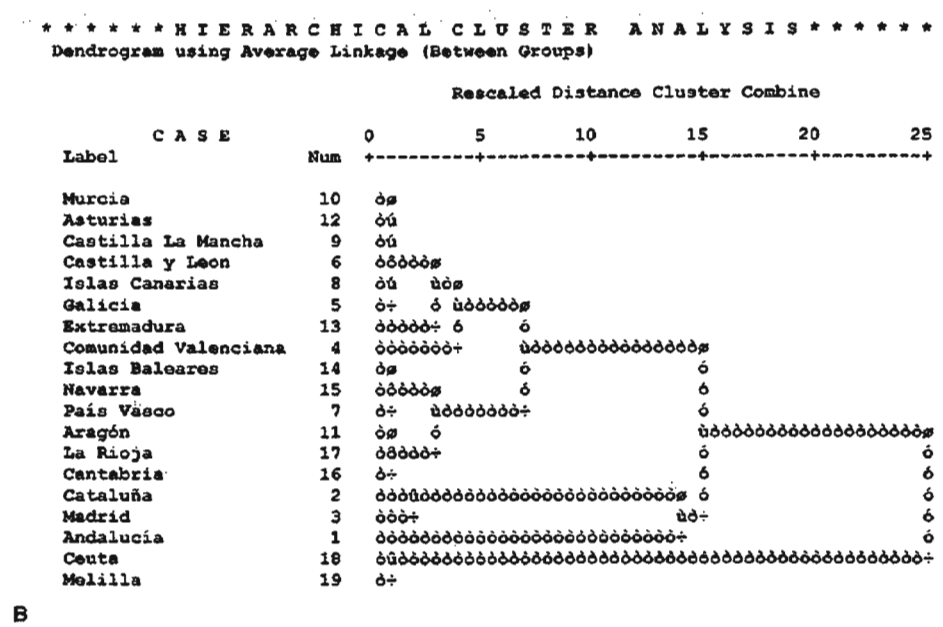
STATA permite la estandarización de variables a través de la instrucción **egen**. La instrucción se escribe como:

```
egen stdPIB=std (PIB)

egen stdhab=std (hab)

egen stdmorti=std (morti)
```

Esta estandarización es necesaria para que todas las variables tengan el mismo impacto, con independencia de sus unidades de medida. Para ilustrar esta recomendación, se presentan seguidamente dos dendrogramas de un clúster jerárquico (fig. 17.3). El primero clasifica las comunidades autónomas según PIB, número de habitantes y densidad de población sin estandarizar; en el segundo se han estandarizado las variables.




A




B

**Figura 17.3** Dendrogramas con variables sin estandarizar (A) y con variables estandarizadas (B).

Como puede observarse, el resultado del segundo dendrograma parece más lógico que el del primero, ya que se utilizaron las variables estandarizas en unas escalas que son comparables. Intuitivamente se tendería a esperar que Cataluña y Madrid pudieran formar un clúster, y que Ceuta y Melilla integraran otro, por sus características semejantes de renta *per capita*, densidad de población y número de habitantes.

## 17.6. REQUISITOS PARA LA APLICACIÓN DE LOS MÉTODOS DE ANÁLISIS DE CLÚSTER

1. Las observaciones deben ser independientes entre sí.
2. Las variables utilizadas para crear clústeres han de ser cuantitativas o dicotómicas. Si una de las variables utilizadas fuese categórica, no se podría usar el clúster jerárquico ni el de K medias, y habría que recurrir al de conglomerados en dos fases.
3. Se deben asumir los mismos supuestos que para la correlación, la regresión y el análisis factorial. Sin embargo, la técnica del análisis de conglomerados es muy robusta, por lo que la violación de algún requisito no suele ser importante, especialmente si el tamaño de muestra es grande.
4. El clúster de K medias asume una muestra grande (más de 200 casos).
5. El clúster de K medias es muy sensible a los valores extremos. Como práctica habitual, antes de realizar un clúster de K medias se eliminan los valores extremos o *«outliers»*.

## 17.7. CLÚSTERES DE VARIABLES

El análisis de clústeres también se puede usar para clasificar las variables y agruparlas en conglomerados con homogeneidad intraclúster, de modo similar a lo que se ha hecho anteriormente para clasificar los sujetos. Por ejemplo, sería útil cuando se dispone de una serie de medidas del electroencefalograma de diversas áreas cerebrales y se desea saber cuáles se asemejan más entre sí (**3,4**), o bien si se recogieron escalas de calidad de vida y se pretende valorar cuáles están más próximas entre sí. Aunque para este último ejemplo parecería apropiado un análisis factorial, podrían surgir problemas si el número de casos es reducido, porque los intervalos de confianza o las pruebas de significación difícilmente alcanzarían significación estadística. En este caso, la técnica de clúster es preferible, porque no se realizan pruebas de significación. Desde STATA, estos análisis se llevan a cabo desde la opción:

```
clustermat
```

## 17.8. EJEMPLO DE ANÁLISIS DE CLÚSTER CON STATA

### 17.8.1. A través de instrucciones

Supóngase una base de datos de 40 sujetos en la que se recoge su frecuencia en el consumo de alcohol *(alcohol)*, tabaco *(smoke)* y cannabis *(cannabis)*.

```
. summ alcohol smoke cannabis
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| alcohol | 40 | 4.5 | 2.917674 | 0 | 10 |
| smoke | 40 | 3.725 | 3.558936 | 0 | 10 |
| cannabis | 40 | 2.275 | 2.995616 | 0 | 10 |

En este ejemplo, todas las variables fueron recogidas con las mismas unidades (frecuencia semanal de consumo), por lo que no será necesario proceder a su estandarización previa. Se realizará

un análisis de clúster jerárquico *(simple linkage)* para determinar cómo se agrupan los 40 sujetos[2]. A este análisis de clúster se decide denominarlo «habitos».

```
cluster s alcohol smoke cannabis, name(habitos)
```

El análisis de conglomerados genera tres variables al dar la instrucción. Al haber elegido el nombre «habitos», las tres variables comenzarán con esta palabra.

*habitos_id* (id variable).

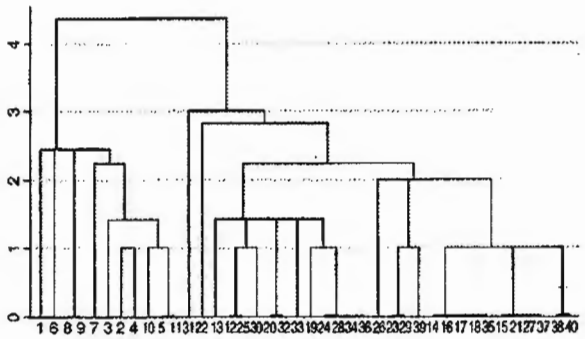
*habitos_ord* (order variable).

*habitos_height* (height variable).

Estas variables, que guardan relación con la manera en que el ordenador ha realizado el análisis de clúster, no suelen tener una utilidad práctica directa.

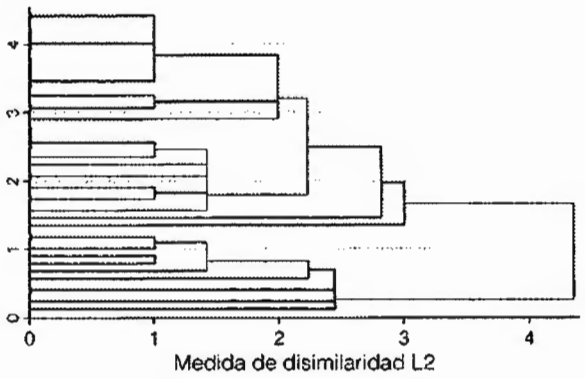
A continuación, puede obtenerse un dendrograma vertical:

```
cluster dend habitos
```




O uno horizontal:

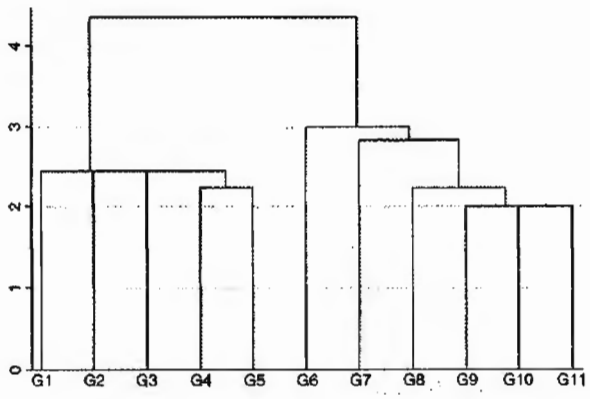
```
cluster dend habitos, horizontal
```




2 Obsérvese que no es necesario escribir la subinstrucción **simplelinkage** completa; basta con emplear su primera letra.

También es posible modificar los ejes para una visualización gráfica más clara. Por ejemplo, se elegirá que solo se presenten las ramas del dendrograma formadas a partir de una distancia de 1,5.

```
cluster dend habitos, cutvalue (1.5)
```




Existen otras posibilidades de análisis de STATA que no han sido mencionadas a lo largo del capítulo; son las instrucciones **generate** y **stop.**

Imagínese que se desea crear la variable *conductas2,* que clasificará a los 40 sujetos en dos clústeres o grupos. STATA usará la siguiente instrucción (5):

```
cluster gen conductas2=group(2)
```

Podría haberse creado no una, sino dos variables (una con dos clústeres y otra con tres). En este caso, la instrucción sería:

```
cluster gen conductas=group(2/3)
```

Se ha seleccionado la subinstrucción **group(*2/3*)** para que genere dos variables conductas: *conductas2* (con dos categorías) y *conductas3* (con tres categorías). La distribución de los sujetos en ambas variables sería la siguiente:

```
. table conductas2 conductas3
```

| conductas2 | conductas3 | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| 1 | 11 | | |
| 2 | | 1 | 28 |

Parece que agrupar los datos en dos conglomerados (variable *conductas2*), uno con 11 sujetos y el otro con 29, resulta más lógico que agruparlos en tres conglomerados (variable *conductas3*) con 11, 1 y 28 sujetos, respectivamente.

Aunque en este ejemplo la elección de un análisis con dos conglomerados se intuye fácilmente, STATA posee una instrucción que establece el número de clústeres idóneo siguiendo dos reglas: el índice de Calinski y Harabasz (índice seudo-F) y el índice de Duda-Hart (índice Je(2)/Je(1)).

Por defecto, STATA calcula el índice de Calinski y Harabasz para diferente número de clústeres. La opción que posea un índice mayor será la idónea. La instrucción se escribirá como:

```
cluster stop
```

```
. cluster stop
```

| Number of clusters | Calinski/ Harabasz pseudo-F |
|---|---|
| 2 | 113.54 |
| 3 | 60.72 |
| 4 | 43.59 |
| 6 | 29.63 |
| 9 | 58.95 |
| 11 | 67.79 |
| 12 | 62.57 |

Obsérvese que el número de conglomerados idónco es 2.

Si se desease calcular otro índice, debería especificarse con la subinstrucción:

```
cluster stop, rule(duda)
```

Este mismo análisis puede realizarse en el método de K medias, a través de la instrucción:

```
cluster k alcohol smoke cannabis, k(2) n(conducta)
```

(Se ha decidido denominar *conducta* a la variable creada según el análisis de K medias, para diferenciarla de la variable *conductas2* obtenida con el análisis jerárquico.)

En este caso, STATA crea directamente la variable *conducta* sin necesidad de aplicar la instrucción **generate**. La distribución de esta variable según este método es:

```
. table conducta
```

| conducta | Freq. |
|---|---|
| 1 | 29 |
| 2 | 11 |

Si se comparan los resultados obtenidos con el modelo jerárquico (variable *conductas2*) con el modelo de las K medias (variable *conducta*), los resultados coinciden.

```
. table conducta conductas2
```

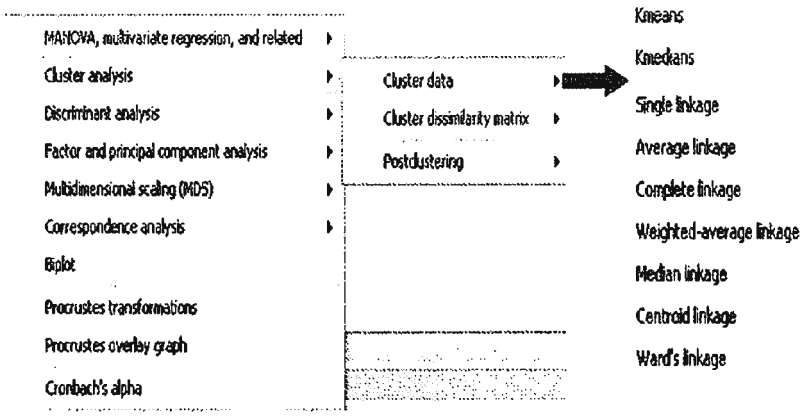
| conducta | conductas2 1 | 2 |
|---|---|---|
| 1 | | 29 |
| 2 | 11 | |

### 17.8.2. A través del menú

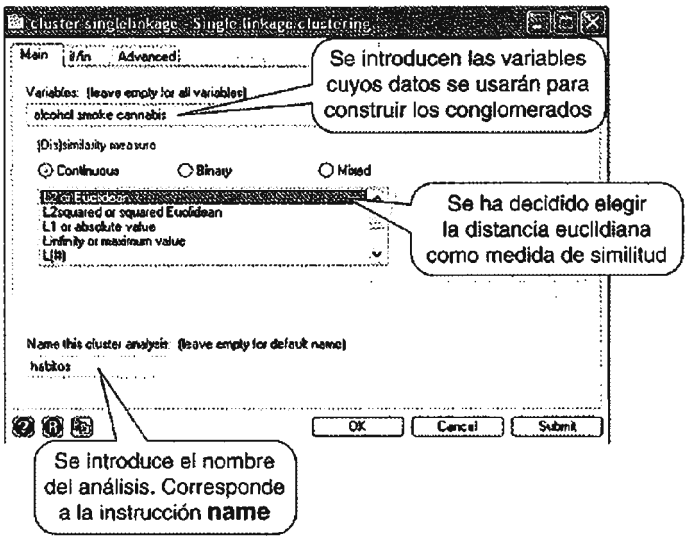
Desde la opción:

**Statistics → Multivariate analysis → Cluster analysis → Cluster data**

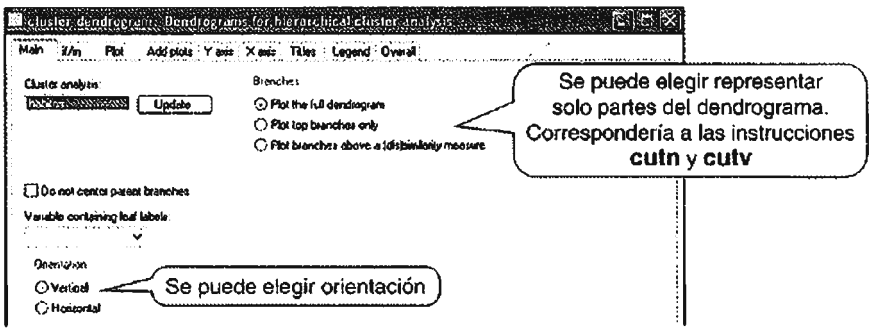
se abre un menú despegable en STATA que permite elegir las opciones de análisis jerárquico o de análisis de K medias/K medianas.




En el ejemplo anterior, se ha elegido el método de la distancia mínima *(single linkage)*, con la distancia euclidiana como medida de similitud, y se realiza un análisis de clúster llamado «habitos» a partir de las variables *alcohol*, *smoke* y *cannabis*.



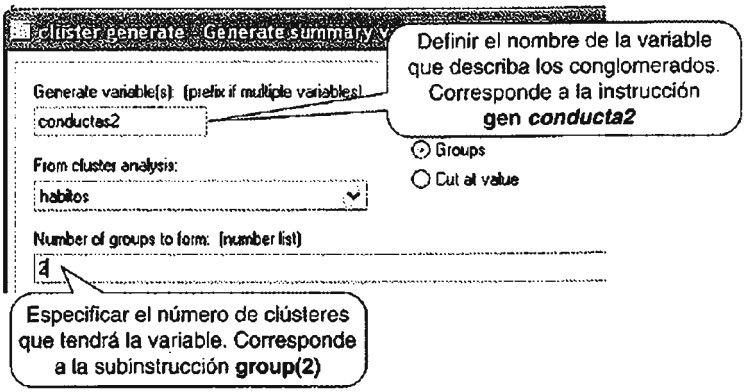

Si se desea representar un dendrograma de este análisis, se elegirá la siguiente secuencia desde el menú:
**Statistics → Multivariate analysis → Cluster analysis-Postclustering → Dendrograms**



Desde el menú:

**Statistics → Multivariate analysis → Cluster analysis → Postclustering → Summary variables from cluster analysis**

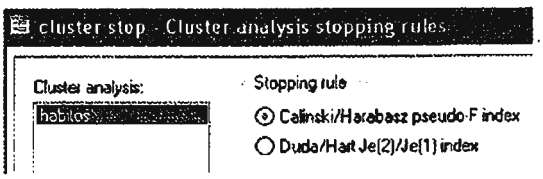
pueden crearse variables que describan los clústeres.



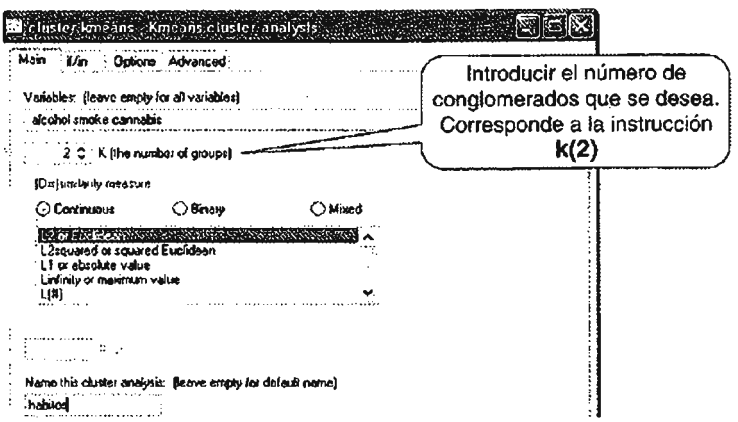

Finalmente, con la instrucción:

**Statistics → Multivariate analysis → Cluster analysis → Postclustering → Cluster analysis stopping rules**

se definirá el número idóneo de clústeres, para construir aplicando diferentes reglas.



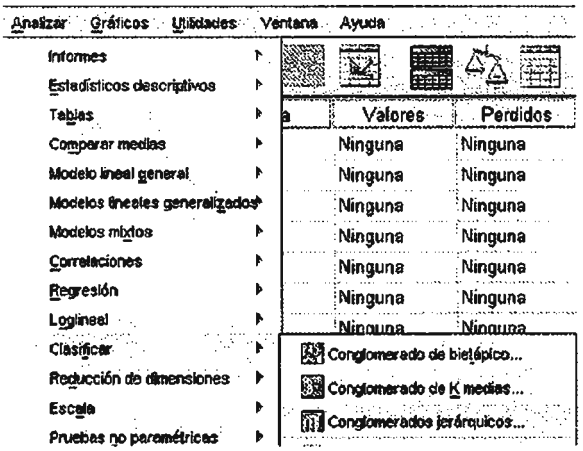

La realización del análisis con el método de K medias sería muy similar. El único cambio en el cuadro de diálogo presentado con esta opción consistiría en incluir el número de conglomerados que se desean (en el ejemplo, dos).



## 17.9. ANÁLISIS DE CLÚSTER CON SPSS

SPSS permite realizar un análisis de conglomerados a través de la opción:

**Analizar → Clasificar**

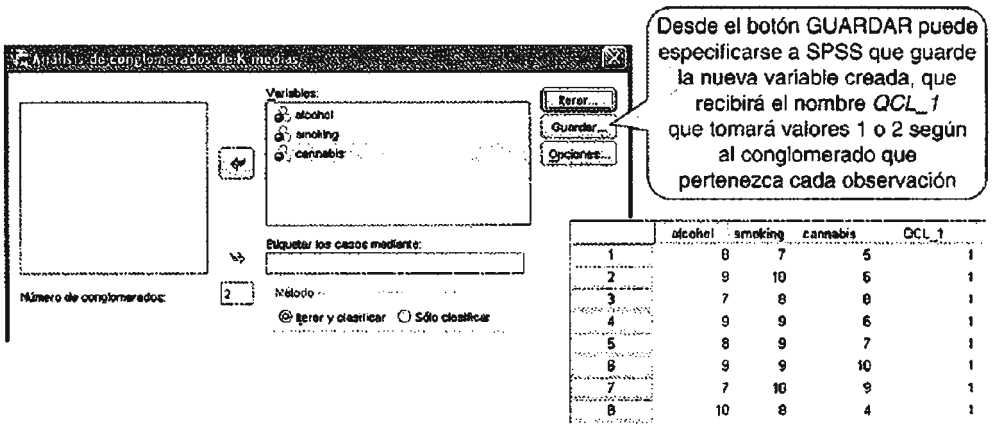



SPSS lleva a cabo tres tipos posibles de análisis de conglomerados. Además de los análisis del clúster de K medias y del clúster jerárquico, es capaz de efectuar un análisis de clúster bietápico, que está indicado para agrupar observaciones procedentes tanto de variables de tipo cuantitativo como cualitativo. La descripción de este procedimiento desborda los objetivos del presente libro.

Con respecto al resto de análisis de conglomerados, la forma de operar es similar a la descrita para el programa STATA. Se describirán los procedimientos de forma breve, aplicando el ejemplo desarrollado con STATA.

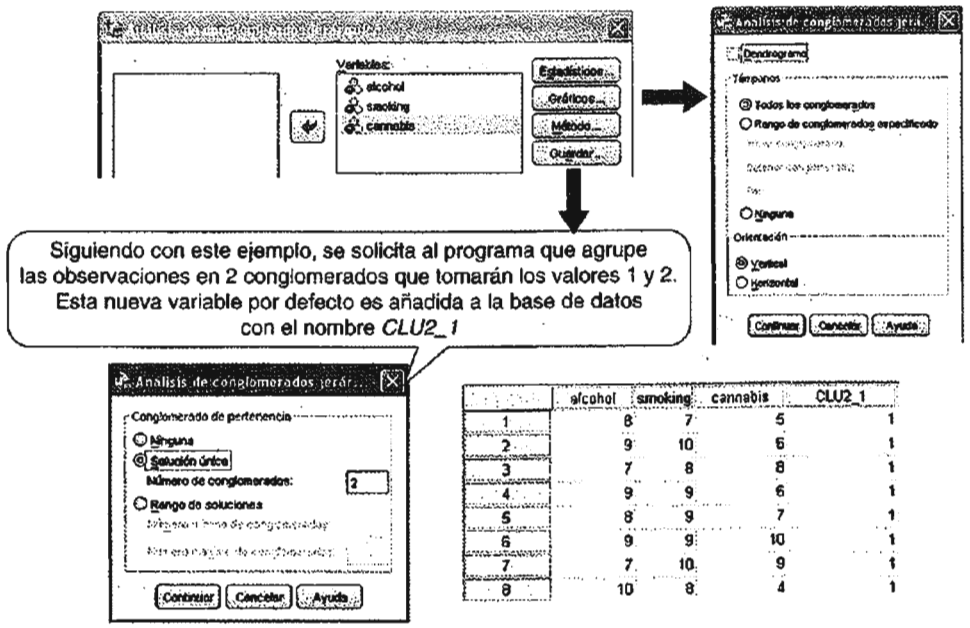
### 17.9.1. Conglomerado de K medias

Una vez seleccionadas las variables de las que se quieren obtener los conglomerados, se decide el número de clústeres o conglomerados para calcular. En este ejemplo, dos:



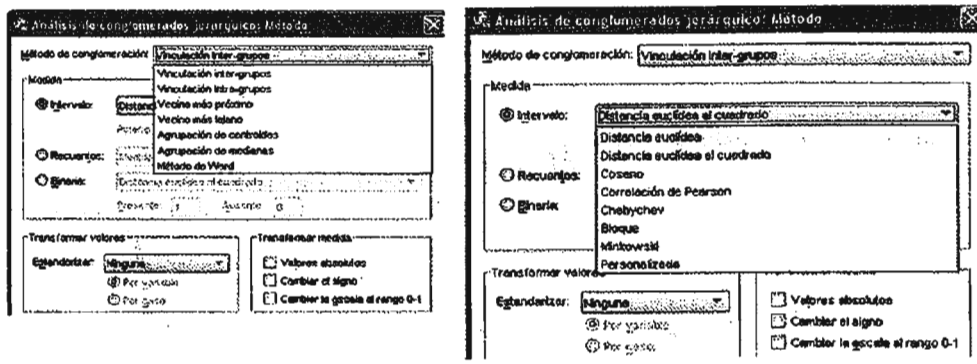
### 17.9.2. Conglomerados jerárquicos

Tras la selección de las variables de las que se derivarán los conglomerados, estos pueden representarse a través de un dendrograma (botón GRÁFICOS) o guardarse con el botón GUARDAR.




| | alcohol | smoking | cannabis | CLU2_1 |
|---|---|---|---|---|
| 1 | 8 | 7 | 5 | 1 |
| 2 | 9 | 10 | 6 | 1 |
| 3 | 7 | 8 | 8 | 1 |
| 4 | 9 | 9 | 6 | 1 |
| 5 | 8 | 9 | 7 | 1 |
| 6 | 9 | 9 | 10 | 1 |
| 7 | 7 | 10 | 9 | 1 |
| 8 | 10 | 8 | 4 | 1 |

SPSS hace posible realizar el análisis jerárquico siguiendo diferentes aproximaciones, al igual que STATA. Para ello, debe acudirse al botón MÉTODO. Se permite elegir el método de unión de conglomerados (método de conglomeración) y la medida de similitud o disimilaridad.




Por defecto, SPSS utiliza la vinculación intergrupos como método de conglomeración (**average linkage** en STATA) y la distancia euclidiana al cuadrado (**L2 squared** en STATA) como medida de similitud. Igualmente, permite estandarizar las variables antes de realizar el análisis de clúster, si fuera necesario.

## 17.10. RESUMEN DE LAS INSTRUCCIONES EN STATA Y SPSS

| Procedimiento | STATA | SPSS |
|---|---|---|
| **Clúster de K medias** | **cluster kmeans** $v_1 v_2 v_p$**, k (#)** | **QUICK CLUSTER** $v_1 v_2 v_p$ **/CRITERIA=CLUSTER(#) /METHOD=KMEANS** |
| Guardar conglomerados | p.d. | **/SAVE CLUSTER** |
| **Clúster jerárquico** | **cluster** | **CLUSTER** $v_1 v_2 v_p$ |
| Medida de similitud: | | |
| • Distancia euclídea o euclidiana | p.d. | **/MEASURE=EUCLID** |
| • Distancia euclídea al cuadrado | **measure (L2squared)** | **/MEASURE=SEUCLID** |
| Método de unión: | | |
| • Mínima distancia | **cluster singlelinkage** $v_1 v_2 v_p$ | **/METHOD SINGLE** |
| • Máxima distancia | **cluster completelinkage** $v_1 v_2 v_p$ | **/METHOD COMPLETE** |
| • Distancia media | **cluster averagelinkage** $v_1 v_2 v_p$ | **/METHOD BAVERAGE** |
| Dendrograma vertical | **cluster dendrogram** | **/PLOT DENDROGRAM VICICLE** |
| Dendrograma horizontal | **cluster dendrogram, horizontal** | **/PLOT DENDROGRAM HICICLE** |
| Guardar conglomerados | **cluster generate** ***nombre de variable*****=group (#)** | **/SAVE CLUSTER (#)** |
| Ejemplo del capítulo: creación de dos clústeres para la variable *conducta2* con el método de la mínima distancia; obtención de un dendrograma vertical | **cluster s** ***alcohol smoke cannabis*** **cluster dend cluster gen** ***conducta2*****=group(*****2*****)** | **CLUSTER** ***alcohol smoke cannabis*** **/METHOD SINGLE /MEASURE=EUCLID /PLOT DENDROGRAM VICICLE /SAVE CLUSTER(*****2*****). EXECUTE. VARIABLE LABELS CLU2_1 '*****conducta2*****'. EXECUTE.** |

#, n.° de conglomerados; p.d., por defecto (el programa calcula el parámetro sin añadir instrucciones); *v*, variables.

## REFERENCIAS


1. Aldenderfer MS, Blashfield RK. Cluster analysis. Thousand Oaks: Sage Publications, Quantitative Applications in the Social Sciences Series; 1984. p. 44.
2. Corter JE. Three models of similarity and association. Thousand Oaks: Sage Publications, Quantitative Applications in the Social Sciences Series; 1996. p. 112.
3. Everitt B, Landau S, Leese M. Cluster analysis. 4th ed. London: Edward Arnold Publishers; 2001.
4. Kachigan SK. Multivariate statistical analysis. New York: Radius Press; 1982.
5. Kaufman L, Rousseeuw PJ. Finding groups in data: An introduction to cluster analysis. New York: John Wiley & Sons; 1990.

# MÉTODOS ESTADÍSTICOS EN METAANÁLISIS 18

*M. Á. Martínez-González, P. A. de la Rosa, A. Gea*

## 18.1. REVISIONES SISTEMÁTICAS Y METAANÁLISIS

Revisar una hipótesis de investigación supone recoger y sintetizar críticamente la investigación original realizada hasta la fecha sobre ese tema. Se persigue hacer una síntesis del *estado de conocimientos (state of the art)*. Esta síntesis, cuando se publica como artículo en una revista científica, se denomina *artículo de revisión* o simplemente *revisión*. Las *revisiones* pueden ser *narrativas* o *sistemáticas*. Las revisiones sistemáticas son cada vez más utilizadas y están reemplazando a las clásicas revisiones narrativas, quizá porque las revisiones meramente narrativas carecían de criterios metodológicos específicos y podían acabar por seleccionar y combinar la investigación previa según el capricho del autor. Resulta paradójico que se revise la evidencia *científica* aplicando procedimientos que *no son científicos*. La revisión *sistemática* sí aplica el método científico y exige establecer unos criterios de búsqueda, selección y combinación de la investigación previa que estén bien definidos y sean absolutamente *reproducibles* por otros autores. Cuando la revisión sistemática incorpora, además, un análisis estadístico para combinar cuantitativamente los resultados de varios estudios independientes, entonces se denomina metaanálisis (**1-6**).

## 18.2. TAREAS PREVIAS AL ANÁLISIS ESTADÍSTICO

La cuidadosa ejecución de cada una de las tareas previas (cuadro 18.1) es más importante que el propio análisis estadístico (**1,2,5-7**). Deben realizarse de modo meticuloso.

## 18.3. ESCALA ADITIVA O MULTIPLICATIVA

El primer paso del metaanálisis es muy parecido a calcular una media ponderada. Sin embargo, antes de ponerse a realizar este cálculo es preciso distinguir dos situaciones. Por una parte, la medida del efecto que se combinará mediante el metaanálisis puede ser de tipo *aditivo* (medias, proporciones, diferencias de medias o diferencias de proporciones) o *multiplicativo* (*odds ratios*, riesgos relativos, razones de riesgos, razones de tasas, *hazard ratios*). En el segundo caso, es necesario trabajar con los logaritmos de la medida del efecto.

## 18.4. EFECTOS ESTANDARIZADOS: *D* DE COHEN

La *d* de Cohen es la diferencia entre dos medias dividida entre la desviación estándar común.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

En $s_p$, el subíndice p significa *ponderada* y equivale a la desviación estándar común para los dos grupos, que, como se recordará, es la que se calcula para una *t* de Student con varianzas homogéneas (v. apartado 6.1.3).

Por ejemplo, si la media de reducción de colesterol LDL en sujetos tratados con estatinas fue 20 mg/dl, la del grupo de control fue 2 mg/dl y la desviación estándar común ($s_p$) de la variable cambio era 36 mg/dl, la *d* de Cohen valdría (20 – 2)/36 = 0,5. Se interpreta diciendo que existen

**CUADRO 18.1 TAREAS PREVIAS A LA REALIZACIÓN DE UN METAANÁLISIS**

1. Formular la pregunta de investigación
2. Definir los criterios de elegibilidad de los estudios
   a. Tipo de participantes
   b. Intervenciones o exposiciones que se van a comparar
   c. Desenlaces o resultados *(outcomes, end-points)*
   d. Diseño del estudio
   e. Requisitos de calidad metodológica
3. Localización de estudios y estrategia de búsqueda: fuentes
   a. Bases de datos electrónicas (PubMed, EMBASE, otras)
   b. Registro de ensayos de la colaboración Cochrane
   c. Bases de datos electrónicas no cubiertas por la Cochrane
   d. Revisión de las referencias bibliográficas de cada artículo
   e. Búsqueda manual en revistas claves para la materia, en libros de congresos y en la literatura gris
   f. Contactar con expertos en la materia y pedirles, a su vez, nombres de más expertos *(snowball sampling)*
4. Selección de estudios
   a. Comprobar la elegibilidad por dos o más observadores
   b. Establecer un algoritmo para resolver desacuerdos
   c. Mantener un listado de estudios excluidos y las razones de su exclusión
5. Valoración de la calidad de los estudios seleccionados
   a. Considerar si pueden valorarla independientemente dos o más evaluadores
   b. Usar preferentemente una lista-guía de requisitos que se exigen y no una escala que cuantifique la calidad
   c. En los ensayos, evaluar siempre el enmascaramiento de la asignación y del desenlace, y el manejo de las pérdidas durante el seguimiento
   d. En los estudios observacionales, valorar siempre el control de la confusión y los sesgos de selección
   e. Considerar si se enmascara para los evaluadores el nombre de los autores y sus instituciones, y de las revistas
6. Extracción de datos para el metaanálisis
   a. Considerar si dos o más observadores realizarán independientemente la extracción de los datos
   b. Diseñar y pilotar el formulario para la recogida de datos de cada estudio
   c. Considerar si se enmascara para los observadores el nombre de los autores y sus instituciones, y de las revistas

0,5 desviaciones estándar de diferencia entre las medias de ambos grupos. La *d* de Cohen puede ser positiva o negativa. En el ejemplo sería negativa y habría que añadirle el signo menos, ya que se trata de *reducciones* de colesterol LDL.

Estas medidas estandarizadas se han usado a menudo en metaanálisis, especialmente en el terreno de la psicología. Se trata de obtener de cada estudio la diferencia estandarizada (es decir, la *d* de Cohen) y después obtener una media ponderada de todas ellas, como se verá más adelante.

## 18.5. MÉTODO DEL INVERSO DE LA VARIANZA: EFECTOS FIJOS

Este método se explicará con un ejemplo imaginario de cada posible situación.

### 18.5.1. Combinar la estimación de proporciones

El siguiente ejemplo plantea un caso poco frecuente. Lo más común es combinar medidas *relativas* (*odds ratios*, riesgos relativos, *hazard ratios*) basadas en un cociente, como se explica en el apartado 18.5.3.

Imagínese que se desean combinar tres estudios que valoran la proporción de pruebas de detección de cáncer de próstata que resultaron positivas (tabla 18.1). Estas proporciones fueron

*Tabla 18.1 Método del inverso de la varianza (efectos fijos) para una sola proporción*

| ESTUDIO | P | LIC 95% | LSC 95% | EE | W = 1/EE² | W × P | PESOS (%) |
|---|---|---|---|---|---|---|---|
| A | 0,09 | 0,06 | 0,12 | 0,015306 | 4.268,44 | 384,16 | 8,16 |
| B | 0,12 | 0,10 | 0,14 | 0,010204 | 9.604 | 1.052,48 | 18,37 |
| C | 0,1 | 0,09 | 0,11 | 0,005102 | 38.416 | 3.841,6 | 73,47 |
| | | | | Sumas= | 52.288,44 | 5.378,24 | |
| | | | | | P. comb. | **0,103** | |
| | | | | | EEc | 0,00437 | |
| | | | | | LIC P. comb. | **0,094** | |
| | | | | | LSC P. comb. | **0,111** | |

del 9, el 12 y el 10% para los estudios A, B y C, respectivamente, con los siguientes intervalos de confianza al 95%:

- Estudio A: **0,09** (IC 95%: 0,06-0,12).
- Estudio B: **0,12** (IC 95%: 0,10-0,14).
- Estudio C: **0,10** (IC 95%: 0,09-0,11).

Lo primero que se requiere es extraer el error estándar (EE) de cada proporción.

$$EE = \frac{LSC\text{-}LIC}{2 \times z_{\alpha/2}}$$

Así, el EE en el estudio A será (0,12-0,06)/(2 × 1,96) = 0,015306, etc. (v. tabla 18.1).

Una vez obtenido el error estándar, se calcula una media ponderada de las tres proporciones, usando como peso $(w_i)$ para cada estudio el inverso de su error estándar al cuadrado:

$$\text{Proporción}_{\text{combinada}} = \frac{\sum w_i p_i}{\sum w_i} = \frac{\sum\left(\frac{1}{EE_i^2} \times p_i\right)}{\sum\left(\frac{1}{EE_i^2}\right)}$$

Los cálculos en este ejemplo serían:

$$\text{Proporción}_{\text{combinada}} = \frac{\frac{0,09}{0,015306^2} + \frac{0,12}{0,010204^2} + \frac{0,1}{0,005102^2}}{\frac{1}{0,015306^2} + \frac{1}{0,010204^2} + \frac{1}{0,005102^2}} = 0,103$$

Para esta proporción combinada (0,103) que resume la información de las 3 investigaciones originales, se debe calcular un intervalo de confianza. Para eso se necesita un error estándar combinado ($EE_c$), cuya fórmula es:

$$EE_c = \sqrt{\frac{1}{\sum w_i}}$$

En este ejemplo, el $EE_c$ sería:

$$EE_c = \sqrt{\frac{1}{1/0,015306^2 + 1/0,010204^2 + 1/0,005102^2}} = 0,00437$$

Una vez que se dispone de una proporción combinada ($P_c$) y de su respectivo error estándar ($EE_c$), se puede calcular un intervalo de confianza al 95% para la proporción que resume los 3 estudios, mediante el procedimiento habitual para calcular intervalos de confianza a las proporciones:

$$IC(\pi_c) = P_c \pm z_{\alpha/2}\, EE_c$$

$$IC\,95\%\ \text{proporción poblacional}\,(\pi_c) = 0{,}103 \pm 1{,}96(0{,}00437) = 0{,}094\ \text{a}\ 0{,}111$$

### 18.5.2. Diferencias de proporciones, estimaciones de medias o diferencias de medias

Se darán exactamente los mismos pasos que para las proporciones:

- Extraer el error estándar (EE) de cada estudio.
- Obtener la media ponderada por el inverso del cuadrado del EE: $w_i = 1/EE_i^2$.
- Calcular el error estándar combinado: $EE_c = (1/\Sigma w_i)^{0.5}$.
- Estimar el intervalo de confianza para la media ponderada.

La tabla 18.2 recoge un supuesto metaanálisis de 3 estudios que comparan un tratamiento frente a placebo y encuentran las siguientes reducciones (IC 95%) en el riesgo de complicaciones (riesgo con tratamiento-riesgo con placebo):

- Estudio A: 0% (IC 95%: –3% a +3%).
- Estudio B: –2% (IC 95%: –3% a –1%).
- Estudio C: –4% (IC 95%: –8% a 0%).

El estudio B sería el único que encontraría un efecto beneficioso significativo del tratamiento frente al placebo. El A correspondería a un efecto nulo. El estudio C estaría justo en el límite de la significación estadística, pues el límite superior del intervalo de confianza al 95% coincide exactamente con el 0. Al integrar estos tres estudios en un metaanálisis, se obtendría la siguiente diferencia de proporciones (DP) combinada:

$$DP_{combinada} = -0{,}019\,(IC\,95\%: -0{,}028\,\text{a}\,-0{,}010)$$

La tabla 18.3 recoge otro metaanálisis simulado de cuatro estudios ficticios que comparaban la calidad de vida (*Quality of life* o QoL, escala de 0 a 100) entre fumadores y no fumadores:

- Estudio A: $QoL_{No\ fumadores} - QoL_{Fumadores} = 5$ (IC 95%: 2 a 8).
- Estudio B: $QoL_{No\ fumadores} - QoL_{Fumadores} = 7$ (IC 95%: 2 a 12).
- Estudio C: $QoL_{No\ fumadores} - QoL_{Fumadores} = 8$ (IC 95%: 6 a 10).
- Estudio D: $QoL_{No\ fumadores} - QoL_{Fumadores} = 4$ (IC 95%: –6 a 14).

***Tabla 18.2** Método del inverso de la varianza (efectos fijos): diferencia de proporciones*

| ESTUDIO | $P_1$-$P_2$ | LIC 95% | LSC 95% | EE | $W = 1/EE^2$ | $W\,(P_1-P_2)$ | PESOS (%) |
|---|---|---|---|---|---|---|---|
| A | **0** | **–0,03** | **0,03** | 0,015306 | 4.268,44 | 0 | 9,47 |
| B | **–0,02** | **–0,03** | **–0,01** | 0,005102 | 38.416 | –768,32 | 85,21 |
| C | **–0,04** | **–0,08** | **0** | 0,020408 | 2.401 | –96,04 | 5,33 |
| | | | | Sumas= | 45.085,44 | –864,36 | |
| | | | | | Dif. comb. | **–0,019** | |
| | | | | | EEc | 0,00471 | |
| | | | | | LIC 95% | **–0,028** | |
| | | | | | LSC 95% | **–0,010** | |

*Tabla 18.3 Método del inverso de la varianza (efectos fijos): diferencia de medias*

| ESTUDIO | $X_1$-$X_2$ | LIC 95% | LSC 95% | EE | W | W × DIF | PESOS (%) |
|---|---|---|---|---|---|---|---|
| A | 5 | 2 | 8 | 1,53061 | 0,43 | 2,13 | 27,03 |
| B | 7 | 2 | 12 | 2,55102 | 0,15 | 1,08 | 9,73 |
| C | 8 | 6 | 10 | 1,02041 | 0,96 | 7,68 | 60,81 |
| D | 4 | −6 | 14 | 5,10204 | 0,04 | 0,15 | 2,43 |
| | | | | Sumas= | 1,58 | 11,05 | |
| | | | | | Dif. comb. | **6,995** | |
| | | | | | EEc | 0,79573 | |
| | | | | | LIC 95% | **5,435** | |
| | | | | | LSC 95% | **8,554** | |

La estimación ponderada de la diferencia de medias en QoL será de 6,995 (IC 95%: 5,435 a 8,554) puntos a favor de los no fumadores.

### 18.5.3. Combinación de medidas relativas (*odds ratios*, razones de riesgos, *hazard ratios*)

Se trata ahora de la situación más frecuente, en la que se desea combinar medidas relativas (*odds ratios*, riesgos relativos, *hazard ratios*). Requiere un paso previo, que consiste en transformar la medida relativa *(ratio)* en su logaritmo, al que en lo sucesivo llamaremos *b*. Se usará un ejemplo con *odds ratios* (que pueden ser intercambiables con riesgos relativos o *hazard ratios*).

$$b = \ln(\text{OR})$$

También se calculará el error estándar[1] teniendo en cuenta los logaritmos:

$$\text{EE}_b = \frac{\ln\left(\frac{\text{LSC}}{\text{LIC}}\right)}{2 \times z_{\alpha/2}}$$

Al final, se revertirá la transformación del modo siguiente:

$$\text{IC}95\%\,\text{OR}_{\text{combinada}} = \text{EXP}(b_{\text{combinada}}) \times \text{EXP}(\pm z_{\alpha/2}\,\text{EE}_c)$$

En lo demás, se procede del modo indicado anteriormente. Se verá más claro con un ejemplo.

La tabla 18.4 recoge un metaanálisis muy sencillo (imaginario) que incluye 3 estudios que comparaban la mortalidad entre usuarios de cannabis y no usuarios:

- Estudio A: *odds ratio* = 1,60 (IC 95%: 0,80-3,20).
- Estudio B: *odds ratio* = 1,50 (IC 95%: 0,50-4,50).
- Estudio C: *odds ratio* = 2,00 (IC 95%: 1,25-3,20).

1 Algunos estudios presentan los resultados, de forma menos práctica e informativa, como estimación puntual y valor *p* en lugar del intervalo de confianza. Por ejemplo, de modo que la diferencia de medias de colesterol entre la intervención y el placebo sea 4 mg/dl y $p_{2\text{colas}} = 0{,}037$. En esta situación podría extraerse así el error estándar:

Si $P_{2\text{colas}} = 0{,}037$, entonces $z = 2{,}086$

Esto se obtiene con STATA mediante **display invnormal**(*0.0372*) que devuelve −2,086.

Despejamos entonces el error, una vez que conocemos z y el efecto encontrado:

$$z = \frac{\text{efecto}}{\textit{error}};\ \text{error} = \frac{4}{2{,}086} = 1{,}918$$

*Tabla 18.4 Método de efectos fijos con una medida relativa (odds ratio [OR])*

| EST. | OR | LIC 95% | LSC 95% | B | EE | W | W × B | PESOS (%) |
|---|---|---|---|---|---|---|---|---|
| A | 1,6 | 0,80 | 3,20 | 0,470 | 0,35365 | 8,00 | 3,76 | 27,99 |
| B | 1,5 | 0,50 | 4,50 | 0,405 | 0,56052 | 3,18 | 1,29 | 11,14 |
| C | 2 | 1,25 | 3,20 | 0,693 | 0,23980 | 17,39 | 12,05 | 60,87 |
| | | | | | Sumas = | 28,57 | 17,10 | |
| | | | | | | b comb. | 0,60 | |
| | | | | | | OR comb. | **1,820** | |
| | | | | | | EEc | 0,18709 | |
| | | | | | | LIC 95% | **1,261** | |
| | | | | | | LSC 95% | **2,626** | |

Los valores de b para cada estudio serán:

- $b_A = \ln(1,6) = 0,470$.
- $b_B = \ln(1,5) = 0,405$.
- $b_C = \ln(2) = 0,693$.

Teniendo en cuenta que $2 \times 1,96 = 3,92$, los errores estándar de b serán:

- $EE_{bA} = \ln(3,2/0,8)/3,92 = 0,354$.
- $EE_{bB} = \ln(4,5/0,5)/3,92 = 0,561$.
- $EE_{bC} = \ln(3,2/1,25)/3,92 = 0,240$.

Los pesos ($w_i$) serán:

- $w_A = 1/0,354^2 = 8$.
- $w_B = 1/0,561^2 = 3,18$.
- $w_C = 1/0,240^2 = 17,39$.

Así se calcula la media ponderada de b:

$$b_{\text{combinada}} = \frac{(8 \times 0,47) + (3,18 \times 0,405) + (17,39 \times 0,693)}{(8 + 3,18 + 17,39)} = 0,60$$

Por lo tanto, la OR combinada valdrá: $e^{0,60} = 1,82$.
El error estándar de la estimación combinada será:

$$EE_{\text{combinado}} = \sqrt{\frac{1}{(8 + 3,18 + 17,36)}} = 0,187$$

Finalmente, el intervalo de confianza de la estimación global combinada se obtendrá como:

$$IC95\%\,OR_{\text{combinada}} = 1,82 \times EXP(\pm z_{\alpha/2}\, EE_c) = 1,82 \times EXP(\pm 1,96 \times 0,187) =$$
$$= 1,26 \text{ a } 2,63$$

## 18.6. GRÁFICOS DE BOSQUE *(FOREST PLOT)*

Se ha hecho habitual representar los metaanálisis mediante un gráfico que muestra los efectos encontrados en múltiples estudios que intentan contestar a una misma pregunta de investigación. Este gráfico representa cada intervalo de confianza como una línea horizontal con la estimación puntual como un cuadrado central. El gráfico de bosque, o *forest plot*, es el modo habitual de

presentar un metaanálisis, y esta diseñado sobre todo para mostrar los riesgos relativos u *odds ratios* (o la medida de efecto de que se trate) de cada estudio.

Aunque los *forest plots* pueden adoptar diversas configuraciones, normalmente se presentan como una lista situada a la izquierda con los nombres (muchas veces en orden cronológico) de cada uno de los estudios recogidos, seguida de unas figuras para cada estudio que incluyen unas barras laterales (intervalo de confianza) en torno a un pequeño cuadrilátero o círculo central (estimación puntual del riesgo relativo o la medida de efecto en cuestión). Este gráfico está situado a la derecha del nombre de cada estudio y puede representarse en escala logarítmica cuando se usen *odds ratios* u otras medidas multiplicativas, porque solo así los intervalos de confianza serán simétricos en torno a la estimación puntual (RR u OR de cada estudio). En caso contrario, podría darse indebidamente una importancia excesiva a la parte del intervalo de confianza que es superior a 1, y despreciar también injustificadamente la inferior a 1 (ya que todo su margen de posibles valores quedaría tan solo entre 0 y 1).

El área de cada cuadrilátero suele ser proporcional al peso que tiene el estudio. Se representan dos líneas verticales, una correspondiente al nulo (RR = 1) y otra a la estimación combinada ($OR_p$, por ejemplo) que resulta del metaanálisis. Estas líneas cruzan todos los estudios. Al final, en la parte inferior se representa como un rombo la estimación combinada global del metaanálisis. El rombo comprende todo el intervalo de confianza de la $OR_p$.

La figura 18.1 recoge un ejemplo de *forest plot*. Contiene hallazgos de estudios prospectivos observacionales publicados hasta agosto de 2013 que valoraron el riesgo relativo de enfermedad cardiovascular asociado a mejorar en dos puntos la adherencia a una escala de dieta mediterránea que va desde 0 (pésima conformidad) a 9 (ideal) (**8**).

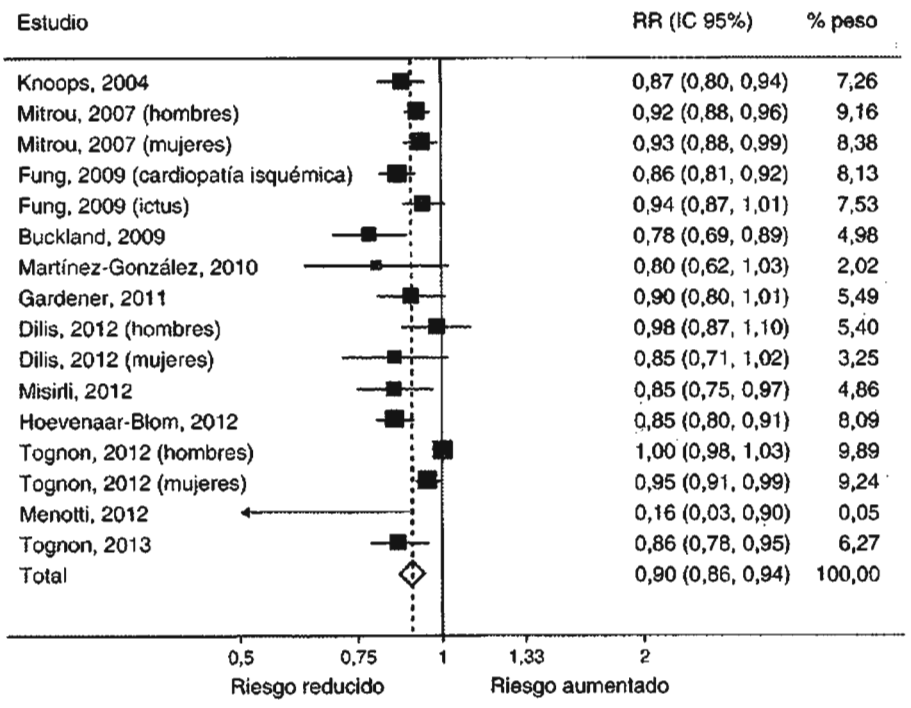



**Figura 18.1** *Forest plot* (gráfico de bosque).

## 18.7. TEST DE HETEROGENEIDAD: ESTADÍSTICO Q

Intentar combinar estudios realizados en diferentes lugares y tiempos, sobre distintas poblaciones y con diseños y criterios diferentes, lleva a encontrarse casi siempre con problemas. El primero es que los resultados de los diversos estudios pueden ser estadísticamente diferentes entre sí, y producir una *heterogeneidad* que no puede ignorarse al hacer un metaanálisis (**7,9**). La detección de esta inconsistencia entre los resultados de los diversos estudios rebajaría la confianza que se puede depositar en la aplicación del tratamiento valorado. Por tanto, siempre debe realizarse un análisis de la heterogeneidad, no solo para detectarla, sino también para intentar explicar las razones de la misma, que a menudo se convierte en la finalidad más importante de un metaanálisis (**6**). Como contraste de hipótesis de la heterogeneidad (hipótesis nula: homogeneidad) se usa el estadístico Q, que sigue una ji cuadrado con $k - 1$ grados de libertad, siendo $k$ el número de estudios (**9**).

$$Q = \sum w_i \left(b - b_{combinado}\right)^2$$

Si el estadístico Q es grande y tiene un valor $p$ significativo o próximo a la significación, se rechazará la homogeneidad de los estudios (y la capacidad de combinarlos se pone en tela de juicio). *Se suele exigir que $p > 0,10$ (y no $p > 0,05$) como falta de evidencia de heterogeneidad.* Por ejemplo, si $p = 0,09$, se *pensará* que *sí* existe heterogeneidad. Esto se debe a que el test de heterogeneidad tiene poca potencia. No ser capaz de rechazar la hipótesis de homogeneidad no implica que no exista heterogeneidad. En el análisis de la heterogeneidad, se debe partir de una reflexión sobre cuáles pueden ser las variables clave *(fuentes de heterogeneidad)* que influyen diferencialmente para explicar, por ejemplo, que unos estudios encuentren una asociación directa y otros una asociación inversa: variables de diseño, exposición, efecto, otros factores de riesgo, variables de persona, tiempo y lugar, etc.

- En el ejemplo de la tabla 18.5 se recogen cinco estudios con las siguientes OR:

0,50 1,00 1,50 3,00 y 2,00

Los respectivos valores de $b$ ($b = \ln(OR)$) serán:

−0,69 0,00 0,41 1,10 y 0,69

***Tabla 18.5 Cálculo del estadístico Q de heterogeneidad***

| EST. | OR | LIC 95% | LSC 95% | B | EE | W | W × B | $W_i(B - B_{comb})^2$ |
|---|---|---|---|---|---|---|---|---|
| A | 0,5 | 0,25 | 1,00 | −0,693 | 0,35 | 8,00 | −5,54 | $8(-0,693 - 0,28)^2 = 7,6$ |
| B | 1 | 0,50 | 2,00 | 0,000 | 0,35 | 8,00 | 0,00 | $8(0,000 - 0,28)^2 = 0,6$ |
| C | 1,5 | 0,50 | 4,50 | 0,405 | 0,56 | 3,18 | 1,29 | $3,18(0,405 - 0,28)^2 = 0,05$ |
| D | 3 | 1,00 | 9,00 | 1,099 | 0,56 | 3,18 | 3,50 | $3,18(1,1 - 0,28)^2 = 2,1$ |
| E | 2 | 1,25 | 3,20 | 0,693 | 0,24 | 17,39 | 12,05 | $17,39(0,693 - 0,28)^2 = 3,0$ |
| | | | | | Sumas = | 39,75 | 11,30 | Q = 13,35 |
| | | | | | | b comb.= | 0,28 | $p = 0,0097$ |
| | | | | | | OR comb.= | 1,329 | |
| | | | | | | EEc= | 0,159 | |
| | | | | | | LIC 95%= | 0,974 | |
| | | | | | | LSC 95%= | 1,813 | |

El metaanálisis obtiene un valor de $b_{combinado} = 0,28$ y, por tanto, una $OR_{combinada} = 1,33$. A partir de estos datos y de los pesos ($w_i$) mostrados en la tabla 18.5 ($w_i = 8,00; 8,00; 3,18; 3,18$ y $17,39$), se puede calcular Q:

$$Q = 8(-0,69-0,28)^2 + 8(0-0,28)^2 + 3,18(0,41-0,28)^2 + 3,18(1,1-0,28)^2 + 17,39(0,69-0,28)^2 = 13,35$$

El estadístico Q = 13,35 resultaría significativo según una ji cuadrado con cuatro grados de libertad ($p = 0,01$), lo que lleva a concluir que se están incluyendo estudios con resultados discordantes entre sí. Esto hace problemático el combinarlos. Se deberían separar en *subgrupos* según características metodológicas o de la población estudiada que permitan averiguar el porqué de las disparidades.

## 18.8. TAU CUADRADO: VARIANZA ENTRE ESTUDIOS

Los estudios resultan más heterogéneos al aumentar la variabilidad entre sus resultados. Esta variabilidad interestudios se estima con una *varianza entre estudios* que se llama *tau cuadrado* ($\tau^2$) y se calcula como:

$$\tau^2 = \frac{Q-k+1}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}} \; (\text{si } \tau^2 < 0, \text{se asume } \tau^2 = 0)$$

A los valores negativos de tau cuadrado se les asigna un 0.

En la tabla 18.6 se recoge el cálculo de tau cuadrado para el ejemplo de la *odds ratio:*

$$\tau^2 = \frac{13,35-5+1}{39,75-\frac{450,5}{39,75}} = 0,329$$

Tau cuadrado es comparable entre los distintos subgrupos de estudios que se puedan separar dentro de un mismo metaanálisis, pero no es comparable entre uno y otro metaanálisis hechos sobre temas diversos.

## 18.9. ÍNDICE I CUADRADO

Al inconveniente de tau cuadrado antes mencionado se suma el hecho de que su valor no tiene una interpretación intuitiva ni otra interpretación cuantitativa directa. Solamente se puede afirmar que un aumento de tau cuadrado refleja mayor heterogeneidad. Una alternativa interesante es el estadístico *I cuadrado* ($I^2$), que cuantifica el grado de heterogeneidad en una misma escala intuitiva y comparable para cualquier metaanálisis (**10**). $I^2$ proporciona una medida del grado de inconsistencia en los resultados de los diferentes estudios incluidos en un metaanálisis. Describe el porcentaje de la variabilidad total entre estudios que es debida a heterogeneidad. Se calcula del modo siguiente:

$$I^2 = \frac{Q-k+1}{Q}$$

En el ejemplo:

$$I^2 = \frac{13,35-5+1}{13,35} = 0,7004$$

Se expresará en tanto por ciento (habitualmente suele bastar con un único decimal): $I^2 = 70,0\%$. Así, puede entenderse que $I^2$ es una medida del grado de heterogeneidad que se mueve en una escala continua que va del 0 al 100% y que es fácilmente interpretable. *Grosso modo*, y por tener

*Tabla 18.6 Modelo de efectos aleatorios para el mismo ejemplo de la tabla 18.5*

| EST. | OR | LIC 95% | LSC 95% | B | EE | W | W × B | Q | $W_i$^2 | $W_i'$ | $WI' \times B$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | **0,5** | **0,25** | 1,00 | −0,693 | 0,35 | 8,00 | −5,54 | 7,638 | 64 | 2,20 | −1,53 |
| B | **1** | **0,50** | 2,00 | 0,000 | 0,35 | 8,00 | 0,00 | 0,646 | 64 | 2,20 | 0,00 |
| C | **1,5** | **0,50** | 4,50 | 0,405 | 0,56 | 3,18 | 1,29 | 0,047 | 10 | 1,55 | 0,63 |
| D | **3** | **1,00** | 9,00 | 1,099 | 0,56 | 3,18 | 3,50 | 2,111 | 10 | 1,55 | 1,71 |
| E | **2** | **1,25** | 3,20 | 0,693 | 0,24 | 17,39 | 12,05 | 2,907 | 302 | 2,59 | 1,79 |
| | | | | | Sumas = | 39,75 | 11,30 | **13,35** | 450,5 | 10,10 | 2,60 |
| | | | | | b (fixed) = | 0,28 | P(heterog.) = | **0,01** | **b random =** | **0,26** | = 2,6/10,1 |
| | | | | | OR (fixed) = | **1,329** | tau^2 = | **0,329** | **OR random =** | **1,294** | |
| | | | | | EE (fixed) 0 | 0,159 | I^2 = | 70,0 | **EE random =** | **0,315** | = 1 = 10,1^0,5 |
| | | | | | LIC 95% = | **0,974** | | | **LIC 95% =** | **0,699** | |
| | | | | | LSC = 95% | **1,813** | | | **LSC 95% =** | **2,398** | |

*A modo de ejemplo, los pesos aleatorios (w') son $1/(EE^2 + \tau^2)$, es decir, 1/(0,352 + 0,329) = 2,20.*

cierta referencia, se puede hablar de ausencia de heterogeneidad si $I^2$ = 0%, baja heterogeneidad ($I^2$ = 25%), heterogeneidad moderada (50%) o heterogeneidad alta (75%) (**10**). En el ejemplo utilizado en la tabla 18.6, la heterogeneidad sería moderada-alta y, además, reuniría claramente los requisitos para considerarla estadísticamente significativa (Q = 13,35, con cuatro grados de libertad, $p$ = 0,01).

## 18.10. GRÁFICO DE L'ABBÉ PARA HETEROGENEIDAD

Se trata de un procedimiento visual para valorar la heterogeneidad (**11**). Es aplicable, sobre todo, al metaanálisis de ensayos clínicos que utilizan una variable dicotómica (ocurrencia o no de un suceso clínico) como desenlace. Se representa la tasa de sucesos clínicos (eventos o *end-points*) en el grupo sometido a tratamiento activo frente a la tasa del grupo control, como se muestra en la figura 18.2.

Cada círculo representa un estudio. El tamaño de los círculos es proporcional al del estudio. En la figura se aprecia que hay un estudio que explica especialmente la heterogeneidad, ya que está situado en la esquina superior izquierda correspondiente a una tasa muy alta de eventos en los sometidos a tratamiento activo, mientras que la tasa de eventos era muy baja en el grupo placebo. En los otros 8 estudios sucede lo contrario: aunque varía la tasa en el grupo placebo, siempre es mayor en esos 8 estudios con el placebo que con el tratamiento. La diagonal corresponde a igualdad de tasas entre tratamiento y control (efecto nulo). Este gráfico se entiende fácilmente de manera intuitiva y permite identificar los estudios responsables de la heterogeneidad, pero requiere indicar en el ordenador los datos de las cuatro casillas de la tabla 2 × 2 de cada estudio. Cuando se trata de estudios que no son aleatorizados, esta identificación no es siempre posible, ya que puede existir confusión y requerirse ajustes multivariables. Se debe elegir siempre el estimador del efecto que esté mejor ajustado.

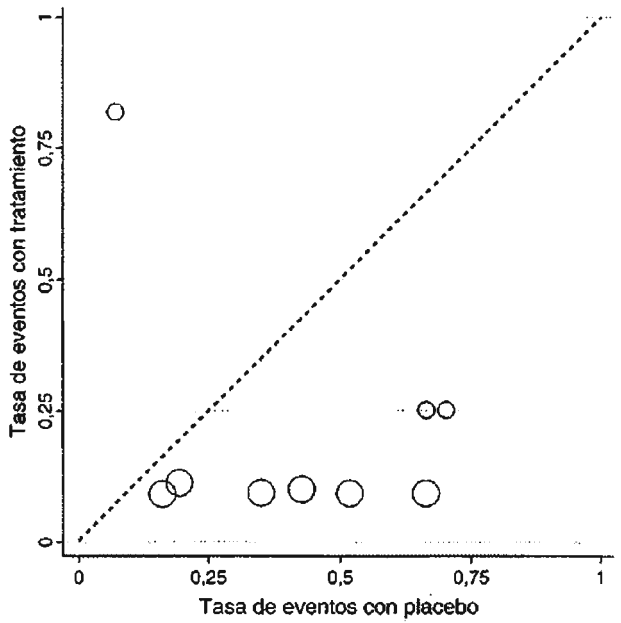



**Figura 18.2** Gráfico de L'Abbé.

## 18.11. METAANÁLISIS DE EFECTOS ALEATORIOS: MÉTODO DE DERSIMONIAN-LAIRD

La heterogeneidad echa a perder la posibilidad de combinar todos los efectos en una estimación única o global. ¿Qué alternativas existen cuando se detecta heterogeneidad? No hay un procedimiento estadístico que *ajuste* por heterogeneidad. Lo que se debe hacer en esta situación es tratar de *identificar las fuentes de heterogeneidad.* Podría suceder, por ejemplo, que los resultados de ensayos financiados por la industria farmacéutica (que pueden tender a dar un mensaje excesivamente optimista) sean distintos que los que cuentan con financiación independiente (**12**). Del mismo modo, otras características de los estudios pueden permitir localizar la explicación de por qué se producen resultados diferentes (**6,7,13**). Esto es lo más interesante.

En presencia de heterogeneidad, se recomienda preferir otro modelo para el análisis, que se llama de efectos aleatorios *(random-effects model),* también conocido como método de DerSimonian-Laird (**14**). Difiere del método de efectos fijos *(fixed-effects model),* que se ha visto hasta ahora en este capítulo.

Aun así, el modelo de efectos aleatorios no *arregla* de ningún modo los problemas de heterogeneidad si estos son graves. Solo consigue que se tenga en cuenta de alguna manera la posible heterogeneidad al añadir la varianza entre estudios ($\tau^2$) al denominador de los pesos. Se llama *de efectos aleatorios* porque este modelo asume que los resultados combinados en el metaanálisis no comprenden la totalidad de la evidencia existente, sino que son una muestra *aleatoria* de todos los posibles resultados de un número muy superior de estudios que no se conocen en su totalidad. El modelo de efectos fijos, en cambio, supone que existe un único efecto en la población y está contemplado en los estudios que se van a combinar. Esta diferencia entre efectos fijos y aleatorios es teórica. Lo práctico es que los pesos se calculan de modo diferente.

En la ponderación no solo se considera el propio error estándar de cada estudio (variabilidad intraestudio), sino también la variabilidad que pueda existir entre los estudios (variabilidad entre estudios o *tau cuadrado*). Los nuevos pesos ($w'$) que se usarán para el modelo de efectos aleatorios son:

$$w_i' = \frac{1}{EE_i^2 + \tau^2}$$

La tabla 18.6 recoge los cálculos al aplicar un modelo de efectos aleatorios al ejemplo presentado en la tabla anterior. El cuadro 18.2 resume las principales ventajas de cada modelo (**5,7**). Cuando el número de estudios es pequeño, se tiende a preferir el modelo de efectos aleatorios. Este modelo estaría indicado cuando no se ha podido resolver la causa de la heterogeneidad (aunque se insiste en que no es una *solución* frente a ella). Debe saberse que un inconveniente del modelo aleatorio es que tiende a asignar demasiado peso relativo a los estudios pequeños.

### CUADRO 18.2 VENTAJAS DEL MODELO DE EFECTOS FIJOS Y EL DE EFECTOS ALEATORIOS

**Modelo de efectos fijos**

- Otorga mucho más peso a los estudios de mayor tamaño, como parece lógico
- Es más sencillo y directo
- No requiere asumir nada sobre representatividad de los estudios incluidos

**Modelo de efectos aleatorios**

- Amplía los intervalos de confianza y, así, previene la falsa imagen de alta precisión que puede transmitir un metaanálisis
- Incorpora la heterogeneidad debida a la variabilidad interestudios
- Asume que solo se posee una muestra aleatoria de los estudios, lo cual es más realista que suponer que se posee toda la información

En las tablas 18.1 a 18.4 se ha añadido una última columna a la derecha que recoge los pesos relativos que se otorgaron a cada estudio. Corresponde a dividir el peso de ese estudio entre la suma total de pesos, que supondría el 100%. Al comparar en la tabla 18.6 los pesos dados en el primer y el último estudio con modelo de efectos fijos *(w)* y con modelo aleatorio (*w'*), se puede comprobar que, con el modelo fijo, el último estudio pesaba más del doble (17,39) que el primero (8,00); sin embargo, esta diferencia casi se anula al usar *w'*. El motivo es que ahora todos los pesos se ven afectados por *tau cuadrado*, que es *constante* para todos los estudios.

## 18.12. ANÁLISIS DE SUBGRUPOS

Separar los estudios en varios subgrupos en función de su diseño, metodología, fecha de realización, fuente de financiación, edad de los participantes, niveles de los factores de riesgo o enfermedades concomitantes, entre otros, puede ayudar a resolver la heterogeneidad, porque se pueden encontrar subgrupos dentro de los cuales los resultados sean homogéneos (6). Así ha sucedido al estratificar en ensayos, estudios de cohortes, y estudios de casos y controles. También se ha conseguido resolver la heterogeneidad cuando se separaron estudios que sólo valoraban como efecto los casos mortales de enfermedad cardiovascular de los que contemplaban eventos cardiovasculares no letales, ya que los primeros dependen no solo de los factores de riesgo, sino también de la calidad de la atención médica. Separar los estudios según su fuente de financiación permite un análisis mucho más sutil del efecto y la historia que puede estar detrás de ciertas disparidades en los resultados. Esta finalidad *analítica* del metaanálisis suele ser más elegante, inteligente e interesante que la búsqueda sintética a toda costa de un estimador único que resuma la totalidad de la evidencia disponible en un solo número (flanqueado por sus límites de confianza). Puede ser desafortunado dirigirse primariamente a tal finalidad sintética, ya que, a menudo, los estudios no se pueden combinar y se preferirá una aproximación mucho más analítica para buscar las variables clave que crean subgrupos que ya son homogéneos entre sí y podrían explicar la heterogeneidad global entre estudios (6). Esto requiere repetir el metaanálisis dentro de cada subgrupo de estudios.

## 18.13. METARREGRESIÓN

La metarregresión es una técnica estadística destinada a valorar las fuentes de heterogeneidad en un metaanálisis. Equivale a un modelo de regresión en el que, como variable dependiente, se utiliza la magnitud del efecto (diferencia de medias, de proporciones o $b = \ln(RR)$) y, como predictores o variables independientes, se introducen una o varias variables que podrían explicar las diferencias entre estudios. Las técnicas de metarregresión requieren trabajar con ponderaciones. No debe olvidarse que se manejan datos agregados y no individuales (la unidad de observación es el estudio) y, por tanto, la falacia ecológica (7,15) puede afectar a este procedimiento.

## 18.14. SESGO DE PUBLICACIÓN: GRÁFICO DE EMBUDO *(FUNNEL PLOT)*

Los metaanálisis suelen basarse solo en estudios publicados. Incluso cuando se intentan sacar a la luz los estudios no publicados, los publicados tendrán más probabilidad de ser localizados. A pesar de estos esfuerzos, siempre habrá estudios que se realizaron y finalmente no se publicaron, o que nunca son localizados. Los estudios publicados pueden diferir *sistemáticamente* de los no publicados, lo cual creará un sesgo *(sesgo de publicación)*. La principal amenaza para la validez de un metaanálisis suele provenir, precisamente, de este *sesgo de publicación* (5,7,16). La trascendencia de este sesgo es relevante para investigadores y lectores: si lo que aparece no representa la realidad, se está distorsionando la transmisión del conocimiento científico. Es importante detectar este sesgo y conocer las razones que lo favorecen.

La selección sesgada de los artículos que acaban publicándose no solo depende de los revisores *(peer-reviewers)* y editores de las revistas; también recae en autores que deciden no enviar sus investigaciones a publicar si no les gustan o no les emocionan sus resultados. Por ejemplo, hay investigadores que hacen muchos estudios, pero solo envían para su publicación aquellos que tienen resultados significativos porque intuyen que, sus artículos serán aceptados (o al menos lo serán con más rapidez) si comunican resultados significativos. Esta creencia, que desgraciadamente responde a la realidad (**16**), crea un sesgo enorme, porque lo que se publica es solo una parte sesgada (y muy tendenciosa) de la verdad. Todavía peor es el caso de quienes desarrollan estudios financiados por alguna industria y piensan que los resultados que encuentran no favorecerán a sus patrocinadores, con lo que acaban por no ver la luz (**12**). Otros investigadores encuentran resultados contrarios al paradigma vigente, y ellos mismos opinan que no van a ser creíbles y no los envían a publicar, aunque hayan aplicado métodos rigurosos.

En cambio, los grandes ensayos aleatorizados y los multicéntricos se publican prácticamente siempre, sean cuales sean sus resultados. Los estudios de gran tamaño muestral rara vez dejan de publicarse, ya que suponen un mayor esfuerzo en su diseño y ejecución y los autores ponen un mayor interés en que los resultados no permanezcan ignorados, ya sean positivos o negativos.

Así se explica que, al seguir la historia de la investigación desarrollada sobre una hipótesis, a menudo se observe un fenómeno de *regresión a la media* (**7,15**). Los primeros artículos sobre un tema, por su mayor novedad, son más fácilmente aceptados, aunque sean imprecisos, de pequeño tamaño y tiendan a dar resultados más exagerados. En cambio, los que se van publicando después se acercan más al valor nulo.

La posibilidad de que un determinado estudio pase en el futuro a formar parte de un metaanálisis ha apoyado la adopción de varias normas: CONSORT (**17**), que detallan la información que debe aportar todo artículo que presente resultados de un ensayo de intervención; STARD (**18,19**), que establecen los contenidos imprescindibles en artículos sobre validación de pruebas diagnósticas, y STROBE (**20**) para estudios observacionales (transversales, casos y controles, y cohortes). Análogamente, las normas PRISMA (**21**) recogen los criterios que deben tenerse en cuenta para escribir un artículo de metaanálisis o revisión sistemática.

Desde hace una década se ha establecido la obligatoriedad de registrar *por adelantado* todo ensayo clínico que se vaya a realizar. Si no se registró con antelación, no se aceptará después para publicación en prestigiosas revistas. El registro previo (**22**) intenta monitorizar todo ensayo en marcha para identificar los que finalmente se publicarán. Así resulta más difícil que alguien realice ensayos y después oculte sus resultados. En el futuro, esta práctica contribuirá a reducir el sesgo de publicación. Este sesgo puede ser muy grave. No solo hay que detectarlo, sino también cuantificarlo.

El gráfico de embudo *(funnel plot)* es uno de los procedimientos más socorridos para valorar el sesgo de publicación. La figura 18.3 presenta un gráfico de embudo en el que se aprecia un sesgo de publicación porque faltan los estudios de pequeño tamaño de la parte inferior derecha.

Se trata de representar el ln(RR) frente al tamaño de muestra del estudio o frente al inverso del error estándar. Si se coloca el ln(RR) en el eje de abscisas, se obtendrá la imagen de un embudo con la parte más estrecha dirigida hacia arriba (más precisión cuando el tamaño de muestra aumenta o el error estándar disminuye). La existencia de una figura simétrica alrededor de un eje que pasa por el valor medio ponderado del ln(RR) habla en favor de la ausencia de sesgo de publicación. Sin embargo, si la figura no es simétrica y aparece amputado uno de sus dos brazos laterales en la parte más baja, se pensará que es probable que los estudios pequeños que iban en esa dirección se hayan quedado sin publicar por no atreverse sus autores a desafiar el paradigma vigente o porque los editores de las revistas los rechazaron no por sus métodos (que es lo que debería contar), sino porque no se creían sus resultados. Esto contribuiría a que se perpetúe el círculo vicioso y se ahonde

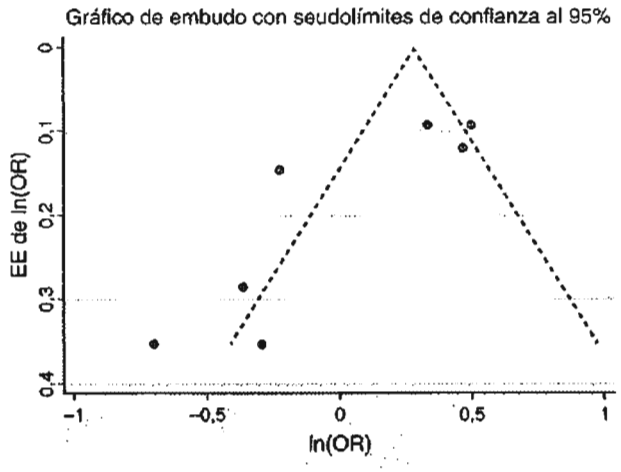



**Figura 18.3** Gráfico de embudo. Se aprecia sesgo de publicación porque faltan estudios de pequeño tamaño con ln(OR) superior a 0,5.

en el sesgo de publicación. Este razonamiento ayuda a entender el motivo por el cual se deben publicar *todos* los estudios finalizados, aunque contengan resultados nulos (no significativos) o contrarios a los esperados, y por qué revisores y editores deben fijarse en la *calidad de los métodos* y no en los resultados para aceptar o no un artículo para su publicación.

## 18.15. SESGO DE PUBLICACIÓN: TEST DE EGGER

El test de Egger es un procedimiento para detectar un sesgo de publicación (**23,24**). Consiste en una regresión lineal simple de la magnitud del efecto, es decir el ln(OR), dividida entre su error estándar, que se usa como variable dependiente, mientras que el inverso del error estándar se usa como variable independiente; la ecuación sería:

$$\frac{\ln(\mathrm{OR})}{\mathrm{EE}} = a + b\frac{1}{\mathrm{EE}}$$

Es decir:

$$y = \ln(\mathrm{OR})/\mathrm{EE}$$
$$x = 1/\mathrm{EE}$$

En lo que hay que fijarse es en la significación estadística de la ordenada en el origen. La ordenada en el origen *(a)* será compatible con 0 cuando el *funnel plot* sea simétrico; en cambio, será significativamente diferente de 0 cuando exista asimetría en el *funnel plot* porque hubo sesgo de publicación.

## 18.16. SESGO DE PUBLICACIÓN: MÉTODOS DE MACASKILL Y DE PETERS

También se puede trazar una recta de regresión entre el tamaño de muestra (variable independiente) y el logaritmo neperiano de la *odds ratio* (variable dependiente), ponderando las observaciones por el inverso de la varianza; es el método de Petra MacAskill (**25**). En ausencia de sesgo, la pendiente de esta recta será cero (línea horizontal). Si la pendiente es estadísticamente significativa (distinta de 0), se considerará que existe sesgo de publicación. En la figura 18.4 se aprecia que faltan estudios

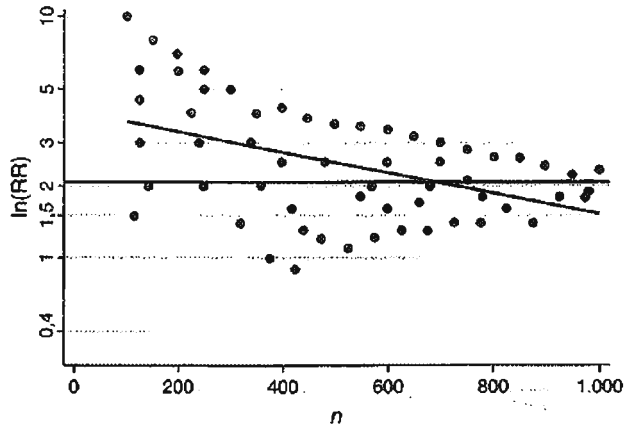



**Figura 18.4** Método de MacAskill. Gráfico en embudo *(funnel plot)* en el que se ajusta una regresión lineal de ln(RR) como variable dependiente sobre *n* como independiente. Se aprecia que la pendiente es diferente de 0 (negativa). Este ejemplo hipotético sugeriría sesgo de publicación.

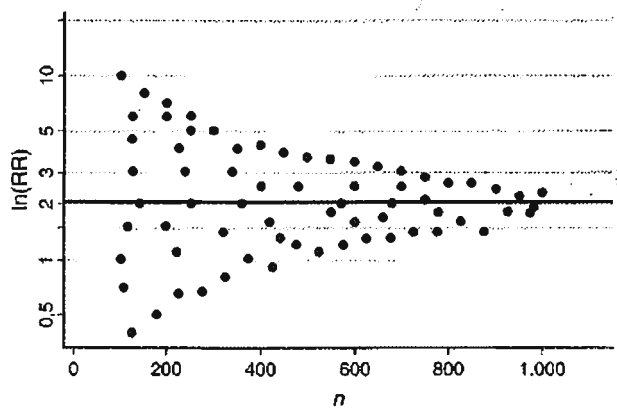



**Figura 18.5** Método de MacAskill. La recta de regresión con el ln(RR) como dependiente y el tamaño de muestra como independiente tiene pendiente 0. Resultados hipotéticos en que no hay sesgo de publicación.

pequeños con OR alrededor de la unidad; si se traza una recta de regresión, la pendiente será negativa, lo que refleja la asimetría del gráfico (en cambio, en la figura 18.5 la pendiente sería 0 y no sugiere un sesgo de publicación). Se ha comprobado que en la regresión es preferible utilizar el inverso del tamaño de muestra, que es el procedimiento de Peters (**7,26**).

## 18.17. SESGO DE PUBLICACIÓN: OTROS MÉTODOS

El método de Begg es similar al de Egger, aunque usa el coeficiente $\tau$ de Kendall en vez de una regresión. Otro método, llamado *trim and fill,* trata de estimar el número de estudios ausentes (existentes, pero no publicados). Intenta rellenar los huecos causantes de la asimetría del *funnel plot* mediante la imputación de los valores simétricos (**5,23**).

## 18.18. METAANÁLISIS ACUMULADO

Se llama metaanálisis *acumulado* a un método consistente en añadir cada vez un estudio más y repetir el metaanálisis con cada nuevo estudio. Así, se actualiza cada vez la estimación del parámetro combinado a medida que se añaden nuevos estudios. El orden en que se incorpora cada nuevo estudio para repetir el metaanálisis suele estar definido por *la fecha de publicación,* aunque también se puede basar en una variable cuantitativa que permita la ordenación de las distintas investigaciones. Los resultados se presentan en forma de gráfico similar al *forest plot,* aunque, en vez de que cada línea corresponda al intervalo de confianza de un solo estudio, se asocia al resumen (estimador combinado de la OR y su intervalo de confianza) *de todo lo publicado hasta esa fecha,* que se actualiza después de cada adición. El metaanálisis se actualiza con la incorporación de cada estudio reciente. Este procedimiento permite valorar la contribución de cada estudio sobre el grado de evidencia disponible hasta ese momento.

## 18.19. USO DE STATA PARA EL METAANÁLISIS

STATA es el *software* ideal para efectuar metaanálisis (27). La orden básica de STATA es **metan,** pero no viene instalada por defecto y es preciso instalarla. La versión actualmente vigente data de 2008. Para los nuevos desarrollos y avances en esta materia, y para aprender a instalar esta opción, se sugiere consultar la siguiente dirección: http://www.stata.com/support/faqs/statistics/meta-analysis/.

El modo más fácil de instalar los programas de metaanálisis en STATA es conectarse a internet, dirigirse a http://www.stata-press.com/data/mais.html y seguir las instrucciones que aparezcan allí. Una vez conectado a internet, también se puede obtener desde STATA tecleando en la ventana *Command* lo siguiente:

```
net install mais
```

Después de instalados los programas de metaanálisis, el ejemplo en escala aditiva de la tabla 18.1 requeriría escribir lo siguiente en un *Do-file* y ejecutarlo:

```
clear
input ///
id str28 Study  p   LIC      LSC
1 "Estudio A"   .09 .06     .12
2 "Estudio B"   .12 .1      .14
3 "Estudio C"   .1  .09     .11
end

g EE=(LSC-LIC)/3.92

metan p EE, lcols (Study) effect("Proportion") ///
textsize (200) astext(60) boxsca(170) boxopt(mcolor(black)) ///
force xlabel(.08, .1, .12, .14, .16)
```

La orden **metan** necesita, al menos, dos argumentos: el efecto y su error estándar. La opción **lcols** indica la variable que ocupará una columna a la izquierda *(left columns),* mientras que **effect** va seguida del nombre (entre paréntesis) que se quiere dar a la medida de efecto. A su vez, la opción **textsize** establece el tamaño de la letra del texto que aparecerá en el gráfico de árbol, **astext** indica el porcentaje de ese gráfico que estará ocupado por texto, y **boxsca** sirve para indicar la escala de los recuadros correspondientes a la estimación puntual de cada estudio. Aquí se ha decidido que los recuadros sean de color negro. La opción **xlabel** fija los rótulos del eje de abscisas. Para realizar metaanálisis de medidas multiplicativas, se debe añadir la opción **eform.** Para pedir un modelo de efectos aleatorios se usará la opción **random.** Ambas deben ir después de la coma.

Por ejemplo, para reproducir el metaanálisis de efectos aleatorios y con *odds ratios* de la tabla 18.6, se usarán las siguientes órdenes:

```
clear
input ///
OR   LIC    LSC
0.5   0.25  1
1     0.5   2
1.5   0.5   4.5
3     1     9
2     1.25  3.2
end
g b=ln(OR)
g EE=ln(LSC/LIC)/3.92
metan b EE, eform random  effect("Odds Ratio") ///
textsize (140) astext(70) boxsca(120) boxopt(mcolor(black)) ///
xlabel(.11, .25, .5, 1, 2, 4, 9)
```

Si se practica con este ejemplo, podrá apreciarse que la escala del *forest plot* resulta simétrica en escala multiplicativa, ya que los rótulos de **xlabel** se han indicado teniendo en cuenta que 1/9 = 0,11; 1/4 = 0,25, etc.

Para mayores detalles, puede consultarse la ayuda de STATA y la compilación realizada por Sterne et al. (27), incluidas las órdenes para valorar la heterogeneidad **(labbe)** y el sesgo de publicación **(metabias, metafunnel)**, que son distintas de **metan.**

## REFERENCIAS


1. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester: John Wiley; 2000.
2. Sterne JAC, Bradburn MJ, Egger M. Meta-analysis in Stata™. En: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. 2nd ed. London: BMJ Publications; 2001. p. 347-69.
3. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis in medicine. New York: Oxford University Press; 1994.
4. Bailar JC. The promise and problems of meta-analysis. N Engl J Med 1997;337:559-61.
5. Delgado-Rodríguez M. Metaanálisis. UD 8. En: Doménech JM, editor. Diseño de estudios sanitarios. Barcelona: Signo; 2001. p. 1-75.
6. Greenland S, O'Rourke K. Meta-analysis. En: Rothman KJ, Greenland S, Lash T, editors. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2008. p. 652-82.
7. Delgado-Rodríguez M, Sillero Arenas M. Revisión sistemática y metaanálisis. En: Martínez-González MA, editor. Conceptos de salud pública y estrategias preventivas: un manual para ciencias de la salud. Barcelona: Elsevier; 2013. p. 55-62.
8. Martínez-González MA, Bes-Rastrollo M. Dietary patterns, Mediterranean diet, and cardiovascular disease. Curr Opin Lipidol 2014;25(1):20-6.
9. Takkouche B, Cadarso-Suarez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. Am J Epidemiol 1999;150:206-15.
10. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557-60.

11. L'Abbe KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med 1987;107:224-33.

12. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008;358:252-60.

13. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med 1997;21(337):536-42.

14. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials 1986;7:177-88.

15. De Irala J, Martínez-González MA, Seguí-Gómez M. Epidemiología aplicada. 2.ª ed. Barcelona: Ariel; 2008.

16. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997;315:640-5.

17. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276(8):637-9.

18. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326(7379):41-4.

19. Ochodo EA, Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. Clin Chem 2013;59(6):917-9.

20. Bastuji-Garin S, Sbidian E, Gaudy-Marqueste C, Ferrat E, Roujeau JC, Richard MA, et al. Impact of STROBE Statement Publication on Quality of Observational Study Reporting: Interrupted Time Series versus Before-After Analysis. PLoS One 2013;8(8):e64733.

21. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reportingitems for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6(7):e1000097.

22. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal. N Engl J Med 2004;351(12):1250-1.

23. Rothstein H, Sutton A, Borenstein M. Publication Bias in Meta-Analysis: Prevention, Assessment Adjustments. Chichester: John Wiley; 2000.

24. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detect by a simple, graphical test. BMJ 1997;315:629-34.

25. MacAskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med 2001;20:641-54.

26. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006;295:676-80.

27. Sterne JAC, Harris RJ, Harbord RM, Steichen TJ. Meta-Analysis in Stata: An Updated Collection from the Stata Journal. College Station: Stata Press; 2009.

# 19 OTROS MÉTODOS BIOESTADÍSTICOS

*M. Á. Martínez-González, P. A. de la Rosa, A. Gea*

## 19.1. MÉTODOS DE REMUESTREO: *BOOTSTRAP, JACKKNIFE*

### 19.1.1. *Bootstrap*

El *bootstrap* es una técnica de remuestreo (tomar muchas submuestras de la muestra que ya se tiene). Se utiliza, principalmente, para valorar la precisión en la estimación de parámetros (cálculo de intervalos de confianza y test de significación estadística). No requiere asumir ninguna distribución teórica de los datos poblacionales, con lo cual es de gran utilidad cuando no exista un método paramétrico o bien cuando no se puedan asumir los requisitos del método paramétrico existente. Cuanto mayor sea la representatividad de la muestra, más fiables serán los resultados obtenidos con este método (1).

El *bootstrap* consiste en tomar repetidas submuestras con reemplazo al azar a partir de la muestra original. Todas las submuestras deben ser de igual tamaño que la muestra original. La característica diferencial del *bootstrap* con respecto a otros métodos de remuestreo es que el remuestreo se realiza *con reemplazo*, de tal forma que en cada nueva muestra —insistimos en que es siempre de igual tamaño muestral que la original— habrá tantos sujetos repetidos (muestreados varias veces) como sujetos que no hayan sido seleccionados. Por ejemplo, si la muestra original tiene cinco sujetos (A, B, C, D y E), podrían obtenerse las siguientes submuestras (subm.) al azar por *bootstrap*:

*original:* ***A B C D E***

*subm. 1:* B B C D D

*subm. 2:* A C C D E

*subm. 3:* A B C E E

*subm. 4:* A A B D E

etc.

Este proceso del remuestreo se repite un número elevado de veces[1] y se calcula el estimador (p. ej., la media) en cada una de estas submuestras. La desviación estándar de la distribución de los estimadores calculados en las distintas submuestras equivaldría al error estándar.

Una vez estimado el error estándar, existen varias aproximaciones para calcular el intervalo de confianza:

- Aproximación a la normal: se asume que los estimadores siguen una normal, con lo cual se utiliza el error estándar obtenido con *bootstrap* en las fórmulas que ya han sido expuestas en capítulos anteriores.
- Método de los percentiles: se obtienen el $P_{2,5}$ y el $P_{97,5}$ de la distribución de los estimadores muestrales obtenida empíricamente por *bootstrap* y se asignan como límites del intervalo de confianza al 95%. Se mantiene así el carácter no paramétrico de esta técnica.

1 Debido a la magnífica capacidad de cálculo de los ordenadores actualmente, se recomienda no bajar de 1.000 repeticiones.

- Métodos de los percentiles corregidos por sesgo y corregidos por sesgo incorporando una constante de aceleración: se introducen unas correcciones al método anterior. El segundo es el menos restrictivo.

#### 19.1.1.1. *Bootstrap* con STATA

Los alumnos de la asignatura de bioestadística se examinaron de sus conocimientos previos antes de empezar la asignatura. ¿Existe correlación entre dichos conocimientos previos *(pretest)* y la nota final de la asignatura *(final)*? Se calcula el coeficiente de correlación con la instrucción **correlate** y el intervalo de confianza con el método *bootstrap*.

En la instrucción **bootstrap** es necesario hacer referencia al estimador cuyo error estándar se quiere averiguar. Con la instrucción **return list** después de la orden (**correlate,** en el ejemplo), STATA devuelve la lista de todos los estadísticos almacenados y el modo de referirse a ellos (en este caso, **r(rho)**).

La opción **reps** indica el número de submuestras que se toman, **bca** sirve para que se calcule el intervalo de confianza corregido por sesgo, y la incorporación de la constante de aceleración, **seed(#)**, establece la semilla de aleatorización para asegurar la futura reproducibilidad de los resultados; si no se estableciese, el azar haría que, al repetir el proceso, pudiese encontrarse otro resultado. Como puede comprobarse, en este caso la aproximación del cálculo del intervalo de confianza por medio de la normal no es válida (no existen coeficientes de correlación superiores a 1) (fig. 19.1).

### 19.1.2. *Jackknife*

El *jackknife* es otro procedimiento de remuestreo. A diferencia de *bootstrap*, *jackknife* toma todas las posibles submuestras de tamaño $n - 1$ (o $n - k$). Cuando el tamaño es $n - 1$, debido a los requisitos de aplicación no es válido para el cálculo de estadísticos de posición (sí lo es $n - k$, ya que se relajan estos requisitos de aplicación).

En STATA le corresponde la instrucción **jackknife.** Siguiendo con el ejemplo anterior, para calcular el intervalo de confianza para la media de la variable **pretest** habría que introducir la siguiente instrucción:

```
jackknife r(mean): summarize    pretest

Jackknife results                                 Number of obs      =        10
                                                  Replications       =        10

      command:  summarize pretest
        _jk_1:  r(mean)
          n():  r(N)

------------------------------------------------------------------------------
             |             Jackknife
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _jk_1 |         2   .7333333     3.55   0.006     .9410849    4.258915
------------------------------------------------------------------------------
```

Cabe destacar que, para muchas de las órdenes de estimación de parámetros en STATA, existe la opción de calcular los intervalos de confianza utilizando tanto *jackknife* como *bootstrap*. Por ejemplo, en una regresión se podría añadir a la instrucción **regress** la opción que se enuncia a continuación:

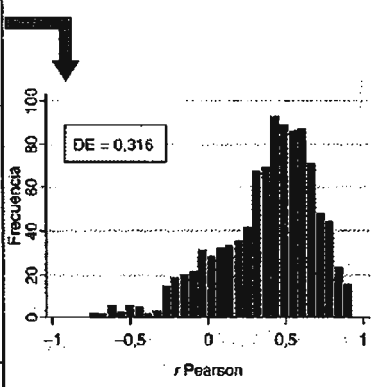
```
vce(bootstrap, reps(1000) bca seed(1))
```

correlate *pretest final*

| | pretest | final |
|---|---|---|
| pretest | 1.0000 | |
| final | 0.4376 | 1.0000 |

bootstrap r(rho), reps(1000) bca seed(123): cor *pretest final*

| Muestra | | Submuestra 1 | | Submuestra 2 | | Submuestra 1.000 | |
|---|---|---|---|---|---|---|---|
| *pretest* | *final* | *pretest* | *final* | *pretest* | *final* | *pretest* | *final* |
| 0 | 5 | 1 | 8 | 0 | 4 | 3 | 9 |
| 2 | 6 | 3 | 7 | 3 | 7 | 4 | 5 |
| 4 | 5 | 8 | 10 | 2 | 10 | 2 | 10 |
| 3 | 7 | 4 | 5 | 8 | 10 | 3 | 3 |
| 3 | 9 | 4 | 5 | 3 | 7 | 2 | 6 |
| 1 | 8 | 1 | 8 | 4 | 5 | 2 | 10 |
| 2 | 10 | 0 | 5 | 1 | 8 | 1 | 8 |
| 0 | 4 | 3 | 9 | 3 | 7 | 3 | 3 |
| 3 | 3 | 0 | 4 | 2 | 6 | 3 | 9 |
| 8 | 10 | 8 | 10 | 2 | 6 | 2 | 10 |
| $r = 0{,}4376$ | | $r_1 = 0{,}6234$ | | $r_2 = 0{,}5052$ | | $r_{1.000} = -0{,}4850$ | |




estat bootstrap, all

```
Bootstrap results                                 Number of obs      =        10
                                                  Replications       =      1000

      command:  correlate pretest final
        _bs_1:  r(rho)
```

| | Observed Coef. | Bias | Bootstrap Std. Err. | [95% Conf. | Interval] | |
|---|---|---|---|---|---|---|
| _bs_1 | .43755307 | −.0628217 | .31647274 | −.1827221 | 1.057828 | (N) |
| | | | | −.4178035 | .8316773 | (P) |
| | | | | −.4348283 | .8280459 | (BC) |
| | | | | −.2498859 | .8836523 | (BCa) |

```
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval
(BCa)  bias-corrected and accelerated confidence interval
```

**Figura 19.1** *Bootstrap* con STATA. Al tratarse de un coeficiente de correlación en una muestra pequeña, la aproximación normal (N) no es válida.

## 19.2. MÉTODO DE CAPTURA-RECAPTURA PARA INDAGAR EL TAMAÑO DE UNA POBLACIÓN

El método de muestreo por captura y recaptura es usado para estimar el tamaño de una población completa. En un principio se aplicaba este método, sobre todo, a las poblaciones biológicas, pero actualmente existen muchas aplicaciones del método para estimar el tamaño de las poblaciones humanas (2).

Consiste en capturar una muestra ($n_1$) de individuos de una población, etiquetarlos, anotar su número y devolverlos a la población. Posteriormente se vuelve a muestrear la población eligiendo otra muestra ($n_2$), en la que se hallará la coincidencia de volver a encontrar a algunos de los individuos etiquetados *(t)* en la primera muestra ($n_1$). La proporción de individuos etiquetados en la segunda muestra *(p)* debería ser representativa de la proporción de individuos etiquetados en la población.

Por tanto, se puede calcular el tamaño de la población mediante la siguiente fórmula:

$$\text{Total de la población} = n_1/p = n_1/(t/n_2) = n_1 \times n_2/t$$

Esta fórmula estima el tamaño del total de la población. Para calcular el intervalo de confianza se usa una varianza del estimador con esta fórmula:

$$s^2 = (n_1+1)(n_2+1)/(t+1)-1$$

Ejemplo: un epidemiólogo realiza un registro de personas indigentes en una ciudad, registrando a 50 personas en una base de datos. Dos meses más tarde, repite el registro y localiza a 220 personas, 35 de las cuales ya estaban etiquetadas porque fueron registradas en el estudio previo. La población total estimada de personas indigentes en esa ciudad sería:

$$\text{Total} = n_1 \times n_2/t = 50 \times 220/35 = 314{,}28$$

El total de personas indigentes en esa ciudad sería 314. La varianza sería:

$$s^2 = (n_1+1)(n_2+1)/(t+1)-1 = (50+1)(220+1)/(35+1)-1 = 312{,}08$$

Por tanto, el intervalo de confianza al 95% de la población total valdría:

$$\text{IC}95\%: 314{,}28 \pm 1{,}96 \times (312{,}08)^{0{,}5} = (279{,}65 \text{ a } 348{,}90)$$

Para estimar el tamaño poblacional con mayor precisión se pueden seguir haciendo muestreos. No obstante, el cálculo de los tamaños poblacionales mediante un sistema que incluya más de dos muestreos se extiende más allá de los objetivos de este capítulo.

La fórmula anterior es útil para calcular el tamaño poblacional en poblaciones *cerradas*, que son aquellas que tienen un tamaño constante durante el estudio. Las poblaciones abiertas, en cambio, se definen como aquellas en las que ocurren adiciones (nacimientos, inmigraciones) y deleciones (muertes, emigraciones) durante el estudio. Un ejemplo gráfico para comprender la diferencia entre poblaciones abiertas y cerradas es comparar el autobús con el avión como poblaciones. En un autobús, la población es abierta, ya que a lo largo del trayecto suben y bajan viajeros. En cambio, el avión es una población cerrada, dado que ningún pasajero abandona o se incorpora a la aeronave durante el trayecto. Para calcular el tamaño poblacional de poblaciones abiertas existen otros métodos estadísticos de captura y recaptura, pero son más complejos que lo explicado anteriormente.

Aunque originalmente esta técnica estaba destinada al recuento de poblaciones animales, se está aplicando en el ámbito de la epidemiología como un método eficiente para estimar el tamaño de poblaciones de difícil acceso (p. ej., número de consumidores ilegales de sustancias, personas sin techo, prostitutas, etc.) (3).

No es necesario que los sucesivos muestreos sean realizados por el mismo grupo de investigadores. Se puede recurrir a organizaciones que posean bases de datos con una muestra de la población de estudio (p. ej., asociaciones benéficas que atiendan a usuarios de drogas parenterales). Sin embargo, las bases de datos de estas organizaciones pueden llevar a sesgos de selección, al contener una mayor proporción de personas que buscan asistencia, lo que causaría una infraestimación del tamaño poblacional (4).

## 19.3. ANÁLISIS DE DECISIONES

El análisis de decisiones consta de seis pasos:

1. Definir el problema.
2. Definir los objetivos.
3. Estructurar el problema.
4. Incluir probabilidades.
5. Análisis cuantitativo.
6. Interpretación.

### 19.3.1. Definir el problema

Antes de tomar una decisión, hay que plantear el problema con todas las posibles decisiones. En el ejemplo, se trata de decidir si se realiza o no cribado prenatal poblacional del síndrome de Down en todas las mujeres embarazadas (**5,6**). Se ha asumido que la triple prueba sérica tiene un bajo valor predictivo positivo (VPP) < 5% y que, con la amniocentesis o la biopsia de vellosidades coriónicas, se pierde el 1% de los fetos (**5-10**). Cabría preguntarse qué beneficios aporta el cribado y si estos superan a los riesgos, sobre todo al no existir una alternativa terapéutica que resuelva la alteración genética. Los supuestos que aquí se han asumido y las alternativas técnicas que van apareciendo pueden hacer variar en los próximos años el árbol de probabilidad y las probabilidades asumidas (**11**). De todos modos, el parámetro clave es la probabilidad de que el hijo de la embarazada padezca síndrome de Down (prevalencia, probabilidad pretest), que es muy baja (en torno a 0,001), y los valores de sensibilidad (en torno al 85%) y especificidad (en torno al 95%) de las diferentes pruebas no invasivas (*non-invasive prenatal testing*, NIPT). Se ha simplificado mucho el problema por motivos de espacio y didácticos. Puede encontrarse un análisis más detallado en otras fuentes (**6,11**).

### 19.3.2. Definir los objetivos

Hay que plantear qué se pretende alcanzar con la decisión: reducción de costes, aumento de la calidad de vida, disminución de la mortalidad, etc. A veces se puede definir más de un objetivo. Estas metas pueden tener carácter objetivo (coste económico, supervivencia, etc.) o subjetivo (preferencia o *utilidad*). En este último caso, se les ha de adjudicar un valor. La *utilidad* consiste en la preferencia que da una persona a una situación sobre otra. La *utilidad* puede variar de un individuo a otro, por lo cual es importante definir un valor de *utilidad* aceptable para la población ante un análisis de decisiones con la *utilidad* como objetivo. Se debe asignar una utilidad a cada posible desenlace del proceso. En el ejemplo, los posibles desenlaces serían:

- Recién nacido normal.
- Recién nacido con síndrome de Down (falso negativo en el cribado).
- Eutanasia prenatal.
- Pérdida fetal como consecuencia del cribado.

Podría pensarse en otras dos posibilidades más:

- Cribado positivo para síndrome de Down, pero el embarazo prosigue hasta su nacimiento.
- Pérdida fetal debida al cribado de un feto con síndrome de Down.

Estas dos últimas alternativas se descartaron con el fin de simplificar el ejemplo: la primera por la inconsistencia de recabar información prenatal para luego no cambiar de opinión, y la segunda, por su bajísima probabilidad.

Una vez enumeradas las alternativas, como individuo (o como sociedad) se puede decidir optimizar uno o varios de estos objetivos: disminución de pérdidas fetales, aumento de la probabilidad de que los recién nacidos no tengan síndrome de Down, etc. Además, se puede otorgar un valor subjetivo *(utilidad)* a cada una de estas posibilidades e intentar maximizar la cantidad de utilidad que se gana.

### 19.3.3. Estructurar el problema

En esta fase se deben definir las alternativas posibles de resultados ante nuestra decisión, que se representarán en un árbol de decisiones. Cada rama del árbol se dividirá en dos o más ramas a partir de un *nudo de decisión* (allí donde se debe decidir, representado por un cuadrado) o un *nudo de probabilidad* (allí donde las cosas ocurren por azar, sin que nosotros tomemos una decisión, representado por un círculo). Cada rama puede volver a ramificarse mediante nuevos nudos. Generalmente, un árbol de decisión comienza con un nudo de decisión, para ramificarse posteriormente con nudos de probabilidad. En el ejemplo, el árbol tiene un nudo de decisión: hacer o no cribado. Después, progresa hacia el primer nudo probabilístico, que contiene las probabilidades de que la embarazada acepte o no el test. Si la embarazada acepta el test, aparece un nuevo nudo probabilístico en función del VPP de la prueba de cribado (5%), para clasificar el embarazo en alto riesgo o bajo riesgo. Algunas de las embarazadas de alto riesgo accederán a someterse a amniocentesis o biopsia de las vellosidades coriónicas, lo que origina un nuevo nudo probabilístico. Finalmente, el proceso termina con los cuatro posibles desenlaces propuestos en el apartado anterior (fig. 19.2).

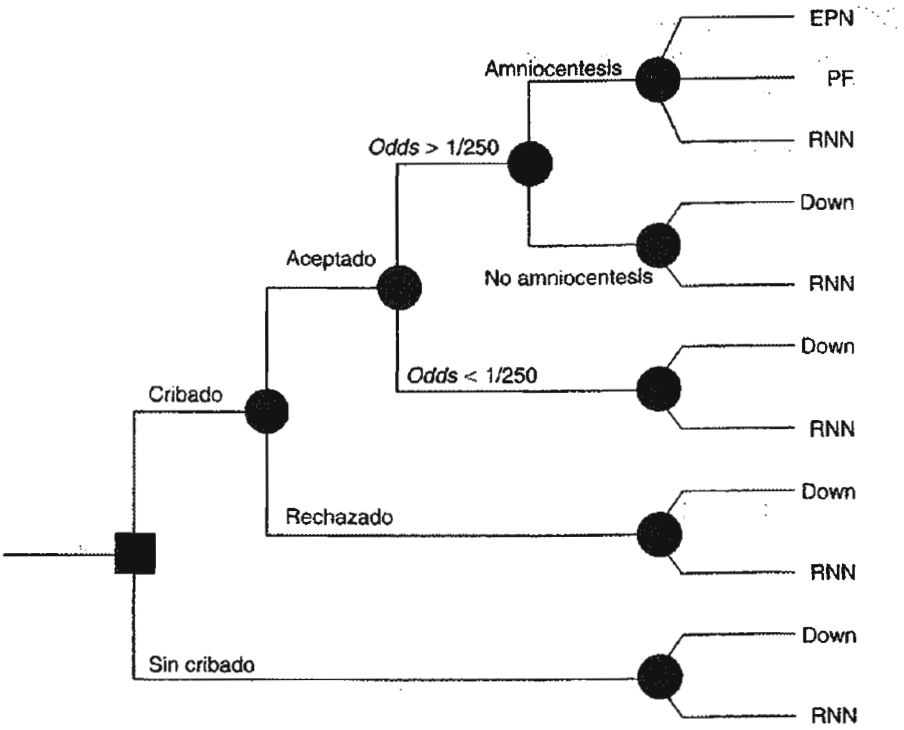



**Figura 19.2** Árbol de probabilidad para un análisis de decisiones. Down, nacimiento de un niño con síndrome de Down; EPN, eutanasia prenatal intencional; PF, pérdida fetal no intencional; RNN, nacimiento de un recién nacido normal.

*Tabla 19.1 Datos de probabilidades utilizados para el ejemplo del árbol de decisión*

| | PROBABILIDAD (%) |
|---|---|
| **Aceptación de triple prueba** | |
| Rechazarán hacerse el cribado | 20 |
| Aceptarán el cribado | 80 |
| **Validez de las pruebas de cribado no invasivas** | |
| Valor predictivo positivo | 5 |
| Resultarán de alto riesgo ($odds_{postest} > 1/250$) | 5 |
| Resultarán de bajo riesgo ($odds_{postest} \leq 1/250$) | 95 |
| **Amniocentesis** | |
| Riesgo de pérdida fetal | 1 |
| Embarazos de alto riesgo que rechazarán amniocentesis | 25 |
| | **ODDS** |
| ***Odds* postest de presentar síndrome de Down** | |
| Si el cribado las clasificó de alto riesgo | 1/75 |
| Si el cribado las clasificó de bajo riesgo | 1/2.800 |

### 19.3.4. Incluir probabilidades

Se asignará a cada suceso su probabilidad de que ocurra (tabla 19.1). Se puede expresar en forma de probabilidad o en forma de *odds*. Estas probabilidades se deben obtener a partir de la mejor evidencia científica disponible, preferiblemente no usando una sola fuente, sino todas las disponibles que sean de alta calidad. Las probabilidades de todas las ramas que componen un nudo deben sumar 1.

En algunas patologías es posible que los enfermos tengan recaídas, se cronifiquen, etc. Se pueden elaborar nudos de probabilidad para estas cuestiones, lo cual daría lugar a árboles de decisiones más complicados, con posibilidades de retornos y de procesos cíclicos. En estos casos, es más útil afrontar el árbol de decisiones mediante unos modelos denominados *procesos de Markov*, cuya explicación excede los objetivos de este capítulo. Normalmente se representan en el árbol como una M mayúscula rodeada de un círculo. Aunque los cálculos se podrían realizar manualmente, se suele recurrir a programas informáticos.

### 19.3.5. Análisis de decisión: estimación de los desenlaces

Para realizar el análisis se multiplica cada desenlace por la respectiva probabilidad y, posteriormente, se suman de derecha a izquierda todas las ramas que surgen de cada división (*repliegue* del árbol) (fig. 19.3).

En el ejemplo, se han hecho los cálculos sobre dos hipotéticas cohortes de 100.000 embarazadas que recorrerían cada una de las dos ramas del árbol que parten del nudo de decisión. Una rama simula que se ofrece el cribado a todas las embarazadas y otra rama, que se decide no ofrecerles el cribado de manera universal. Si en el análisis solo se cuentan *vidas humanas*, el objetivo sería que naciese el mayor número de niños *sanos*. En la rama que simula que se ofrece el cribado a todas las embarazadas nacerían 99.870 niños sanos y 60 con síndrome de Down. En el grupo sin cribado nacerían 99.990 niños sanos y 100 niños con síndrome de Down (las diferencias se explican por las pérdidas fetales y la eutanasia prenatal). Por tanto, si el objetivo fuera aumentar el número de niños sanos, la decisión ha de ser no ofrecer el cribado universal.

No obstante, si el objetivo fuera maximizar la *calidad de vida global*, habría que asignar utilidades o preferencias subjetivas a cada uno de los posibles desenlaces. Los métodos para averiguar tales preferencias son complejos y exceden los objetivos de este capítulo. En el ejemplo, se ha supuesto

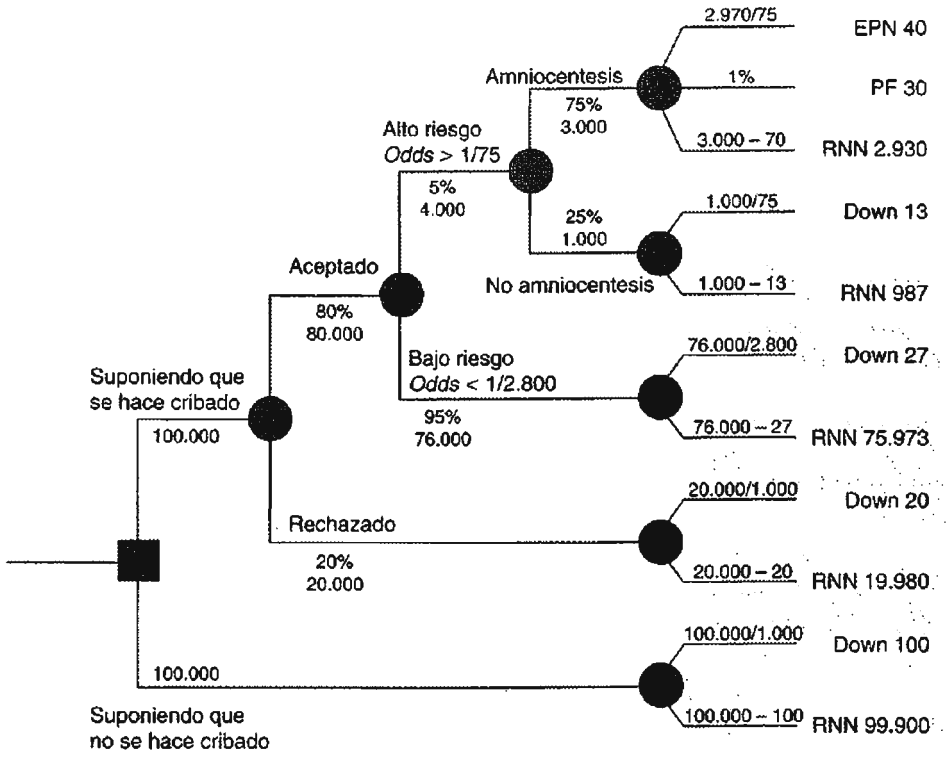



**Figura 19.3** Asignación de las probabilidades y estimación del número de desenlaces en el ejemplo de análisis de la decisión.

que se asignaría una utilidad de 1 (máxima utilidad posible) al recién nacido sano y una utilidad de 0 a la pérdida fetal no intencional. Resulta discutible qué valor o utilidad asignar a los recién nacidos con síndrome de Down. Si se le asigna una utilidad de 0,5 (valor intermedio entre nacer totalmente sano y la pérdida fetal), se obtendría una mayor utilidad esperada global si se decide no realizar el cribado prenatal frente a la implantación del cribado (tabla 19.2):

- Utilidad esperada tras implantación del cribado: $99.870 \times 1 + 60 \times 0,5 = 99.900$.
- Utilidad esperada si *no* se implanta el cribado: $99.990 \times 1 + 90 \times 0,5 = 99.950$.

Solo asignando al recién nacido con síndrome de Down una utilidad inferior a –0,75 (bastante inferior a la pérdida fetal), la decisión de implantar el cribado tendría una utilidad esperada mayor que el rechazo del mismo.

El ejercicio de cambiar los supuestos (p. ej., la utilidad) para ver cómo varía el resultado del análisis se conoce como *análisis de sensibilidad*, y valora la solidez y la estabilidad de los resultados.

### 19.3.6. Interpretación del análisis de la decisión

El análisis de decisión es una herramienta orientativa que permite hacer explícitas las preferencias y valorar las consecuencias de diferentes cursos de acción, así como la probabilidad de obtener unos resultados u otros según las decisiones que se deseen asumir. No es, por tanto, un método normativo que sustituya a la ética a la hora de decir lo que se debe hacer (12).

*Tabla 19.2 Utilidades esperadas en el ejemplo de análisis de decisión*

| | EVENTOS | UTILIDADES | UTILIDADES ESPERADAS |
|---|---|---|---|
| **Con cribado** | | | |
| EPN | 40 | 0 | 0 |
| PF | 30 | 0 | 0 |
| Síndrome de Down | 60 | 0,5 | 30 |
| RNN | 99.870 | 1 | 99.870 |
| | | Total | 99.900 |
| **Sin cribado** | | | |
| EPN | 0 | 0 | 0 |
| PF | 0 | 0 | 0 |
| Síndrome de Down | 100 | 0,5 | 50 |
| RNN | 99.900 | 1 | 99.900 |
| | | Total | 99.950 |

*EPN, eutanasia prenatal intencional; PF, pérdida fetal no intencional; RNN, nacimiento de un recién nacido normal. En este ejercicio, se asigna a los nacimientos de niños con síndrome de Down una* utilidad *de 0,5.*

## 19.4. MODELOS FLEXIBLES DE REGRESIÓN CON INTERVALOS DE CONFIANZA *(SPLINES)*

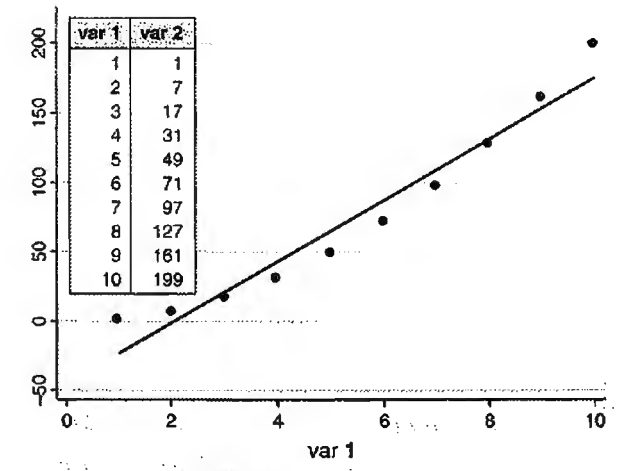
Siempre que se recojan datos, se recomienda registrarlos de la manera más informativa y desagregada posible. Por ejemplo, es preferible siempre calcular el índice de masa corporal (IMC) como variable cuantitativa continua, porque se recogió el peso y la talla, que recoger únicamente si el paciente era obeso o no. Ahora bien, una vez llegados a la fase de análisis, cuando se desea estudiar una variable de exposición de naturaleza cuantitativa, en los métodos convencionales existen dos opciones:

- Introducir la variable como cuantitativa y estudiar el efecto que tiene el cambio en una unidad sobre la variable de desenlace. Siguiendo con el ejemplo del IMC, y si se estudiase su relación con la fibrilación auricular, en una regresión logística se obtendría la OR correspondiente al efecto de un incremento en 1 kg/m² en el IMC sobre la *odds* de fibrilación auricular. Esta aproximación no es útil cuando se sospecha que puede haber *una relación no lineal.*
- Como alternativa, categorizar la variable IMC y ver la OR para cada categoría con respecto a una categoría que se deja como referencia (OR = 1). Así se podría observar una cierta relación no lineal.

Esta segunda alternativa de la categorización será preferible cuando se sospeche una relación no lineal. Sin embargo, no resuelve todos los problemas. Existen al menos dos problemas potenciales con la categorización:

- La arbitrariedad en la elección del punto de corte, ya que distintos puntos de corte podrían arrojar resultados muy diferentes. Por ejemplo, se podría hallar una OR significativa para un IMC > 35 kg/m² cuando se definen categorías basadas en puntos de corte exactos (<20, 20-25, 25-30, 30-35, >35 kg/m²), pero esta asociación podría perderse si las categorías se basasen en quintiles. Todo dependerá de cómo esté distribuido el IMC en la muestra.
- La categorización asume implícitamente que la OR será la misma dentro de cada categoría. En el ejemplo, la primera categorización no distinguiría entre el riesgo asociado a tener IMC = 30,01 y el relacionado con tener IMC = 34,99 kg/m². Probablemente, ambos riesgos diferirán. Lo mismo sucedería con el riesgo de quien tiene IMC = 15,5 y el de quien posee 19,99 kg/m². Desde el punto de vista biológico, parece poco pertinente considerarlos iguales.

Los modelos flexibles de regresión (en inglés, *splines*) intentan dar solución a este problema. De todos modos, también ellos pueden presentar sus limitaciones, y lo ideal sería combinar estos métodos flexibles con la categorización tradicional (13).

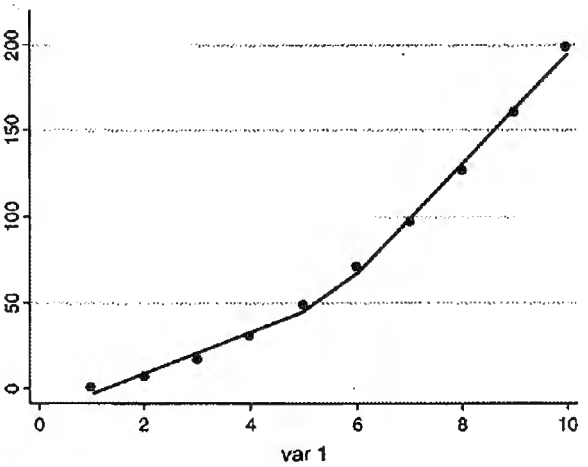


| var 1 | var 2 |
|---|---|
| 1 | 1 |
| 2 | 7 |
| 3 | 17 |
| 4 | 31 |
| 5 | 49 |
| 6 | 71 |
| 7 | 97 |
| 8 | 127 |
| 9 | 161 |
| 10 | 199 |



```
tw sc var2 var1 || lfit var2 var1
```

A





```
mkspline s=var1, cubic nknots(3)
regress var2 s*
predict y
tw sc var2 var1 || line y var1
```

B

**Figura 19.4** Valoración de la relación dosis-respuesta y creación de *splines*. **A.** *Spline* lineal. **B.** *Spline* cúbico.

Los *splines* más utilizados son los de tipo cúbico *(restricted cubic splines)*. Una explicación matemática detallada del modelo de *splines* (13-16) excedería los objetivos de este capítulo (fig. 19.4).

Los modelos flexibles de *splines* estudian el efecto sobre el desenlace de cada valor observado de la variable cuantitativa independiente en comparación con el valor de referencia, y tienen en cuenta cómo se asocian los valores próximos (alisamiento).

En la figura anterior se presenta un modelo sencillo para una regresión lineal simple. Sin embargo, además de la predicción, es interesante representar las bandas correspondientes al intervalo de confianza.

A continuación se presenta un ejemplo con STATA para una regresión logística. Para utilizar una base de datos más completa, se usará una de las disponibles en la web y a las que se puede acceder desde STATA, con la instrucción **webuse lbw.**

También se puede encontrar esta base de datos (lbw.dta) en las direcciones:

http://www.stata-press.com/data/r9/rmain.html

http://www.unav.edu/departamento/preventiva/recursos_bioestadistica

Esta base de datos corresponde a un estudio de casos y controles de niños con bajo peso al nacer (*low* = 1 para los casos y *low* = 0 para los controles). Las exposiciones son algunas características de la madre. Supóngase que se pretende estudiar el efecto de la edad de la madre sobre el riesgo de que el niño nazca con bajo peso. Se podría pensar que la relación sería no lineal.

En primer lugar, hay que crear nuevas variables que representen la edad y permitan adaptarse a una forma alisada para la relación dosis-respuesta. Para conseguirlo se usará la primera orden, que es **mkspline,** y se encuentra en:

**Data → Create or change data → Other variable-creation commands → Linear and cubic spline construction**

Se debe seleccionar una nueva variable de destino (asignándole el nombre que se desee) y una variable origen, que es una independiente cuantitativa que ya existía. Aquí se hace con la edad:

```
mkspline age_s = age, nknots( 4) cubic displayknots
```

La instrucción **mkspline** genera una serie de variables (los *splines*) llamadas *age_s*, seguidas de un número, que representan la edad. El número 4 indica que se ha deseado que haya cuatro puntos de inflexión. Habrá tantas variables nuevas como puntos de inflexión menos uno (*age_s1*, *age_s2* y *age_s3*).

En esta instrucción es necesaria la opción **cubic** para elegir *splines* cúbicos y, además, se puede variar el número de puntos de inflexión con la opción **nknots(#)**. Se pide que enseñe los puntos de inflexión elegidos con la opción **displayknots,** como se ha hecho aquí.

Después de la instrucción anterior, se debe emplear:

```
mat knots = r(knots)
```

Con esto se pide que se almacene en la memoria del programa la matriz con los puntos de inflexión, que será necesaria más adelante.

Antes de seguir con el estudio de la relación no lineal, habría que comprobar que, efectivamente, los datos son compatibles con dicha relación no lineal entre la exposición y el desenlace. Para esto se realiza la regresión correspondiente, añadiendo como variables independientes todas las creadas con **mkspline,** seguido de la orden *testparm* para las mismas variables:

```
logit low age_s* smoke race
testparm age_s*
```

```
( 1)  [low]age_s1 = 0
( 2)  [low]age_s2 = 0
( 3)  [low]age_s3 = 0

           chi2(  3) =    4.57
         Prob > chi2 =    0.2059
```

Con esta instrucción se realiza un test de hipótesis para contrastar si el polinomio (representado por las nuevas variables *age_s1*, *age_s2* y *age_s3*) aporta algo comparado con la variable original.

En este caso, como ocurrirá muchas veces, el resultado no es estadísticamente significativo ($p = 0{,}2059$), con lo que la información que aporta la transformación de la variable no es relevante con respecto a la original. Llegados a este punto, habría que abandonar los modelos flexibles y volver a los métodos clásicos. Sin embargo, y solo con finalidad didáctica, se presenta a continuación cómo se realizaría el resto del análisis, si el resultado de este test hubiera sido estadísticamente significativo.

En primer lugar, debido a que se necesita hacer referencia a cada uno de los niveles de la variable original (*age* en el ejemplo), se puede utilizar la instrucción **levelsof,** y luego referirse a esos niveles (que STATA habrá guardado internamente como **r(levels)**). Además, se puede añadir el prefijo **quietly** para que en la ventana de resultados no aparezca toda la lista de valores de la variable *age*.

```
levelsof age
```

Así se obtienen todos los valores únicos de edad contenidos en la base de datos.

A continuación, se deberá usar la orden **xbrcspline,** pero se trata de una orden que no viene por defecto en STATA[2], y se puede descargar desde: http://ideas.repec.org/c/boc/bocode/s457092.html.

Una vez descargada e instalada, la orden **xbrcspline** permite determinar, para cada posible valor observado de edad y obtenido con la instrucción **levelsof,** una estimación de la OR y sus límites de confianza, que se almacenarán en las variables que se escriban dentro del paréntesis después de **gen.**

```
xbrcspline age_s, values(`r(levels)') ///
mat knots(knots) eform gen(edad or lb ub)
```

Esta instrucción devolverá la siguiente salida:

```
. xbrcspline age_s, values(`r(levels)') ///
> matknots(knots) eform gen(edad or lb ub)

Reference value for age = 14
```

| age | exp(XB) | LB | UB |
|---|---|---|---|
| 14 | 1.00 | 1.00 | 1.00 |
| 15 | 0.83 | 0.60 | 1.15 |
| 16 | 0.69 | 0.36 | 1.33 |
| 17 | 0.58 | 0.22 | 1.53 |
| 18 | 0.50 | 0.14 | 1.75 |
| 19 | 0.45 | 0.10 | 1.98 |
| 20 | 0.44 | 0.09 | 2.21 |
| 21 | 0.47 | 0.09 | 2.44 |
| 22 | 0.54 | 0.11 | 2.68 |
| 23 | 0.63 | 0.13 | 2.91 |
| 24 | 0.72 | 0.17 | 3.12 |
| 25 | 0.78 | 0.19 | 3.20 |
| 26 | 0.77 | 0.19 | 3.05 |
| 27 | 0.70 | 0.18 | 2.72 |
| 28 | 0.59 | 0.15 | 2.31 |
| 29 | 0.47 | 0.11 | 1.92 |
| 30 | 0.35 | 0.08 | 1.61 |
| 31 | 0.26 | 0.05 | 1.38 |
| 32 | 0.19 | 0.03 | 1.21 |
| 33 | 0.13 | 0.02 | 1.09 |
| 34 | 0.09 | 0.01 | 1.00 |
| 35 | 0.07 | 0.00 | 0.93 |
| 36 | 0.05 | 0.00 | 0.87 |
| 45 | 0.00 | 0.00 | 0.61 |

Se ha estimado una OR, con sus límites de confianza para cada posible valor de la edad, tomando como referencia el valor mínimo de la edad (14 años). Si se desease usar otro valor como referencia, bastaría sustituir la anterior orden por:

2 La instrucción *xbrcspline* se introdujo por Nicola Orsini, del Instituto Karolinska, como una orden opcional de STATA en el simposio de usuarios de STATA de países nórdicos y bálticos de 2009. Se puede encontrar más información en: http://www.stata.com/meeting/sweden09/se09_orsini.pdf.

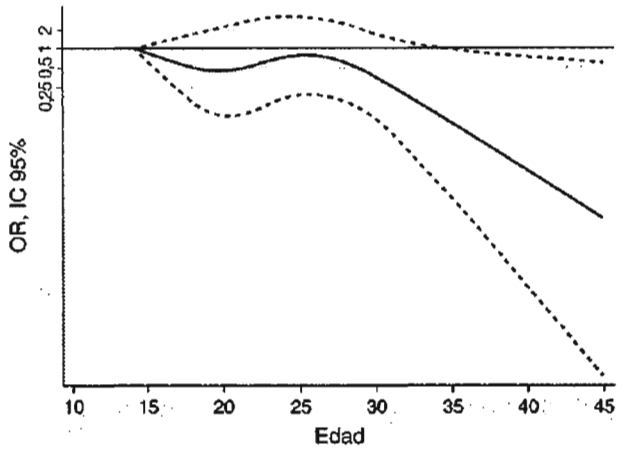



**Figura 19.5** Gráfico de la relación entre la edad y el riesgo *(odds ratio)* de bajo peso al nacer estimado mediante un modelo flexible *(restricted cubic splines)*.

```
xbrcspline age_s, values(`r(levels)') ref(23) ///
mat knots(knots) eform gen(edad or lb ub)
```

Ahora se ha tomado como referencia un valor próximo a la media de edad de la base de datos (edad media = 23,2). El valor que se desee adoptar como referencia debe existir realmente en la base de datos.

Por último, solo queda la representación gráfica de este modelo flexible. Se representan con una línea continua los valores de la estimación puntual de la OR y, con líneas discontinuas, sus límites de confianza al 95%. Además, se añade una línea horizontal en OR = 1, que representa el valor nulo para la OR.

```
tw (line lb ub or edad , lp(- - l) lc(black black
black)), ///
legend(off) yline(1) ytit("OR, 95% CI") ///
ysca(log range(.15 2)) xlab(10(5)45) ylab(.25 .5 1 2)
```

Véase la figura 19.5.

## 19.5. VALORES PERDIDOS *(MISSING)* Y MÉTODOS DE IMPUTACIÓN

Se consideran valores perdidos o faltantes *(missing)* todos aquellos datos que deberían estar y, sin embargo, faltan en la base de datos. Algunos de estos valores tienen que estar necesariamente ausentes y, por lo tanto, no suponen un problema (p. ej., el número de cigarrillos fumados al día en una persona nunca fumadora o la edad de la menarquia en un varón). Sin embargo, el resto de valores perdidos suponen un problema que a veces hay que afrontar en la fase de análisis de datos. La mejor aproximación con respecto a los valores faltantes es la *prevención*, como sucede siempre: mejor prevenir que curar. Una alta calidad y meticulosidad al recoger datos reducirá los

valores perdidos y evitará futuros problemas. Esta es la mejor solución. Aun así, es frecuente que existan valores faltantes, y a veces es inevitable, por muy cuidadosa que sea la recogida de datos, ya que intervienen factores que el investigador no puede controlar totalmente, como abandonos, contestación incompleta de cuestionarios, etc.

### 19.5.1. Exploración de valores perdidos en STATA

Se pueden describir los valores perdidos de un grupo de variables con una instrucción sencilla:

```
misstable summarize varlist
```

Con esta instrucción se obtendrá el número de valores perdidos de cada una de las variables que se listen a continuación. Si alguna de las variables no tiene valores faltantes, no aparecerá en la tabla de resultados.

```
. misstable summarize var1 var2 var3
```

| | | | | Obs<. | | |
|---|---|---|---|---|---|---|
| Variable | Obs=. | Obs>. | Obs<. | Unique values | Min | Max |
| var1 | 542 | | 5,150 | 2 | 0 | 1 |
| var2 | 188 | | 5,504 | >500 | 11.74885 | 36.00221 |

Una posible opción que se puede añadir después de una coma a la orden **misstable summarize** es **generate** (*varname*), por ejemplo:

```
misstable sum  var1 var2 var3,  generate( perd_)
```

Se creará así una variable nueva por cada variable de la lista *(var1, var2, var3)* que tenga valores faltantes, con el mismo nombre que tenga la variable a la que hace referencia, pero añadiendo el prefijo *perd_* (o cualquier otro prefijo que se elija). Estas nuevas variables *(perd_var1, perd_var2, perd_var3)* valdrán 0 cuando la observación esté recogida y 1 cuando esté perdida.

Otra instrucción interesante es:

```
misstable  pattern varlist
```

En este caso, ofrece una descripción del patrón que siguen los valores perdidos en un conjunto de variables.

```
. misstable pattern var1 var2 var3

Missing-value patterns
  (1 means complete)

             |  Pattern
     Percent |  1  2
-------------+-------
         90% |  1  1
             |
           6 |  1  0
           3 |  0  0
-------------+-------
        100% |

Variables are  (1) var2  (2) var1
```

Siguiendo con el ejemplo anterior, habría un 90% de observaciones completas, un 6% a las que les falta la variable *var1* y un 3% que tiene perdidas ambas variables *var1* y *var2*.

¿Cómo afrontar el problema de los valores faltantes una vez que han ocurrido? En general, la opción establecida por defecto en los programas de estimación en muchos paquetes estadísticos es eliminar todas las observaciones que tengan un valor perdido en alguna de las variables que intervienen en el modelo (los programas a veces informan de que lo han hecho con mensajes u órdenes como *casewise deletion* o *listwise deletion*). Con esta aproximación se dejaría de utilizar mucha información. Por ejemplo, si en un modelo multivariable que incluye 10 variables independientes hay un 10% de sujetos con su valor perdido en al menos una variable, la muestra probablemente quedaría reducida a menos de la mitad. Así se pierde mucha potencia, y lo peor es que, probablemente, se introducirá un sesgo de selección, pues los sujetos que tienen valores perdidos suelen ser distintos de los que no los tienen. En consecuencia, casi nunca se puede asumir que los valores faltantes son completamente explicables por el azar (*missing completely at random* o MCAR). Tal supuesto sería muy difícilmente sostenible.

En contraposición al patrón de valores faltantes completamente al azar (*missing completely at random* o MCAR) hay otro posible supuesto, que es el de valores perdidos al azar (*missing at random* o MAR). El supuesto MCAR supone que *solo el azar*, y nada más, explica por qué unos sujetos tienen valores perdidos y otros no. Tal supuesto sería muy difícilmente asumible en cualquier escenario realista de investigación. Un supuesto menos exigente y más realista y asumible es que los valores perdidos seguirían un patrón de valores faltantes al azar (*missing at random*, MAR) una vez conocidas ciertas variables que representan características asociadas a una mayor o menor probabilidad de que haya valores faltantes. Por ejemplo, ser varón o ser fumador puede asociarse a dejar más respuestas en blanco en un cuestionario. El sexo y el tabaco serían predictores de valores perdidos. Si estas variables son capaces de explicar el patrón de valores faltantes, bastaría conocer tales variables para reemplazar de algún modo los valores en cuestión. A menudo es posible asumir que, una vez controladas o ajustadas ciertas variables clave, lo que resta para explicar los valores *missing* es solo el azar, y entonces se estaría en una situación de MAR, mucho más realista. Parece asumible que algunas de las variables observadas permitirían predecir parcialmente el hecho de que exista un valor perdido. En cualquier caso, el supuesto MAR implica que se han recogido adecuadamente los predictores de los valores faltantes. Casi todo lo que sigue supone asumir que el patrón de valores perdidos es de tipo MAR.

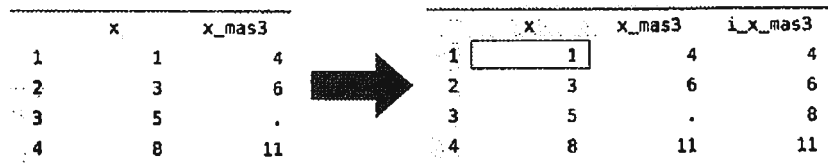
Una vez asumido este supuesto, se podría optar por asignar un valor (imputar) a todas aquellas observaciones faltantes, basándose en los datos similares que sí se han observado. Cuando se desconoce el valor para una observación, el primer recurso al que se podría acudir es asignar a ese valor desconocido la media (o mediana) del resto de observaciones que sí se conocen. Sin embargo, es más frecuente que dejen de contestar las personas con valores extremos que las que tienen valores próximos a la media, con lo cual esta no sería una buena aproximación, precisamente porque no se puede asumir el supuesto MCAR.

Dando un paso más, se podrían predecir los valores faltantes a partir de las variables que se conocen y que, según se sabe, están relacionadas con la variable faltante. Por ejemplo, si se sabe que el sexo, la edad y el peso son predictores de la talla, se podría usar un modelo de regresión lineal para hallar el valor de talla predicho para quien tiene una determinada edad, sexo y peso, pero cuya estatura se ignora.

### 19.5.2. Imputación simple

En la imputación simple se seleccionan variables que predigan bien la variable con valores perdidos y se introducen en un modelo de regresión como variables independientes, y como variable dependiente la que tiene valores faltantes. Se asigna entonces el valor predicho por esta regresión a la observación perdida (fig. 19.6).

Este método tiene el problema de que reduce artificialmente la varianza de la variable imputada. Sin embargo, podría bastar cuando se trate de una variable que solo se usa para ajustar y haya pocos valores faltantes (<5% o <10%, según otros autores).



| | x | x_mas3 |
|---|---|---|
| 1 | 1 | 4 |
| 2 | 3 | 6 |
| 3 | 5 | . |
| 4 | 8 | 11 |

| | x | x_mas3 | i_x_mas3 |
|---|---|---|---|
| 1 | 1 | 4 | 4 |
| 2 | 3 | 6 | 6 |
| 3 | 5 | . | 8 |
| 4 | 8 | 11 | 11 |

impute x_mas3 x, gen(i_x_mas3)

[impute *vardep varlist* , gen(*new_var*)]

**Figura 19.6** Imputación simple con STATA.

### 19.5.3. Imputación múltiple

El método de imputación múltiple resuelve el problema de reducir artificialmente (y erróneamente) la varianza de las variables imputadas, que es lo que ocurría con la imputación simple. La imputación múltiple ya no reducirá la varianza de la variable imputada.

Este método consiste en imputar a partir de variables observadas como se hacía antes el valor faltante, pero ahora no una, sino muchas veces; se incorpora siempre un componente de variabilidad al azar. Cada vez que se realiza una imputación, se generará una nueva base de datos. Así, este método de imputación multiplicará la base de datos con un valor imputado al azar distinto en cada base nueva. Entonces se lleva a cabo la estimación en cada uno de esos escenarios y, por último, se combinan todas las estimaciones para obtener un resultado único. La variabilidad en la estimación calculada en las diversas bases de datos sirve para aumentar la variabilidad global; así no se reduce erróneamente la varianza.

Se pueden encontrar las expresiones matemáticas de combinación de las varianzas intrabase de datos e interbase de datos en algunas revisiones sobre el tema (17). La prevención y tratamiento de los valores *missing* requiere cada vez mayor atención en la investigación biomédica (18).

La imputación múltiple se basa (al igual que la simple) en que los valores perdidos siguen un patrón al azar (MAR), pero *no* completamente al azar, con lo cual algunas de las variables observadas permitirían predecir parcialmente el hecho de que haya un valor perdido.

Se deben introducir como predictoras en la imputación las siguientes variables:

- Todas aquellas que se piensen introducir en el modelo multivariable final, incluida la variable dependiente. Se ha discutido que esta introducción podría formar parte de un razonamiento circular y resultar *tendenciosa,* pero hoy en día se considera todo lo contrario: si no se aprovechase la variable de desenlace como predictor de los valores *missing* de las variables independientes, los resultados se sesgarían hacia el nulo, lo que llevaría a una imputación subóptima, porque el desenlace de un estudio puede estar relacionado con el hecho de que haya valores faltantes.
- Todas aquellas variables que, sin ser interesantes para el modelo de regresión, estén relacionadas con que exista un valor perdido. Se puede comprobar cuáles son estas variables llevando a cabo una regresión logística, en la que se introducen como independientes todas las variables candidatas (p. ej., lugar de procedencia, estado civil, etc.) y, como dependiente, la variable *perd_var* (v. apartado 19.5.1).
- Todas aquellas variables que puedan tener relación con la variable que se imputará.

Como consideraciones generales, las variables que se introducen como predictoras en la imputación no deben tener muchos valores faltantes ellas mismas. Además, el número de variables

que se introducirán no debe ser demasiado elevado, ya que empezará a existir colinealidad entre ellas y se complicará el modelo sin ningún beneficio práctico.

En STATA, la familia de instrucciones que corresponde a la imputación múltiple es *mi*. La secuencia habitual de órdenes que deben indicarse es:

```
mi set flong                          #formato : m bases datos
mi register imputed   varlist_i #designar var. a imputar
mi register regular   varlist  #designar predictores
mi impute mvn  varlist_i ///
varlist , add(20) rseed(1)      #imputar nuevos valores
mi estimate: regress   y ///
x1 x2 x3...                           #modelo
```

En primer lugar, hay que establecer cómo se crearán y almacenarán las sucesivas bases de datos para la imputación múltiple. Con la primera instrucción (**mi set**) se dice a STATA cómo almacenar las nuevas bases de datos que se creen después de la original. Si se carece de espacio suficiente, se debe sustituir **flong** por **mlong** y únicamente se añadirán las observaciones con valores imputados, sin repetir cada vez el resto de la base de datos.

STATA crea tres nuevas variables:

1. *_mi_miss:* # identificará con un 1 las observaciones imputadas y con 0 el resto.
2. *_mi_m:* # numera las m bases de datos que se van creando.
3. *_mi_id:* # número de identificación para cada sujeto (repetido en cada base).

A continuación se dará la orden **mi register imputed,** que irá seguida del listado de variables *(varlist_i)* que tengan valores *missing* y que sea preciso imputar. Esta orden designa en qué variables se realizará imputación. Una vez que se indican al programa las variables que tiene que imputar, STATA asigna un 1 en la variable *_mi_miss* a las observaciones que están perdidas al menos para alguna de ellas. Después, con la instrucción **mi register regular** se indica qué variables *(varlist)* no tienen valores perdidos o no van a ser imputadas. Se designan así las variables que no requieren imputación y que se usarán para predecir los valores imputados.

Seguidamente se utiliza la instrucción **mi impute mvn *varlist_i*, add(*20*) rseed(*1*)**[3]. Se introducen en la lista de variables todas aquellas que tengan que ser imputadas (es decir, las mismas que se incluyeron con **mi register imputed**); en la opción *add* se indica el número de bases de datos nuevas que se crearán (se recomienda que este número no sea inferior a 20, pues ahora es factible con la capacidad de los ordenadores actuales), y con la opción *rseed(#)* se establece la semilla de aleatorización igual que en otras instrucciones con componentes aleatorios, para asegurar que los resultados sean reproducibles e idénticos cuando se vuelva a ejecutar esta sintaxis. En la imputación es posible que a algunas variables se les imputen valores

3 En lugar de **mvn,** que se basa en aproximaciones a la normal y se utiliza para variables continuas, se podría utilizar **chained** para variables categóricas o cuantitativas discretas.

implausibles o que variables cualitativas o cuantitativas discretas acaben con valores decimales. Aunque parezca un problema, la recomendación es no *arreglar* estos valores (redondeando o aproximando al valor plausible más próximo), ya que se puede introducir un sesgo y empeorar la situación.

Por último, se completa el análisis al especificar el modelo que se desee aplicar usando ya los datos imputados. Para ello, delante de la orden convencional se incluye el prefijo:

```
mi estimate:
```

Puede elegirse, entre otras, alguna de las siguientes opciones según el modelo que se desee estimar:

```
mi estimate: regress y x1 x2 x3
mi estimate: logit caso x1 x2 x3
mi estimate: poisson caso x1 x2 x3,
exposure(person_years)
mi stset followup, failure(death==1)
mi estimate: stcox x1 x2 x3
```

Las últimas dos líneas, como es habitual en STATA, son las necesarias para un modelo de Cox. Una vez ejecutada la orden correspondiente, STATA realizará esa estimación en cada una de las bases de datos que se han creado y a continuación fusionará los resultados y ofrecerá el resultado global, integrando la variabilidad entre bases de datos en el error estándar de los coeficientes. Este procedimiento penaliza en la estimación la variabilidad entre las distintas imputaciones, de tal forma que tampoco aumente la potencia artificialmente. En la figura 19.7 se puede ver un ejemplo sencillo de imputación múltiple paso a paso.

Actualmente se deben preferir los métodos basados en la *imputación múltiple* y no usar otras aproximaciones a las que se ha recurrido con frecuencia en el pasado, como son usar una variable *dummy* (como si fuese una categoría más) para quienes tienen valores perdidos en esa variable, reemplazar los valores perdidos con el último valor recogido para ese sujeto en esa variable *(last value carried forward)* cuando se trata de medidas repetidas u otras aproximaciones basadas en buscar al vecino más parecido, y copiarle su dato para sustituir el valor perdido (**19**).

## 19.6. PONDERACIÓN POR EL INVERSO DE LA VARIANZA Y MODELOS ESTRUCTURALES MARGINALES

Imagínese el ejemplo de la tabla 19.3, en el que se valora si la exposición al alcohol incrementa el riesgo cardiovascular. Se aprecia que la estimación global o cruda del riesgo relativo está fuertemente confundida por el tabaco, ya que, cuando no se estratifica por tabaco, el alcohol parece comportarse como un fuerte factor de riesgo. Esto es falso (está confundido), ya que, *dentro* de cada estrato de exposición al tabaco, el alcohol apunta a todo lo contrario: es un protector. Es un caso típico de confusión, ya que los consumidores de alcohol tienen mayor probabilidad de estar expuestos al tabaco, y el tabaco es un fuerte factor de riesgo de la enfermedad estudiada.

```
mi set mlong
mi register imputed imc
mi register regular ///
sexo tabaco pas
```

| | id | sexo | imc | tabaco | pas | _mi_m | _mi_id | _mi_miss |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 25.7 | 1 | 110 | 0 | 1 | 0 |
| 2 | 2 | 1 | 30.2 | 0 | 125 | 0 | 2 | 0 |
| 3 | 3 | 1 | 19.6 | 0 | 90 | 0 | 3 | 0 |
| 4 | 4 | 0 | 22.3 | 1 | 100 | 0 | 4 | 0 |
| 5 | 5 | 0 | . | 0 | 95 | 0 | 5 | 1 |
| 6 | 6 | 0 | 17.4 | 1 | 95 | 0 | 6 | 0 |
| 7 | 7 | 0 | . | 1 | 120 | 0 | 7 | 1 |
| 8 | 8 | 1 | 36.1 | 1 | 135 | 0 | 8 | 0 |
| 9 | 9 | 0 | 23.9 | 0 | 100 | 0 | 9 | 0 |
| 10 | 10 | 1 | 18.2 | 0 | 90 | 0 | 10 | 0 |

```
mi impute mvn imc, add(20) rseed(1)
```

| | id | sexo | imc | tabaco | pas | _mi_m |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 25.7 | 1 | 110 | 0 |
| 2 | 2 | 1 | 30.2 | 0 | 125 | 0 |
| 3 | 3 | 1 | 19.6 | 0 | 90 | 0 |
| 4 | 4 | 0 | 22.3 | 1 | 100 | 0 |
| 5 | 5 | 0 | . | 0 | 95 | 0 |
| 6 | 6 | 0 | 17.4 | 1 | 95 | 0 |
| 7 | 7 | 0 | . | 1 | 120 | 0 |
| 8 | 8 | 1 | 36.1 | 1 | 135 | 0 |
| 9 | 9 | 0 | 23.9 | 0 | 100 | 0 |
| 10 | 10 | 1 | 18.2 | 0 | 90 | 0 |
| 11 | 5 | 0 | 18.1603 | 0 | 95 | 1 |
| 12 | 7 | 0 | 30.3743 | 1 | 120 | 1 |
| 13 | 5 | 0 | 43.874 | 0 | 95 | 2 |
| 14 | 7 | 0 | 38.708 | 1 | 120 | 2 |
| 15 | 5 | 0 | 13.275 | 0 | 95 | 3 |
| 16 | 7 | 0 | 21.7318 | 1 | 120 | 3 |

...hasta 20 veces.

```
Multivariate imputation                       Imputations =       20
Multivariate normal regression                     added =       20
Imputed: m=1 through m=20                        updated =        0

Prior: uniform                                Iterations =     2000
                                                burn-in =      100
                                                between =      100
```

| | Observations per m | | | |
|---|---|---|---|---|
| Variable | Complete | Incomplete | Imputed | Total |
| imc | 8 | 2 | 2 | 10 |

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)

```
mi estimate: regress pas imc tabaco sexo
```

```
Multiple-imputation estimates                 Imputations      =         20
Linear regression                             Number of obs    =         10
                                              Average RVI      =     0.7277
                                              Largest FMI      =     0.7339
                                              Complete DF      =          6
DF adjustment:   Small sample                 DF:     min      =       1.64
                                                      avg      =       2.27
                                                      max      =       2.57
Model F test:       Equal FMI                 F(   3,    1.3) =       2.84
Within VCE type:          OLS                 Prob > F         =     0.3565
```

| pas | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| imc | 1.565076 | .8202508 | 1.91 | 0.224 | -2.825581 | 5.955732 |
| sexo | 7.737752 | 10.17197 | 0.76 | 0.510 | -27.91081 | 43.38632 |
| tabaco | 12.0752 | 9.995251 | 1.21 | 0.332 | -24.60128 | 48.75169 |
| _cons | 58.51235 | 16.96211 | 3.45 | 0.056 | -2.973062 | 119.9978 |

**Figura 19.7** Imputación múltiple con STATA.

Se podría corregir la confusión con un modelo multivariable estándar de regresión logística o de regresión de Poisson en el que se ajuste por tabaco. No obstante, existe otro método de afrontar y corregir este problema. Se trata de asignar una ponderación variable a cada sujeto, eligiendo para ello los pesos de tal modo que desaparezca la asociación entre alcohol y tabaco, y el factor de confusión (tabaco en el ejemplo) se distribuya por igual entre expuestos y no expuestos al alcohol.

*Tabla 19.3 Relación del consumo de alcohol con el riesgo cardiovascular*

| TODO EL ESTUDIO | CASOS DE ECV | TOTAL | RIESGO RELATIVO |
|---|---|---|---|
| **Alcohol** | | | |
| Sí | 430 | 11.000 | $RR = \frac{\left(\frac{430}{11.000}\right)}{\left(\frac{204}{9.000}\right)} = 1,72$ |
| No | 204 | 9.000 | |
| **Estratificado por TABACO** | | | |
| **No fumadores** | | | |
| Alcohol | Casos de ECV | Total | Riesgo relativo |
| Sí | 30 | 3.000 | $RR = \frac{\left(\frac{30}{3.000}\right)}{\left(\frac{84}{7.000}\right)} = 0,83$ |
| No | 84 | 7.000 | |
| **Fumadores** | | | |
| Alcohol | Casos de ECV | Total | Riesgo relativo |
| Sí | 400 | 8.000 | $RR = \frac{\left(\frac{400}{8.000}\right)}{\left(\frac{120}{2.000}\right)} = 0,83$ |
| No | 120 | 2.000 | |

**Modelo crudo (Poisson)**

| ECV | IRR | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| alcohol | 1.724599 | .1466167 | 6.41 | 0.000 | 1.4599 | 2.037291 |
| _cons | .0226667 | .001587 | -54.09 | 0.000 | .0197602 | .0260006 |

**Modelo ponderado por IPW (Poisson)**

| ECV | IRR | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| alcohol | .8333334 | .0766272 | -1.98 | 0.047 | .6959028 | .9979046 |
| _cons | .036 | .0027507 | -43.51 | 0.000 | .030993 | .0418159 |

Con STATA se darán los siguientes pasos sucesivos:

```
logit alcohol tabaco # usa la exposición como v. dep.
predict p_cond       # calcula probabilidad de consumir
                     alcohol condicional a tabaco
replace p_cond=1-p_cond ///
if alcohol==0        # p. condicional de no expuestos
quietly sum alcohol  # descripción silenciosa alcohol
gen P_marg=r(mean)   # genera probabilidad marginal
replace P_marg=1-P_marg ///
if alcohol==0              # p. marginal de no expuestos
g IPW=P_marg/p_cond
logistic cas alc ///
[pweight=IPW], vce(r)      # modelo logístico ponderado
poisson cas alc ///
[pweight=IPW], irr vce(r)  # modelo de Poisson ponderado
```

Con esta secuencia de órdenes se crea una base de datos *fantasma* que distribuye la población, de modo que el uso del alcohol queda libre de confusión. Es *fantasma* porque, al ponderarla, se consigue que los sujetos que antes estaban infrarrepresentados ahora reciban mucho peso, y los que estaban sobrerrepresentados ahora se ponderen a la baja. Todo el secreto está en ponderarlos por el *inverso* de su probabilidad de estar expuestos (o no estarlo).

La primera orden (**logit**) ajusta un modelo de regresión logística. De él se obtiene, con la segunda orden (**predict**), la probabilidad predicha de ser consumidor de alcohol en función del tabaco. Esta probabilidad predicha solo se aplica a quienes de hecho consumen alcohol. En quienes no consumen se aplicará su complementario, que es la probabilidad predicha de *no* consumir alcohol (siguiente paso). A cada sujeto se le asigna una probabilidad de estar expuesto en función de los factores de confusión. A los expuestos se les asigna la probabilidad predicha de estar expuesto, y a los no expuestos, la de no estarlo. A cada uno lo suyo. Aquí, para introducir este método del modo más sencillo posible se ha usado solo un factor de confusión (el tabaco), pero en la primera orden (**logit**) podría haberse incluido como variables independientes un número muy amplio de posibles factores de confusión; entonces, las probabilidades predichas estarían condicionadas a cada posible combinación de esos factores de confusión.

Una vez calculadas las dos probabilidades condicionales (la de los expuestos y la de los no expuestos), bastaría con ponderar y usar como pesos el inverso de estas probabilidades para liberar el ejemplo de la confusión. Este tipo de métodos se conoce como *inverse probability*

*weight*, porque se basan en ponderar por el inverso de la probabilidad de la variable de exposición.

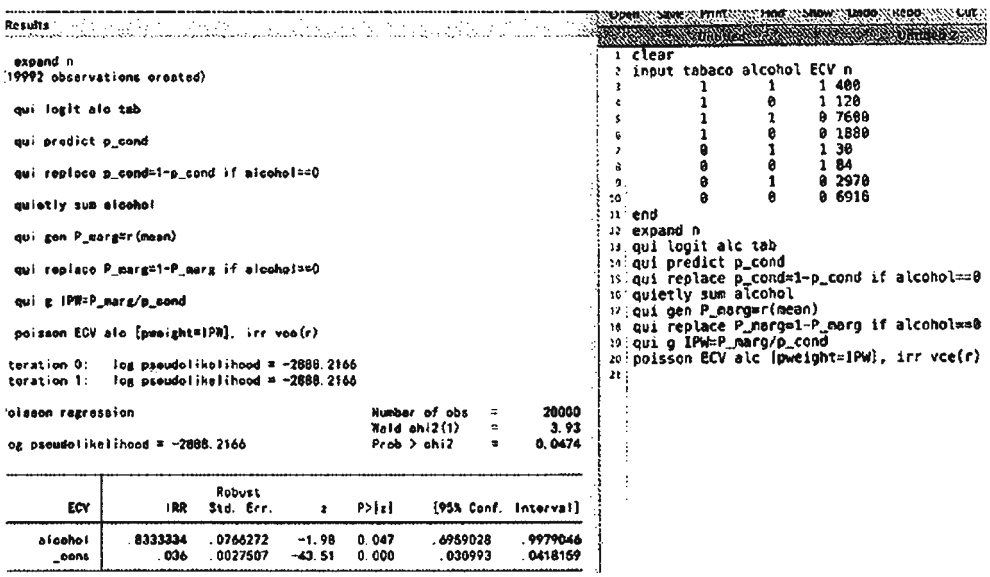
$$\text{Pesos(IPW)} = \frac{1}{\text{P(exposición} \mid \text{confusores)}}$$

Sin embargo, se requiere dar un paso más: hay que estabilizar los pesos para no influir artificialmente en las varianzas. Para ello, lo ideal es multiplicar el peso por la probabilidad *marginal* de exposición, que es simplemente la proporción de la muestra que está expuesta (para los expuestos) y la proporción no expuesta (para los no expuestos).

$$\text{Pesos estabilizados(IPW)} = \frac{\text{P(exposición)}}{\text{P(exposición} \mid \text{confusores)}}$$

La figura 19.8 indica todos los pasos que habría que dar en STATA para reproducir este ejemplo (ridículamente simple, por otra parte).

Estos métodos de ponderación por el inverso de la varianza son especialmente útiles en diseños observacionales (no experimentales) cuando se desean analizar como si fuesen un ensayo aleatorizado. Requieren siempre usar la opción de varianza robusta (**vce(robust)** en STATA). Pueden resultar de alta utilidad y constituirse en la única aproximación factible ante las frecuentes amenazas del sesgo por indicación (hay tratamientos médicos que se indican selectivamente a pacientes más graves, lo cual puede hacer que el tratamiento parezca peor), sobre todo cuando se están usando mediciones repetidas en el tiempo de exposición a ese tratamiento y el desenlace es también una medición repetida de variaciones en la gravedad o en la ocurrencia de complicaciones. Los modelos multivariables que aplican estos métodos o procedimientos análogos se denominan



```
Results
. expand n
(19992 observations created)
. qui logit alc tab
. qui predict p_cond
. qui replace p_cond=1-p_cond if alcohol==0
. quietly sum alcohol
. qui gen P_marg=r(mean)
. qui replace P_marg=1-P_marg if alcohol==0
. qui g IPW=P_marg/p_cond
. poisson ECV alc [pweight=IPW], irr vce(r)
Iteration 0:   log pseudolikelihood = -2888.2166
Iteration 1:   log pseudolikelihood = -2888.2166
Poisson regression                              Number of obs   =      20000
                                                Wald chi2(1)    =       3.93
Log pseudolikelihood = -2888.2166               Prob > chi2     =     0.0474
```

| ECV | IRR | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| alcohol | .8333334 | .0766272 | -1.98 | 0.047 | .6959028 | .9979046 |
| _cons | .036 | .0027507 | -43.51 | 0.000 | .030993 | .0418159 |

```
Open  Save  Print  Find  Show  Undo  Redo  Cut
Untitled
1  clear
2  input tabaco alcohol ECV n
3          1       1     1 400
4          1       0     1 120
5          1       1     0 7600
6          1       0     0 1880
7          0       1     1 30
8          0       0     1 84
9          0       1     0 2970
10         0       0     0 6916
11 end
12 expand n
13 qui logit alc tab
14 qui predict p_cond
15 qui replace p_cond=1-p_cond if alcohol==0
16 quietly sum alcohol
17 qui gen P_marg=r(mean)
18 qui replace P_marg=1-P_marg if alcohol==0
19 qui g IPW=P_marg/p_cond
20 poisson ECV alc [pweight=IPW], irr vce(r)
21
```

**Figura 19.8** Ejemplo (ridículamente simple; v. tabla 19.3) de ponderación por el inverso de la probabilidad de exposición hecho con STATA.

modelos estructurales marginales *(marginal structural models)* o de tipo g-estimación. También sirven para analizar ensayos (**20**).

Las limitaciones de espacio impiden profundizar más en estos modelos, pero pueden encontrarse explicaciones más detalladas (sin dejar de ser *amigables*) en referencias recientes (**21**).

## 19.7. ÍNDICES DE PROPENSIÓN *(PROPENSITY SCORES)*

Los métodos basados en índices de propensión *(propensity scores)* se han ido usando cada vez más para controlar múltiples factores de confusión en investigación observacional, es decir, cuando no se pueden usar diseños experimentales aleatorizados. Los diseños observacionales *no* se suelen considerar como el estándar de oro de la inferencia causal. Sin embargo, cuando se analizan bien, si no es ético o no es factible realizar un ensayo (**22**), entonces son suficientemente fuertes como para proporcionar evidencias causales firmes (**23**).

Los *propensity scores* permiten combinar un gran número de posibles factores de confusión en una sola variable (el *score*). En principio se asumirá que se trata de la propensión a estar expuesto en función de una serie de covariables. Los *propensity scores* se definen como la probabilidad para cada sujeto de estar expuesto a un tratamiento (o factor, en general) específico, dadas sus covariables medidas previamente al tratamiento. Son probabilidades condicionales. La condición es el patrón de covariables que presenta ese sujeto. Si se estudiase la supervivencia asociada a un tratamiento oncológico y solo interesasen tres covariables previas (edad, sexo y estadio tumoral), el *propensity score* se definiría como:

$$p(\text{tratamiento} \mid \text{sexo}, \text{edad}, \text{estadio})$$

Esta probabilidad se puede hallar con dos pasos. El primero consiste en preparar un modelo logístico con el tratamiento como variable dependiente, y el segundo, en usar una orden (**predict** en STATA) para extraer para cada sujeto su valor predicho de estar sometido al tratamiento en función de estas tres covariables, según el modelo logístico. Los *propensity scores* oscilarán así entre 0 y 1 y reflejarán la probabilidad estimada, basada en el sexo, la edad y el estadio tumoral de que ese sujeto reciba el tratamiento de interés.

Más adelante se pueden formar estratos en función de los *propensity scores.* Dentro de cada estrato de propensión, algunos pacientes habrán recibido el tratamiento de interés y otros no, pero todos tendrán unas probabilidades estimadas similares de recibirlo dadas sus covariables observadas. Al estimar la asociación tratamiento-enfermedad dentro de estos niveles homogéneos de los *propensity scores,* en teoría, y con el supuesto de que no hay más factores de confusión no medidos, se puede alcanzar un escenario que se ha denominado de *aleatorización virtual,* en el que subgrupos de pacientes comparables constituyen grupos tratados y no tratados que se pueden analizar sin confusión (**24**).

Los *propensity scores* tienen la ventaja de que reducen la cantidad de covariables independientes que han de incluirse en el modelo, pues se puede ajustar por el *propensity score* nada más. Esto ha permitido, en ocasiones, el ajuste por un gran número de variables de una manera eficiente en algunos estudios observacionales (**25**). Las estrategias analíticas de ajuste pueden ser dividir el *propensity score* en categorías (quintiles o deciles), como *splines,* o como una variable continua que asumiría que el *propensity score* es un predictor lineal e introducirlo de esta manera en un modelo convencional de regresión lineal, logística, de Poisson o de Cox. Otra alternativa consiste en usar un diseño emparejado por el *propensity score* y emparejar a cada participante expuesto con otro no expuesto usando el *propensity score* como variable para seleccionar la pareja (emparejamiento 1:1). Una tercera posibilidad en análisis de supervivencia sería el uso de un modelo de regresión de Cox

en el que se utilice el *propensity score* como variable de estratificación. También se pueden usar los *propensity scores* para aplicar el método de ponderación por el inverso de la probabilidad, que se ha explicado en el apartado anterior. En las simulaciones que se han hecho no hay diferencias importantes entre estas diversas aproximaciones (**26**), aunque otras simulaciones optan por el *inverse probability weighting* ante efectos no uniformes del tratamiento (**25**). Lo que sí se pierde es la capacidad de identificar y distinguir en el análisis otros predictores del desenlace distintos de la exposición principal y el conjunto del *propensity score*.

## 19.8. ECUACIONES DE ESTIMACIÓN GENERALIZADAS *(GENERALIZED ESTIMATING EQUATIONS, GEE)*

Las ecuaciones de estimación generalizadas se engloban dentro de los análisis longitudinales (v. capítulo 9). Son análisis de tipo longitudinal todos aquellos que miden en más de una ocasión la variable de desenlace (respuesta) en contraposición a los transversales, en los que se mide una única vez. El objetivo de todo análisis longitudinal es estudiar el modo en que los cambios dentro del mismo individuo predicen la variable de desenlace a lo largo del tiempo, y relacionar esta variable de desenlace con los niveles de distintas covariables, que también pueden haber sido medidas repetidas veces en el tiempo en cada sujeto. Dado que se estudian los cambios *dentro* de cada individuo, aquellos factores que permanezcan constantes (tanto si se han medido como si no) quedarán controlados en el análisis, lo que lleva a estimaciones mucho más precisas y más parecidas al modelo contrafáctico (cada sujeto es su propio control), que permiten aproximarse a la verdadera causalidad. Además, se tiene en cuenta y se penaliza la posible correlación entre las distintas observaciones dentro del mismo sujeto (**27**).

Para realizar este análisis con STATA, lo primero que hay que hacer es definir el panel de datos, con la instrucción **xtset** seguida de las variables que identifican al sujeto *(id)* y al tiempo *(visita)*. Los datos tienen que encontrarse en formato alargado *(long)*, es decir, cada fila de la base de datos se corresponderá con una observación, y no con un sujeto, y existirán filas repetidas (una por cada medición repetida) para cada sujeto (fig. 19.9).

La instrucción general que se ha de utilizar es **xtgee,** seguida de la variable dependiente y las variables independientes. Como variables independientes se pueden introducir en el modelo tanto aquellas que no varían a lo largo del seguimiento (p. ej., el sexo) como las que varían durante el seguimiento y de las cuales se tiene información (p. ej., el estado civil).

En función de la respuesta o desenlace que se valore como resultado (variable dependiente), se dispone de distintas opciones para la instrucción **xtgee,** con respecto a la distribución de la variable dependiente *(family)* y la función por la que se relaciona **(link)** (tabla 19.4).

Además, hay que indicar a STATA cuál es la estructura de correlación que existe entre las observaciones del mismo sujeto **(cor).** Si se especifica una estructura de correlación independiente

| | id | visita | tabaco | ecv |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 2 | 0 | 0 |
| 3 | 2 | 1 | 0 | 0 |
| 4 | 2 | 2 | 0 | 1 |
| 5 | 3 | 1 | 0 | 0 |
| 6 | 3 | 2 | 0 | 0 |
| 7 | 4 | 1 | 1 | 0 |
| 8 | 4 | 2 | 1 | 1 |
| 9 | 5 | 1 | 1 | 0 |
| 10 | 5 | 2 | 1 | 0 |

**Figura 19.9** Formato de la base de datos para usar ecuaciones de estimación generalizadas (GEE).

*Tabla 19.4 Opciones de modelos en ecuaciones de estimación generalizada (GEE) en STATA*

| VARIABLE DEPENDIENTE | *FAMILY* | *LINK* |
|---|---|---|
| Cuantitativa continua | gaussian | identity |
| Cualitativa dicotómica | binomial | logit |
| Cualitativa dicotómica | poisson | log |

(con la opción **independent**), se obtendrán instrucciones equivalentes a **regress, logit** o **poisson,** según el caso. Esto no es lo adecuado, ya que seguramente habrá alguna correlación intrasujeto que sea preciso corregir. La estructura de correlación con los requisitos de aplicación más laxos es la desestructurada (**unstructured**), con lo que se podría utilizar en cualquier caso. Sin embargo, supone una pequeña pérdida de potencia comparada con otras estructuras (siempre que estas se ajusten bien a los datos). Desde el punto de vista práctico, los resultados que se obtienen con las distintas estructuras de correlación (excepto la independiente) son razonablemente parecidos.

A continuación se presenta un resultado obtenido en STATA con un ejemplo ficticio muy simple en el que únicamente hay dos variables: el tabaco (variable independiente) y la enfermedad cardiovascular (variable dependiente), ambas codificadas como 0 = no, 1 = sí. En primer lugar se define el panel de datos, la segunda instrucción corresponde a la descripción del panel, y la última es propiamente la orden de estimación.

```
xtset id visita
xtdescribe
xtgee ecv tabaco, ///
family(binomial) link(logit) vce(robust) cor(uns) eform
```

```
GEE population-averaged model                   Number of obs      =       800
Group and time vars:              id visita     Number of groups   =       400
Link:                                 logit     Obs per group: min =         2
Family:                             binomial                    avg =       2.0
Correlation:                  unstructured                    max =         2
                                                Wald chi2(1)       =      2.75
Scale parameter:                         1      Prob > chi2        =    0.0974

                                  (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Semirobust
         ecv | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tabaco |   2.418462   1.288684     1.66   0.097     .8510875    6.872332
       _cons |   .0127226   .0056964    -9.75   0.000     .0052901    .0305979
------------------------------------------------------------------------------
```

Desde el punto de vista de la epidemiología, es interesante establecer un *período de inducción* o tiempo causalmente necesario que debe transcurrir entre la exposición y el desenlace. En el ejemplo anterior, se comparaba el hábito tabáquico con el desarrollo de enfermedad cardiovascular en ese período. Sin embargo, si las visitas fuesen anuales, se podría relacionar el hábito tabáquico en la visita $k$ con la enfermedad cardiovascular en la siguiente visita ($k + 1$), dejando 1 año como período mínimo de inducción (fig. 19.10).

1. Se recoge una base de datos en la que cada fila corresponde a un sujeto.
2. Se nombran las variables de la siguiente manera:

| Visita | Hábito tabáquico | Enfermedad cardiovascular |
|---|---|---|
| 1 | tabaco 1 | Partimos de personas sanas |
| 2 | tabaco 2 | ecv1 |
| 3 | tabaco 3 | ecv2 |
| 4 | | ecv3 |

3. Se cambia la base de datos de formato ancho a formato largo (cada fila una observación). Se crea automáticamente la variable *visita* que contendrá el número que había detrás de cada variable:

| | id | tabaco1 | tabaco2 | tabaco3 | ecv1 | ecv2 | ecv3 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 1 | 1 | 1 | 1 | . | . |

reshape long tabaco ecv, i(id) j(visita)

| | id | visita | tabaco | ecv |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 2 | 0 | 0 |
| 3 | 1 | 3 | 1 | 1 |
| 4 | 2 | 1 | 0 | 0 |
| 5 | 2 | 2 | 0 | 0 |
| 6 | 2 | 3 | 0 | 0 |
| 7 | 3 | 1 | 1 | 1 |
| 8 | 3 | 2 | 1 | . |
| 9 | 3 | 3 | 1 | . |

4. Se procede al análisis con la instrucción **xtgee**.

**Figura 19.10** Nomenclatura de las variables para establecer un período de inducción mínimo de 1 año.

Como práctica aconsejable para no perder el contacto con los datos, se recomienda, antes de hacer el análisis de ecuaciones de estimación generalizada, ajustar el modelo de regresión correspondiente (lineal múltiple, logística, Poisson) para cada visita independientemente y evaluar la consistencia de los resultados.

```
bysort visit: logistic ecv tabaco
```

## REFERENCIAS


1. Chernick MR. Bootstrap Methods: A Practitioner's Guide. New York: Wiley; 1999.
2. Pollock KH. Capture-recapture Models: an Overview. Vitoria, Eskuratzea Eta Berreskuratzea Bidezko Laginketa. Euskal Estatistika-Erakundea (Instituto Vasco de Estadística); 1995.
3. Post LA, Zhang H, Barboza GE, Conner T. Simulations Demonstrate Feasibility of Capture-Recapture. Hawai, Proceeding of International Conference on Statistics and Mathematics. International Conference on Statistics and Mathematics; 2006.
4. Corrao G, Bagnardi V, Vittadini G, Favilli S. Capture-recapture methods to size alcohol related problems in a population. J Epidemiol Community Health 2000;54(8):603-10.

5. Seguí-Gómez M, Núñez-Córdoba JM, Guillén-Grima F. Evaluación económica y análisis de decisiones. En: Martínez-González MA, editor. Conceptos de salud pública y estrategias preventivas: un manual para ciencias de la salud. Barcelona: Elsevier; 2013. p. 75-82.

6. Ortega Benito JM. Los marcadores séricos en el diagnóstico prenatal del síndrome de Down: la prueba triple. Med Clin (Barc) 1995;105:264-8.

7. Discroll DA, Gross S. Prenatal screening for aneuploidy. N Engl J Med 2009;360:2556-62.

8. Khoshnood B, De Vigan C, Vodovar V, Goujard J, Goffinet F. A population-based evaluation of the impact of antenatal screening for Down's syndrome in France. 1981-2000. BJOG 2004;111:485-90.

9. Rousseau T, Amar E, Ferdynus C, Thauvin-Robinet C, Gouyon JB, Sagot P. Variations in the prevalence of Down's syndrome in the French population between 1978 and 2005. J Gynecol Obstet Biol Reprod (Paris) 2010;39:290-6.

10. Morris JK, Alberman E. Trends in Down's syndrome live births and antenatal diagnoses in England and Wales from 1989 to 2008: analysis of data from the National Down Syndrome Cytogenetic Register. BMJ 2009;339:b3794.

11. O'Leary P, Maxwell S, Murch A, Hendrie D. Prenatal screening for Down syndrome in Australia: Costs and benefits of current and novel screening strategies. Aust N Z J Obstet Gynaecol 2013;53(5):425-33.

12. Thorton JG, Lilfort RJ. Decision analysis for medical managers. BMJ 1995;310:791-4.

13. Steenland K, Deddens JA. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. Epidemiology 2004;15(1):63-70.

14. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995;6:356-65.

15. Figueiras A, Cadarso-Suarez C. Application of Nonparametric Models for Calculating Odds Ratios and Their Confidence Intervals for Continuous Exposures. Am J Epidemiol 2001;154(3):264-75.

16. Greenland S, Michels KB, Robins JM, Poole C, Willett WC. Presenting statistical uncertainty in trends and dose-response relations. Am J Epidemiol 1999;149:1077-86.

17. Raghunathan TE. What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 2004;25:99-117.

18. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med 2012;367(14):1355-60.

19. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393.

20. Hernán MA, Hernández-Díaz S, Robins JM. Randomized Trials Analyzed as Observational Studies. Ann Intern Med 2013 Sep 10. [Epub ahead of print].

21. Mortimer KM, Neugebauer R, van der Laan M, Tager IB. An Application of Model-Fitting Procedures for Marginal Structural Models. Am J Epidemiol 2005;162:382-8.

22. Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 2003;327:1459-61.

23. Hill AB. The Environment and Disease: Associationor Causation? Proc R Soc Med 1965;58:295-300.

24. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.

25. Mangano DT, Tudor IC, Dietzel C. Multicenter Study of Perioperative Ischemia Research Group; Ischemia Research and Education Foundation. The Risk Associated with Aprotinin in Cardiac Surgery. N Engl J Med 2006;354:353-65.

26. Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect. Am J Epidemiol 2006;163:262-70.

27. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New Jersey: Wiley, 2004.

## *Tabla de la distribución normal tipificada*

Dentro de la tabla se proporciona el valor de $p$ para $+z_a$ o para $-z_a$ (área de una sola cola). Se deben buscar los dos primeros dígitos de $z_a$ en la primera columna vertical y el último dígito de $z_a$ en la primera fila horizontal.

| $z_a$ | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0,5000 | 0,4960 | 0,4920 | 0,4880 | 0,4840 | 0,4801 | 0,4761 | 0,4721 | 0,4681 | 0,4641 |
| **0,1** | 0,4602 | 0,4562 | 0,4522 | 0,4483 | 0,4443 | 0,4404 | 0,4364 | 0,4325 | 0,4286 | 0,4247 |
| **0,2** | 0,4207 | 0,4168 | 0,4129 | 0,4090 | 0,4052 | 0,4013 | 0,3974 | 0,3936 | 0,3897 | 0,3859 |
| **0,3** | 0,3821 | 0,3783 | 0,3745 | 0,3707 | 0,3669 | 0,3632 | 0,3594 | 0,3557 | 0,3520 | 0,3483 |
| **0,4** | 0,3446 | 0,3409 | 0,3372 | 0,3336 | 0,3300 | 0,3264 | 0,3228 | 0,3192 | 0,3156 | 0,3121 |
| **0,5** | 0,3085 | 0,3050 | 0,3015 | 0,2981 | 0,2946 | 0,2912 | 0,2877 | 0,2843 | 0,2810 | 0,2776 |
| **0,6** | 0,2743 | 0,2709 | 0,2676 | 0,2643 | 0,2611 | 0,2578 | 0,2546 | 0,2514 | 0,2483 | 0,2451 |
| **0,7** | 0,2420 | 0,2389 | 0,2358 | 0,2327 | 0,2296 | 0,2266 | 0,2236 | 0,2206 | 0,2177 | 0,2148 |
| **0,8** | 0,2119 | 0,2090 | 0,2061 | 0,2033 | 0,2005 | 0,1977 | 0,1949 | 0,1922 | 0,1894 | 0,1867 |
| **0,9** | 0,1841 | 0,1814 | 0,1788 | 0,1762 | 0,1736 | 0,1711 | 0,1685 | 0,1660 | 0,1635 | 0,1611 |
| **1,0** | 0,1587 | 0,1562 | 0,1539 | 0,1515 | 0,1492 | 0,1469 | 0,1446 | 0,1423 | 0,1401 | 0,1379 |
| **1,1** | 0,1357 | 0,1335 | 0,1314 | 0,1292 | 0,1271 | 0,1251 | 0,1230 | 0,1210 | 0,1190 | 0,1170 |
| **1,2** | 0,1151 | 0,1131 | 0,1112 | 0,1093 | 0,1075 | 0,1056 | 0,1038 | 0,1020 | 0,1003 | 0,0985 |
| **1,3** | 0,0968 | 0,0951 | 0,0934 | 0,0918 | 0,0901 | 0,0885 | 0,0869 | 0,0853 | 0,0838 | 0,0823 |
| **1,4** | 0,0808 | 0,0793 | 0,0778 | 0,0764 | 0,0749 | 0,0735 | 0,0721 | 0,0708 | 0,0694 | 0,0681 |
| **1,5** | 0,0668 | 0,0655 | 0,0643 | 0,0630 | 0,0618 | 0,0606 | 0,0594 | 0,0582 | 0,0571 | 0,0559 |
| **1,6** | 0,0548 | 0,0537 | 0,0526 | 0,0516 | 0,0505 | 0,0495 | 0,0485 | 0,0475 | 0,0465 | 0,0455 |
| **1,7** | 0,0446 | 0,0436 | 0,0427 | 0,0418 | 0,0409 | 0,0401 | 0,0392 | 0,0384 | 0,0375 | 0,0367 |
| **1,8** | 0,0359 | 0,0351 | 0,0344 | 0,0336 | 0,0329 | 0,0322 | 0,0314 | 0,0307 | 0,0301 | 0,0294 |
| **1,9** | 0,0287 | 0,0281 | 0,0274 | 0,0268 | 0,0262 | 0,0256 | 0,0250 | 0,0244 | 0,0239 | 0,0233 |
| **2,0** | 0,0228 | 0,0222 | 0,0217 | 0,0212 | 0,0207 | 0,0202 | 0,0197 | 0,0192 | 0,0188 | 0,0183 |
| **2,1** | 0,0179 | 0,0174 | 0,0170 | 0,0166 | 0,0162 | 0,0158 | 0,0154 | 0,0150 | 0,0146 | 0,0143 |
| **2,2** | 0,0139 | 0,0136 | 0,0132 | 0,0129 | 0,0125 | 0,0122 | 0,0119 | 0,0116 | 0,0113 | 0,0110 |
| **2,3** | 0,0107 | 0,0104 | 0,0102 | 0,0099 | 0,0096 | 0,0094 | 0,0091 | 0,0089 | 0,0087 | 0,0084 |
| **2,4** | 0,0082 | 0,0080 | 0,0078 | 0,0075 | 0,0073 | 0,0071 | 0,0069 | 0,0068 | 0,0066 | 0,0064 |
| **2,5** | 0,0062 | 0,0060 | 0,0059 | 0,0057 | 0,0055 | 0,0054 | 0,0052 | 0,0051 | 0,0049 | 0,0048 |
| **2,6** | 0,0047 | 0,0045 | 0,0044 | 0,0043 | 0,0041 | 0,0040 | 0,0039 | 0,0038 | 0,0037 | 0,0036 |
| **2,7** | 0,0035 | 0,0034 | 0,0033 | 0,0032 | 0,0031 | 0,0030 | 0,0029 | 0,0028 | 0,0027 | 0,0026 |
| **2,8** | 0,0026 | 0,0025 | 0,0024 | 0,0023 | 0,0023 | 0,0022 | 0,0021 | 0,0021 | 0,0020 | 0,0019 |
| **2,9** | 0,0019 | 0,0018 | 0,0018 | 0,0017 | 0,0016 | 0,0016 | 0,0015 | 0,0015 | 0,0014 | 0,0014 |
| **3,0** | 0,0013 | 0,0013 | 0,0013 | 0,0012 | 0,0012 | 0,0011 | 0,0011 | 0,0011 | 0,0010 | 0,0010 |
| **3,1** | 0,0010 | 0,0009 | 0,0009 | 0,0009 | 0,0008 | 0,0008 | 0,0008 | 0,0008 | 0,0007 | 0,0007 |
| **3,2** | 0,0007 | 0,0007 | 0,0006 | 0,0006 | 0,0006 | 0,0006 | 0,0006 | 0,0005 | 0,0005 | 0,0005 |
| **3,3** | 0,0005 | 0,0005 | 0,0005 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0004 | 0,0003 |
| **3,4** | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0003 | 0,0002 |
| **3,5** | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 | 0,0002 |

**Ejemplos** (se han sombreado):

- En una distribución normal, el 50% de los individuos está por encima de la media (≥0,00 desviaciones estándar por encima de la media; $z = 0,00$; columna 2, fila 1, $p = 0,500$).
- La proporción de individuos que está al menos a 1,15 desviaciones típicas bajo la media es del 12,51% ($z = -1,15$; columna 7, fila 12, $p = 0,1251$).

*Tabla inversa de la distribución normal tipificada*

Dentro de la tabla se proporciona el valor de $z_\alpha$ dependiendo del valor de $p$ (área bajo la curva), bien en una sola cola o en cada una de las dos colas. Esta es la tabla donde se deben buscar los valores de $z_\alpha$ para estimar intervalos de confianza o para calcular el tamaño muestral.

| $p$ (2 colas) $\alpha$ | $p$ (1 cola) $\alpha/2$ | $z$ | $p$ (2 colas) $\alpha$ | $p$ (1 cola) $\alpha/2$ | $z$ |
|---|---|---|---|---|---|
| 0,50 | 0,25 | 0,6745 | 0,24 | 0,12 | 1,1750 |
| 0,49 | 0,245 | 0,6903 | 0,23 | 0,115 | 1,2004 |
| 0,48 | 0,24 | 0,7063 | 0,22 | 0,11 | 1,2265 |
| 0,47 | 0,235 | 0,7225 | 0,21 | 0,105 | 1,2536 |
| 0,46 | 0,23 | 0,7388 | 0,20 | 0,10 | 1,2816 |
| 0,45 | 0,225 | 0,7554 | 0,19 | 0,095 | 1,3106 |
| 0,44 | 0,22 | 0,7722 | 0,18 | 0,09 | 1,3408 |
| 0,43 | 0,215 | 0,7892 | 0,17 | 0,085 | 1,3722 |
| 0,42 | 0,21 | 0,8064 | 0,16 | 0,08 | 1,4051 |
| 0,41 | 0,205 | 0,8239 | 0,15 | 0,075 | 1,4395 |
| 0,40 | 0,2 | 0,8416 | 0,14 | 0,07 | 1,4758 |
| 0,39 | 0,195 | 0,8596 | 0,13 | 0,065 | 1,5141 |
| 0,38 | 0,19 | 0,8779 | 0,12 | 0,06 | 1,5548 |
| 0,37 | 0,185 | 0,8965 | 0,11 | 0,055 | 1,5982 |
| 0,36 | 0,18 | 0,9154 | 0,10 | 0,05 | 1,6449 |
| 0,35 | 0,175 | 0,9346 | 0,09 | 0,045 | 1,6954 |
| 0,34 | 0,17 | 0,9542 | 0,08 | 0,04 | 1,7507 |
| 0,33 | 0,165 | 0,9741 | 0,07 | 0,035 | 1,8119 |
| 0,32 | 0,16 | 0,9945 | 0,06 | 0,03 | 1,8808 |
| 0,31 | 0,155 | 1,0152 | 0,05 | 0,025 | 1,9600 |
| 0,30 | 0,15 | 1,0364 | 0,04 | 0,02 | 2,0537 |
| 0,29 | 0,145 | 1,0581 | 0,03 | 0,015 | 2,1701 |
| 0,28 | 0,14 | 1,0803 | 0,02 | 0,01 | 2,3263 |
| 0,27 | 0,135 | 1,1031 | 0,01 | 0,005 | 2,5758 |
| 0,26 | 0,13 | 1,1264 | 0,001 | 0,0005 | 3,2905 |
| 0,25 | 0,125 | 1,1503 | 0,0001 | 0,00005 | 3,8906 |

**Ejemplos** (se han sombreado):

- En una distribución normal, hay una probabilidad = 0,2 de encontrar a un individuo a 0,8416 desviaciones estándar o más por encima de la media ($z$ = 0,8416) y una probabilidad = 0,4 de encontrar a alguien que se distancie 0,8416 desviaciones estándar o más de la media, sumando a los que están a ≥0,8416 desviaciones estándar por encima y a los que están a ≥0,8416 desviaciones estándar por debajo de la media (columna 3, fila 11).
- La proporción de individuos que está al menos a 1,96 desviaciones típicas por encima de la media es del 2,5%. Si se suman los que están por encima y por debajo de 1,96 desviaciones estándar, hay un 5% de individuos a una distancia superior o igual a 1,96 desviaciones estándar de la media ($z$ = 1,96; columna 6, fila 20).

Dentro de la tabla se presentan los valores de la *t*. Se deben tener en cuenta los grados de libertad y el error alfa, a una o dos colas.

| gl | $\alpha = 0{,}10$ | $\alpha/2 = 0{,}025$ $\alpha = 0{,}05$ | $\alpha/2 = 0{,}01$ $\alpha = 0{,}02$ | $\alpha/2 = 0{,}005$ $\alpha = 0{,}01$ | gl | $\alpha = 0{,}10$ | $\alpha/2 = 0{,}025$ $\alpha = 0{,}05$ | $\alpha/2 = 0{,}01$ $\alpha = 0{,}02$ | $\alpha/2 = 0{,}005$ $\alpha = 0{,}01$ |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 6,3137 | 12,7062 | 31,8210 | 63,6559 | **41** | 1,6829 | 2,0195 | 2,4208 | 2,7012 |
| **2** | 2,9200 | 4,3027 | 6,9645 | 9,9250 | **42** | 1,6820 | 2,0181 | 2,4185 | 2,6981 |
| **3** | 2,3534 | 3,1824 | 4,5407 | 5,8408 | **43** | 1,6811 | 2,0167 | 2,4163 | 2,6951 |
| **4** | 2,1318 | 2,7765 | 3,7469 | 4,6041 | **44** | 1,6802 | 2,0154 | 2,4141 | 2,6923 |
| **5** | 2,0150 | 2,5706 | 3,3649 | 4,0321 | **45** | 1,6794 | 2,0141 | 2,4121 | 2,6896 |
| **6** | 1,9432 | 2,4469 | 3,1427 | 3,7074 | **46** | 1,6787 | 2,0129 | 2,4102 | 2,6870 |
| **7** | 1,8946 | 2,3646 | 2,9979 | 3,4995 | **47** | 1,6779 | 2,0117 | 2,4083 | 2,6846 |
| **8** | 1,8595 | 2,3060 | 2,8965 | 3,3554 | **48** | 1,6772 | 2,0106 | 2,4066 | 2,6822 |
| **9** | 1,8331 | 2,2622 | 2,8214 | 3,2498 | **49** | 1,6766 | 2,0096 | 2,4049 | 2,6800 |
| **10** | 1,8125 | 2,2281 | 2,7638 | 3,1693 | **50** | 1,6759 | 2,0086 | 2,4033 | 2,6778 |
| **11** | 1,7959 | 2,2010 | 2,7181 | 3,1058 | **51** | 1,6753 | 2,0076 | 2,4017 | 2,6757 |
| **12** | 1,7823 | 2,1788 | 2,6810 | 3,0545 | **52** | 1,6747 | 2,0066 | 2,4002 | 2,6737 |
| **13** | 1,7709 | 2,1604 | 2,6503 | 3,0123 | **53** | 1,6741 | 2,0057 | 2,3988 | 2,6718 |
| **14** | 1,7613 | 2,1448 | 2,6245 | 2,9768 | **54** | 1,6736 | 2,0049 | 2,3974 | 2,6700 |
| **15** | 1,7531 | 2,1315 | 2,6025 | 2,9467 | **55** | 1,6730 | 2,0040 | 2,3961 | 2,6682 |
| **16** | 1,7459 | 2,1199 | 2,5835 | 2,9208 | **56** | 1,6725 | 2,0032 | 2,3948 | 2,6665 |
| **17** | 1,7396 | 2,1098 | 2,5669 | 2,8982 | **57** | 1,6720 | 2,0025 | 2,3936 | 2,6649 |
| **18** | 1,7341 | 2,1009 | 2,5524 | 2,8784 | **58** | 1,6716 | 2,0017 | 2,3924 | 2,6633 |
| **19** | 1,7291 | 2,0930 | 2,5395 | 2,8609 | **59** | 1,6711 | 2,0010 | 2,3912 | 2,6618 |
| **20** | 1,7247 | 2,0860 | 2,5280 | 2,8453 | **60** | 1,6706 | 2,0003 | 2,3901 | 2,6603 |
| **21** | 1,7207 | 2,0796 | 2,5176 | 2,8314 | **61** | 1,6702 | 1,9996 | 2,3890 | 2,6589 |
| **22** | 1,7171 | 2,0739 | 2,5083 | 2,8188 | **62** | 1,6698 | 1,9990 | 2,3880 | 2,6575 |
| **23** | 1,7139 | 2,0687 | 2,4999 | 2,8073 | **63** | 1,6694 | 1,9983 | 2,3870 | 2,6561 |
| **24** | 1,7109 | 2,0639 | 2,4922 | 2,7970 | **64** | 1,6690 | 1,9977 | 2,3860 | 2,6549 |
| **25** | 1,7081 | 2,0595 | 2,4851 | 2,7874 | **65** | 1,6686 | 1,9971 | 2,3851 | 2,6536 |
| **26** | 1,7056 | 2,0555 | 2,4786 | 2,7787 | **66** | 1,6683 | 1,9966 | 2,3842 | 2,6524 |
| **27** | 1,7033 | 2,0518 | 2,4727 | 2,7707 | **67** | 1,6679 | 1,9960 | 2,3833 | 2,6512 |
| **28** | 1,7011 | 2,0484 | 2,4671 | 2,7633 | **68** | 1,6676 | 1,9955 | 2,3824 | 2,6501 |
| **29** | 1,6991 | 2,0452 | 2,4620 | 2,7564 | **69** | 1,6672 | 1,9949 | 2,3816 | 2,6490 |
| **30** | 1,6973 | 2,0423 | 2,4573 | 2,7500 | **70** | 1,6669 | 1,9944 | 2,3808 | 2,6479 |
| **31** | 1,6955 | 2,0395 | 2,4528 | 2,7440 | **71** | 1,6666 | 1,9939 | 2,3800 | 2,6469 |
| **32** | 1,6939 | 2,0369 | 2,4487 | 2,7385 | **72** | 1,6663 | 1,9935 | 2,3793 | 2,6458 |
| **33** | 1,6924 | 2,0345 | 2,4448 | 2,7333 | **73** | 1,6660 | 1,9930 | 2,3785 | 2,6449 |
| **34** | 1,6909 | 2,0322 | 2,4411 | 2,7284 | **74** | 1,6657 | 1,9925 | 2,3778 | 2,6439 |
| **35** | 1,6896 | 2,0301 | 2,4377 | 2,7238 | **75** | 1,6654 | 1,9921 | 2,3771 | 2,6430 |
| **36** | 1,6883 | 2,0281 | 2,4345 | 2,7195 | **76** | 1,6652 | 1,9917 | 2,3764 | 2,6421 |
| **37** | 1,6871 | 2,0262 | 2,4314 | 2,7154 | **77** | 1,6649 | 1,9913 | 2,3758 | 2,6412 |
| **38** | 1,6860 | 2,0244 | 2,4286 | 2,7116 | **78** | 1,6646 | 1,9908 | 2,3751 | 2,6403 |
| **39** | 1,6849 | 2,0227 | 2,4258 | 2,7079 | **79** | 1,6644 | 1,9905 | 2,3745 | 2,6395 |
| **40** | 1,6839 | 2,0211 | 2,4233 | 2,7045 | **80** | 1,6641 | 1,9901 | 2,3739 | 2,6387 |

**Ejemplo** (se ha sombreado):

- En una distribución *t* de Student, por ejemplo, en una distribución muestral de medias, con muestras de tamaño 30 (29 grados de libertad), hay una probabilidad = 0,05 de encontrar la media de una muestra a 2,0452 errores estándar de la media poblacional, o más lejos aún, en cualquiera de las dos direcciones, por arriba o por abajo (es una probabilidad a dos colas). El 97,5% de las medias de las muestras de tamaño 30 estarán en el intervalo $-\infty$, +2,0452 errores estándar de la media poblacional (columna 3, fila 29).

*Tabla de la ji cuadrado ($\chi^2$)*

En **negrita**, a principio de cada casilla, se indican los grados de libertad. El error alfa corresponde al indicado en la primera columna (0,10; 0,05; 0,025; 0,01, etc.).

| **g.l.** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| 0,10 | 2,706 | 4,605 | 6,251 | 7,779 | 9,236 | 10,645 | 12,017 | 13,362 |
| 0,05 | 3,841 | 5,991 | 7,815 | 9,488 | 11,070 | 12,592 | 14,067 | 15,507 |
| 0,025 | 5,024 | 7,378 | 9,348 | 11,143 | 12,833 | 14,449 | 16,013 | 17,535 |
| 0,01 | 6,635 | 9,210 | 11,345 | 13,277 | 15,086 | 16,812 | 18,475 | 20,090 |
| 0,005 | 7,879 | 10,597 | 12,838 | 14,860 | 16,750 | 18,548 | 20,278 | 21,955 |
| 0,001 | 10,828 | 13,816 | 16,266 | 18,467 | 20,515 | 22,458 | 24,322 | 26,124 |
| **g.l.** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 0,10 | 14,684 | 15,987 | 17,275 | 18,549 | 19,812 | 21,064 | 22,307 | 23,542 |
| 0,05 | 16,919 | 18,307 | 19,675 | 21,026 | 22,362 | 23,685 | 24,996 | 26,296 |
| 0,025 | 19,023 | 20,483 | 21,920 | 23,337 | 24,736 | 26,119 | 27,488 | 28,845 |
| 0,01 | 21,666 | 23,209 | 24,725 | 26,217 | 27,688 | 29,141 | 30,578 | 32,000 |
| 0,005 | 23,589 | 25,188 | 26,757 | 28,300 | 29,819 | 31,319 | 32,801 | 34,267 |
| 0,001 | 27,877 | 29,588 | 31,264 | 32,909 | 34,528 | 36,123 | 37,697 | 39,252 |
| **g.l.** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** |
| 0,10 | 24,769 | 25,989 | 27,204 | 28,412 | 29,615 | 30,813 | 32,007 | 33,196 |
| 0,05 | 27,587 | 28,869 | 30,144 | 31,410 | 32,671 | 33,924 | 35,172 | 36,415 |
| 0,025 | 30,191 | 31,526 | 32,852 | **34,170** | 35,479 | 36,781 | 38,076 | 39,364 |
| 0,01 | 33,409 | 34,805 | 36,191 | 37,566 | 38,932 | 40,289 | 41,638 | 42,980 |
| 0,005 | 35,718 | 37,156 | 38,582 | 39,997 | 41,401 | 42,796 | 44,181 | 45,559 |
| 0,001 | 40,790 | 42,312 | 43,820 | 45,315 | 46,797 | 48,268 | 49,728 | 51,179 |
| **g.l.** | **25** | **26** | **27** | **28** | **29** | **30** | **40** | **50** |
| 0,10 | 34,382 | 35,563 | 36,741 | 37,916 | 39,087 | 40,256 | 51,805 | 63,167 |
| 0,05 | 37,652 | 38,885 | 40,113 | 41,337 | 42,557 | 43,773 | 55,758 | 67,505 |
| 0,025 | 40,646 | 41,923 | 43,195 | 44,461 | 45,722 | 46,979 | 59,342 | 71,420 |
| 0,01 | 44,314 | 45,642 | 46,963 | 48,278 | 49,588 | 50,892 | 63,691 | 76,154 |
| 0,005 | 46,928 | 48,290 | 49,645 | 50,993 | 52,336 | 53,672 | 66,766 | 79,490 |
| 0,001 | 52,620 | 54,052 | 55,476 | 56,892 | 58,301 | 59,703 | 73,402 | 86,661 |
| **g.l.** | **60** | **70** | **80** | **90** | **100** | | | |
| 0,10 | 74,397 | 85,527 | 96,578 | 107,565 | 118,498 | | | |
| 0,05 | 79,082 | 90,531 | 101,87 | 113,145 | 124,342 | | | |
| 0,025 | 83,298 | 95,023 | 106,629 | 118,136 | 129,561 | | | |
| 0,01 | 88,379 | 100,425 | 112,329 | 124,116 | 135,807 | | | |
| 0,005 | 91,952 | 104,215 | 116,321 | 128,299 | 140,169 | | | |
| 0,001 | 99,607 | 112,317 | 124,839 | 137,208 | 149,449 | | | |

**Ejemplo** (se ha sombreado):

- En una prueba de ji cuadrado, con 20 grados de libertad, cuyo valor hallado sea de 34,17, la probabilidad de encontrar un resultado tan alejado o más de la hipótesis nula es de 0,025 (valor $p$ = 0,025). Si se encontrase en la prueba un valor de ji cuadrado superior a 34,17, el valor $p$ sería inferior a 0,025.

*Tabla de la F de Snedecor* para un error alfa = 0,05*

| g.l. DENOMINADOR | g.l. EN EL NUMERADOR | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 161,446 | 199,499 | 215,707 | 224,583 | 230,160 | 233,988 | 236,767 | 238,884 | 240,543 | 241,882 |
| 2 | 18,513 | 19,000 | 19,164 | 19,247 | 19,296 | 19,329 | 19,353 | 19,371 | 19,385 | 19,396 |
| 3 | 10,128 | 9,552 | 9,277 | 9,117 | 9,013 | 8,941 | 8,887 | 8,845 | 8,812 | 8,785 |
| 4 | 7,709 | 6,944 | 6,591 | 6,388 | 6,256 | 6,163 | 6,094 | 6,041 | 5,999 | 5,964 |
| 5 | 6,608 | 5,786 | 5,409 | 5,192 | 5,050 | 4,950 | 4,876 | 4,818 | 4,772 | 4,735 |
| 6 | 5,987 | 5,143 | 4,757 | 4,534 | 4,387 | 4,284 | 4,207 | 4,147 | 4,099 | 4,060 |
| 7 | 5,591 | 4,737 | 4,347 | 4,120 | 3,972 | 3,866 | 3,787 | 3,726 | 3,677 | 3,637 |
| 8 | 5,318 | 4,459 | 4,066 | 3,838 | 3,688 | 3,581 | 3,500 | 3,438 | 3,388 | 3,347 |
| 9 | 5,117 | 4,256 | 3,863 | 3,633 | 3,482 | 3,374 | 3,293 | 3,230 | 3,179 | 3,137 |
| 10 | 4,965 | 4,103 | 3,708 | 3,478 | 3,326 | 3,217 | 3,135 | 3,072 | 3,020 | 2,978 |
| 11 | 4,844 | 3,982 | 3,587 | 3,357 | 3,204 | 3,095 | 3,012 | 2,948 | 2,896 | 2,854 |
| 12 | 4,747 | 3,885 | 3,490 | 3,259 | 3,106 | 2,996 | 2,913 | 2,849 | 2,796 | 2,753 |
| 13 | 4,667 | 3,806 | 3,411 | 3,179 | 3,025 | 2,915 | 2,832 | 2,767 | 2,714 | 2,671 |
| 14 | 4,600 | 3,739 | 3,344 | 3,112 | 2,958 | 2,848 | 2,764 | 2,699 | 2,646 | 2,602 |
| 15 | 4,543 | 3,682 | 3,287 | 3,056 | 2,901 | 2,790 | 2,707 | 2,641 | 2,588 | 2,544 |
| 16 | 4,494 | 3,634 | 3,239 | 3,007 | 2,852 | 2,741 | 2,657 | 2,591 | 2,538 | 2,494 |
| 17 | 4,451 | 3,592 | 3,197 | 2,965 | 2,810 | 2,699 | 2,614 | 2,548 | 2,494 | 2,450 |
| 18 | 4,414 | 3,555 | 3,160 | 2,928 | 2,773 | 2,661 | 2,577 | 2,510 | 2,456 | 2,412 |
| 19 | 4,381 | 3,522 | 3,127 | 2,895 | 2,740 | 2,628 | 2,544 | 2,477 | 2,423 | 2,378 |
| 20 | 4,351 | 3,493 | 3,098 | 2,866 | 2,711 | 2,599 | 2,514 | 2,447 | 2,393 | 2,348 |
| 21 | 4,325 | 3,467 | 3,072 | 2,840 | 2,685 | 2,573 | 2,488 | 2,420 | 2,366 | 2,321 |
| 22 | 4,301 | 3,443 | 3,049 | 2,817 | 2,661 | 2,549 | 2,464 | 2,397 | 2,342 | 2,297 |
| 23 | 4,279 | 3,422 | 3,028 | 2,796 | 2,640 | 2,528 | 2,442 | 2,375 | 2,320 | 2,275 |
| 24 | 4,260 | 3,403 | 3,009 | 2,776 | 2,621 | 2,508 | 2,423 | 2,355 | 2,300 | 2,255 |
| 25 | 4,242 | 3,385 | 2,991 | 2,759 | 2,603 | 2,490 | 2,405 | 2,337 | 2,282 | 2,236 |
| 26 | 4,225 | 3,369 | 2,975 | 2,743 | 2,587 | 2,474 | 2,388 | 2,321 | 2,265 | 2,220 |
| 27 | 4,210 | 3,354 | 2,960 | 2,728 | 2,572 | 2,459 | 2,373 | 2,305 | 2,250 | 2,204 |
| 28 | 4,196 | 3,340 | 2,947 | 2,714 | 2,558 | 2,445 | 2,359 | 2,291 | 2,236 | 2,190 |
| 29 | 4,183 | 3,328 | 2,934 | 2,701 | 2,545 | 2,432 | 2,346 | 2,278 | 2,223 | 2,177 |
| 30 | 4,171 | 3,316 | 2,922 | 2,690 | 2,534 | 2,421 | 2,334 | 2,266 | 2,211 | 2,165 |
| 31 | 4,160 | 3,305 | 2,911 | 2,679 | 2,523 | 2,409 | 2,323 | 2,255 | 2,199 | 2,153 |
| 32 | 4,149 | 3,295 | 2,901 | 2,668 | 2,512 | 2,399 | 2,313 | 2,244 | 2,189 | 2,142 |
| 33 | 4,139 | 3,285 | 2,892 | 2,659 | 2,503 | 2,389 | 2,303 | 2,235 | 2,179 | 2,133 |
| 34 | 4,130 | 3,276 | 2,883 | 2,650 | 2,494 | 2,380 | 2,294 | 2,225 | 2,170 | 2,123 |
| 35 | 4,121 | 3,267 | 2,874 | 2,641 | 2,485 | 2,372 | 2,285 | 2,217 | 2,161 | 2,114 |
| 36 | 4,113 | 3,259 | 2,866 | 2,634 | 2,477 | 2,364 | 2,277 | 2,209 | 2,153 | 2,106 |
| 37 | 4,105 | 3,252 | 2,859 | 2,626 | 2,470 | 2,356 | 2,270 | 2,201 | 2,145 | 2,098 |
| 38 | 4,098 | 3,245 | 2,852 | 2,619 | 2,463 | 2,349 | 2,262 | 2,194 | 2,138 | 2,091 |
| 39 | 4,091 | 3,238 | 2,845 | 2,612 | 2,456 | 2,342 | 2,255 | 2,187 | 2,131 | 2,084 |
| 40 | 4,085 | 3,232 | 2,839 | 2,606 | 2,449 | 2,336 | 2,249 | 2,180 | 2,124 | 2,077 |

**Ejemplo** (se ha sombreado):

- En una prueba F, con 3 grados de libertad en el numerador y 10 en el denominador, cuyo valor hallado sea de 3,708, la probabilidad de encontrar un resultado tan alejado o más de la hipótesis nula es de 0,05 (valor $p = 0,05$). Si se encontrase en la prueba un valor de F superior a 3,708, el valor $p$ sería inferior a 0,05.

*(Continúa)*

*Tabla de la F de Snedecor* para un error alfa = 0,05 (cont.)*

| g.l. DENOMINADOR | g.l. EN EL NUMERADOR | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 15 | 20 | 30 | 40 | 50 | 60 | 100 | 120 | 200 | • |
| 1 | 245,949 | 248,016 | 250,096 | 251,144 | 251,774 | 252,196 | 253,043 | 253,254 | 253,676 | 254,317 |
| 2 | 19,429 | 19,446 | 19,463 | 19,471 | 19,476 | 19,479 | 19,486 | 19,487 | 19,491 | 19,496 |
| 3 | 8,703 | 8,660 | 8,617 | 8,594 | 8,581 | 8,572 | 8,554 | 8,549 | 8,540 | 8,527 |
| 4 | 5,858 | 5,803 | 5,746 | 5,717 | 5,699 | 5,688 | 5,664 | 5,658 | 5,646 | 5,628 |
| 5 | 4,619 | 4,558 | 4,496 | 4,464 | 4,444 | 4,431 | 4,405 | 4,398 | 4,385 | 4,365 |
| 6 | 3,938 | 3,874 | 3,808 | 3,774 | 3,754 | 3,740 | 3,712 | 3,705 | 3,690 | 3,669 |
| 7 | 3,511 | 3,445 | 3,376 | 3,340 | 3,319 | 3,304 | 3,275 | 3,267 | 3,252 | 3,230 |
| 8 | 3,218 | 3,150 | 3,079 | 3,043 | 3,020 | 3,005 | 2,975 | 2,967 | 2,951 | 2,928 |
| 9 | 3,006 | 2,936 | 2,864 | 2,826 | 2,803 | 2,787 | 2,756 | 2,748 | 2,731 | 2,707 |
| 10 | 2,845 | 2,774 | 2,700 | 2,661 | 2,637 | 2,621 | 2,588 | 2,580 | 2,563 | 2,538 |
| 11 | 2,719 | 2,646 | 2,570 | 2,531 | 2,507 | 2,490 | 2,457 | 2,448 | 2,431 | 2,404 |
| 12 | 2,617 | 2,544 | 2,466 | 2,426 | 2,401 | 2,384 | 2,350 | 2,341 | 2,323 | 2,296 |
| 13 | 2,533 | 2,459 | 2,380 | 2,339 | 2,314 | 2,297 | 2,261 | 2,252 | 2,234 | 2,206 |
| 14 | 2,463 | 2,388 | 2,308 | 2,266 | 2,241 | 2,223 | 2,187 | 2,178 | 2,159 | 2,131 |
| 15 | 2,403 | 2,328 | 2,247 | 2,204 | 2,178 | 2,160 | 2,123 | 2,114 | 2,095 | 2,066 |
| 16 | 2,352 | 2,276 | 2,194 | 2,151 | 2,124 | 2,106 | 2,068 | 2,059 | 2,039 | 2,010 |
| 17 | 2,308 | 2,230 | 2,148 | 2,104 | 2,077 | 2,058 | 2,020 | 2,011 | 1,991 | 1,960 |
| 18 | 2,269 | 2,191 | 2,107 | 2,063 | 2,035 | 2,017 | 1,978 | 1,968 | 1,948 | 1,917 |
| 19 | 2,234 | 2,155 | 2,071 | 2,026 | 1,999 | 1,980 | 1,940 | 1,930 | 1,910 | 1,878 |
| 20 | 2,203 | 2,124 | 2,039 | 1,994 | 1,966 | 1,946 | 1,907 | 1,896 | 1,875 | 1,843 |
| 25 | 2,089 | 2,007 | 1,919 | 1,872 | 1,842 | 1,822 | 1,779 | 1,768 | 1,746 | 1,711 |
| 30 | 2,015 | 1,932 | 1,841 | 1,792 | 1,761 | 1,740 | 1,695 | 1,683 | 1,660 | 1,622 |
| •35 | 1,963 | 1,878 | 1,786 | 1,735 | 1,703 | 1,681 | 1,635 | 1,623 | 1,598 | 1,558 |
| 40 | 1,924 | 1,839 | 1,744 | 1,693 | 1,660 | 1,637 | 1,589 | 1,577 | 1,551 | 1,509 |
| 45 | 1,895 | 1,808 | 1,713 | 1,660 | 1,626 | 1,603 | 1,554 | 1,541 | 1,513 | 1,470 |
| 50 | 1,871 | 1,784 | 1,687 | 1,634 | 1,599 | 1,576 | 1,525 | 1,511 | 1,484 | 1,438 |
| 60 | 1,836 | 1,748 | 1,649 | 1,594 | 1,559 | 1,534 | 1,481 | 1,467 | 1,438 | 1,389 |
| 70 | 1,812 | 1,722 | 1,622 | 1,566 | 1,530 | 1,505 | 1,450 | 1,435 | 1,404 | 1,353 |
| 80 | 1,793 | 1,703 | 1,602 | 1,545 | 1,508 | 1,482 | 1,426 | 1,411 | 1,379 | 1,325 |
| 90 | 1,779 | 1,688 | 1,586 | 1,528 | 1,491 | 1,465 | 1,407 | 1,391 | 1,358 | 1,302 |
| 100 | 1,768 | 1,676 | 1,573 | 1,515 | 1,477 | 1,450 | 1,392 | 1,376 | 1,342 | 1,283 |
| 150 | 1,734 | 1,641 | 1,535 | 1,475 | 1,436 | 1,407 | 1,345 | 1,327 | 1,290 | 1,223 |
| 200 | 1,717 | 1,623 | 1,516 | 1,455 | 1,415 | 1,386 | 1,321 | 1,302 | 1,263 | 1,189 |
| • | 1,666 | 1,571 | 1,459 | 1,394 | 1,350 | 1,318 | 1,243 | 1,221 | 1,170 | 1,00 |

*Se le llama a veces F de Fisher, o F de Fisher-Snedecor.

Si se desea obtener valores *p* a partir de tests F con diversas combinaciones de grados de libertad en el numerador y grados de libertad en el denominador, puede recurrirse al programa Excel, introduciendo en cualquier casilla la siguiente función:

**=DISTR.F(valor de F;g.l.numerador;g.l.denominador)**

Por ejemplo = **DISTR.F(1,717;15;200)** devuelve 0,0499 (se ha sombreado en la tabla).

Si se desea saber los valores de F correspondientes a un determinado error alfa, debe introducirse en Excel, en cualquier casilla la siguiente función:

**=DISTR.F.INV(alfa;g.l.numerador;g.l.denominador)**

Por ejemplo = **DISTR.F.INV(0,05;15;200)** devuelve 1,717.

# BIOESTADÍSTICA AMIGABLE

3.ª edición

Miguel Ángel Martínez González
Almudena Sánchez-Villegas
Estefanía A. Toledo Atucha
Javier Faulin Fajardo

Con más de doce años de existencia, *Bioestadística amigable* se ha convertido en un manual clásico en la materia, que ha resultado extraordinariamente útil. Abarca la inmensa mayoría de los procedimientos bioestadísticos actualmente utilizados en los artículos publicados en revistas científicas de medicina, farmacia, odontología, enfermería, fisioterapia, nutrición, psicología, veterinaria y otras profesiones.

Se distingue de otras obras similares por su gran orientación práctica. Su mayor ventaja es que cada procedimiento está explicado con un ejemplo numérico resuelto «a mano» y al que se aplican a la vez varios paquetes estadísticos (Stata, SPSS y R/Splus). Se explican las instrucciones que deben darse al ordenador en cada caso y la interpretación de los resultados obtenidos.

En esta tercera edición, sin perder nunca el carácter «amigable» de las explicaciones, se han incorporado:

- Procedimientos avanzados que se están usando cada vez más en investigación sanitaria (metaanálisis, imputación múltiple, modelos flexibles de splines, ecuaciones de estimación generalizada, índices de propensión, modelos marginales estructurales, etc.).
- Métodos de análisis de supervivencia desarrollados en mayor profundidad.
- Múltiples mejoras en temas de análisis de concordancia, validez y pronóstico.
- Más explicaciones sobre el programa Stata (sin abandonar SPSS, que tenía un gran protagonismo en la primera y la segunda edición).

Está orientado a profesionales sanitarios y estudiantes de titulaciones relacionadas con ciencias de la vida, gracias a un enfoque comprensible que prescinde de tecnicismos matemáticos.

*Los Dres. Martínez González y Faulin Fajardo son conocidos catedráticos de Estadística. Los demás autores son investigadores que trabajan o han trabajado en el Departamento de Medicina Preventiva de la Universidad de Navarra.*

www.studentconsult.es

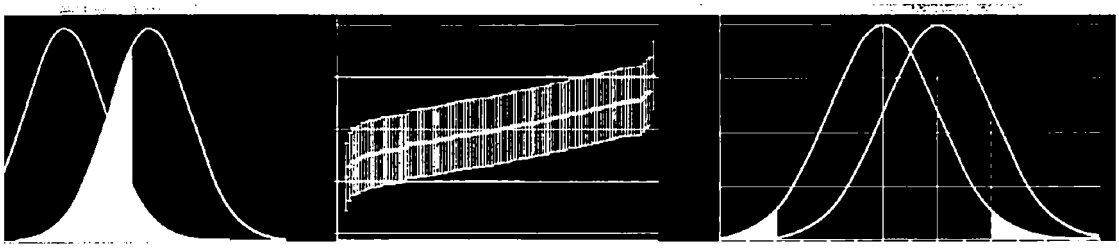
- Cuestiones y problemas resueltos

ISBN 978-84-9022-500-4


IER www.elsevier.es

3.ª edición

# BIOESTADÍSTICA AMIGABLE



Miguel Ángel Martínez González
Almudena Sánchez-Villegas
Estefanía A. Toledo Atucha
Javier Faulin Fajardo

ELSEVIER

CONTENIDO ADICIONAL en línea