Análisis de datos en psicología I

Juan Botella Ausina
Manuel Suero Suñe
Carmen Ximénez Gómez

PIRÁMIDE

Análisis de datos en psicología I

JUAN BOTELLA AUSINA
CATEDRÁTICO DE LA UNIVERSIDAD AUTÓNOMA DE MADRID

MANUEL SUERO SUÑE
PROFESOR TITULAR DE LA UNIVERSIDAD AUTÓNOMA DE MADRID

CARMEN XIMÉNEZ GÓMEZ
PROFESORA TITULAR DE LA UNIVERSIDAD AUTÓNOMA DE MADRID

Análisis de datos en psicología I

EDICIONES PIRÁMIDE

COLECCIÓN «PSICOLOGÍA»

Edición en versión digital

Para cualquier información pueden dirigirse a piramide_legal@anaya.es
Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
ISBN: 978-84-368-2656-2

A Tati y Mario.

J. B.

A mis padres Fernando y Ana María.

M. S.

A Raquel y Javi.

C. X.

Índice

Prefacio

Cuando se publica un libro sobre análisis de datos en psicología hay dos cosas que parece obligado hacer. La primera es justificar el por qué de otro libro más; la segunda es hacer explícitos los agradecimientos a aquellos que han contribuido a que acabe viendo la luz. Comenzaremos por lo primero, dejando los agradecimientos para el final de este prefacio.

El objetivo de este libro es servir de manual en la asignatura Análisis de Datos I, que impartimos en el Grado de Psicología de la Universidad Autónoma de Madrid, ya adaptada a las exigencias del Espacio Europeo de Educación Superior, más conocido como «Plan Bolonia». Su publicación se justifica por las modificaciones que se han ido introduciendo en el programa y por las actualizaciones de las estrategias didácticas que se derivan de nuestra experiencia docente.

Aunque pudiera parecer que la materia de una asignatura como ésta no es algo que cambie con los años, lo cierto es que sí lo hace. El índice tiene algunas variaciones respecto al anterior manual de la asignatura, publicado también por Pirámide (Botella, León, San Martín y Barriopedro, 2001). Éstas tienen que ver con el diferente énfasis que se quiere dar a los distintos procedimientos, pero también con la decisión de abordar una iniciación a la inferencia estadística. Ejemplo de lo primero son las distribuciones de frecuencias. En los libros de hace treinta años éstas recibían mucha atención, pues eran la base para muchos cálculos y para confeccionar representaciones gráficas. Ello obligaba a tratar también la problemática relacionada con la confección de intervalos. Hoy se trabaja con ordenadores y todo esto ha caído en desuso, por lo que su presencia en los manuales ha ido disminuyendo progresivamente.

Respecto a la estadística inferencial, debemos reconocer que las técnicas de contraste de hipótesis son difíciles de asimilar por muchos estudiantes, probablemente porque tienen que acostumbrarse a adoptar decisiones en entornos de incertidumbre. Trabajar estos conceptos al final de la asignatura de Análisis de Datos I, y continuar profundizando en ellos en Análisis de Datos II, es una buena estrategia pedagógica. Los últimos capítulos de este libro han sido ideados para ser empleados en esa primera etapa de asimilación de conceptos de la estadística inferencial.

Otros cambios tienen que ver con cuestiones estrictamente didácticas. Por ejemplo, las propiedades que implican los efectos de las transformaciones lineales

sobre estadísticos como la media y la varianza ya no son tratados en conexión con estos estadísticos, de forma separada. Se tratan de forma combinada en un capítulo dedicado a las transformaciones y que incluye las puntuaciones típicas, una forma de exponerlas que hemos encontrado útil en nuestras clases.

El libro es el resultado de la experiencia docente de los autores durante muchos años con esta difícil materia. De hecho, en bastantes puntos seguramente se notará que el libro está al servicio de la tarea docente que cada año afrontamos. La materia es difícil, pero agradecida. En pocas asignaturas es tan evidente la diferencia entre una buena estrategia didáctica y una inadecuada. Además, una de nuestras mayores satisfacciones es ver cómo algunos estudiantes a los que realmente les cuesta el razonamiento con números finalmente acaban comprendiendo la lógica que subyace en el pensamiento estadístico. A los estudiantes les ocurre lo mismo, pues su satisfacción es mayor cuanto mayor ha sido el esfuerzo realizado para superar la asignatura. Por todo esto hemos dedicado grandes esfuerzos a intentar facilitar el proceso de aprendizaje; muchos de los cambios que se introducen en este libro son el resultado de esos esfuerzos.

No hemos incluido exposiciones relacionadas con el apoyo informático que damos en el curso. La velocidad a la que aparecen las sucesivas versiones de los programas informáticos hace que los manuales que se apoyan en ellos queden pronto obsoletos. Por el contrario, son muy útiles las exposiciones específicas que se dirigen a los módulos básicos y que pueden ser actualizados a una velocidad diferente que los manuales. Tenemos la suerte de contar con la exposición que Ximénez y Revuelta (2011) hacen del uso de SPSS, que está especialmente pensada para estos mismos estudiantes y abarca todos los procedimientos que se incluyen en el programa completo de la asignatura. Nosotros lo utilizamos como complemento docente del presente manual.

Respecto a los agradecimientos, debemos mencionar en primer lugar a nuestros estudiantes, principales sufridores de nuestras limitaciones como docentes, pero a la vez fuente inagotable de ideas para mejorar nuestros recursos didácticos y para señalar los «puntos negros» que necesitan clarificaciones adicionales. También a nuestros compañeros del Departamento en la Universidad Autónoma de Madrid; aunque son muchos los que han contribuido indirectamente a mejorar nuestro trabajo, queremos mencionar explícitamente a los que en algún momento de los últimos años han colaborado con la docencia en la asignatura: Ludgerio Espinosa, Jesús Garrido, Beatriz Gil y Yolanda de Pellegrín.

Madrid, noviembre de 2011.

Juan Botella
Manuel Suero
Carmen Ximénez

1 Conceptos generales

1.1. INTRODUCCIÓN

La palabra «estadística» evoca las tablas de datos y gráficas que tan a menudo se encuentran en libros, en medios de comunicación o en Internet, y que se emplean para comunicar datos económicos, electorales, demográficos o de cualquier otro tipo. Aunque estos resúmenes de datos son, desde luego, el resultado de aplicar técnicas estadísticas, el campo de esta disciplina es bastante más amplio de lo que esos ejemplos sugieren. La estadística no sólo es un conjunto de técnicas para resumir y comunicar informaciones cuantitativas, sino que sirve también, y fundamentalmente, para hacer inferencias, generalizaciones y extrapolaciones de un conjunto relativamente pequeño de datos (observados) a un conjunto mayor (no observados). Una de las aplicaciones más importantes de estas técnicas es el propio trabajo de adquisición de conocimiento mediante la investigación científica, a la que ha proporcionado poderosos instrumentos para el análisis de datos y la toma de decisiones.

En el cuadro 1.1 se describen algunos ejemplos variados de aplicaciones de la estadística. En ellos se llega a un punto en el que es necesario trabajar con un conjunto de números, a veces muy grande, con el que describir aquello que estamos estudiando. Además, en algunos estudios también llega un punto en el que surge la necesidad, o el deseo, de extraer conclusiones, a partir de las observaciones hechas, acerca de los casos que no han participado en el estudio, o de observaciones potenciales que no se han hecho. La estadística proporciona los medios técnicos para realizar estas dos tareas.

Estas dos grandes funciones de la estadística (la organización y descripción de datos, por un lado, y la realización de inferencias por el otro) reflejan la propia historia del desarrollo de esta ciencia. La estadística actual es el producto del encuentro y mutua fecundación en el siglo XIX de dos ramas distintas del conocimiento, la antigua estadística y el cálculo de probabilidades. Etimológicamente, la palabra «estadística» procede de la palabra «estado». Ya en la antigüedad los romanos y los egipcios hicieron intentos por tener un conocimiento preciso del número de sus habitantes y de sus posesiones, es decir, por conocer el estado de sus naciones (de ahí la raíz del término). Para ello hacían censos y recogían datos que posteriormente tenían que resumir de una forma comprensiva para proporcionar informaciones útiles.

CUADRO 1.1

Ejemplos

EJEMPLO 1. La empresa BSX nos encarga un estudio sobre los perfiles de competencias de su plantilla de mandos intermedios, con vistas a futuras promociones. Les administramos varios cuestionarios que miden diversas competencias, entre ellas el grado en que demuestran «capacidad de liderazgo». Al terminar la corrección de esas pruebas contamos con un conjunto de números a partir de los cuales describiremos las competencias de los miembros de esa plantilla.

EJEMPLO 2. Se trata de un estudio sobre la eficacia de un programa de atención a los familiares que ejercen el papel de cuidadores de enfermos con dolor crónico. La abnegada dedicación de los familiares de un enfermo (con frecuencia el marido, la mujer, el padre o la madre) tiene para esos cuidadores un importante coste psicológico, en términos de estrés y otros efectos en su calidad de vida. Para estudiarlo reunimos a los familiares de 30 enfermos con dolor crónico que son atendidos en un determinado centro (centro *A*) y aceptan participar en el estudio. También pedimos a los colegas de otro centro en el que no hay un programa de este tipo que colaboren con nosotros en el estudio, como grupo de control (centro *B*); reúnen también a 30 familiares de otros enfermos con dolor crónico que aceptan participar en el estudio. Evaluamos el nivel de estrés de los 60, pero luego sólo los del centro *A* reciben las sesiones del programa destinado a controlar y reducir el estrés. Valoramos también para cada uno el nivel educativo (sin estudios, primarios, secundarios, universitarios) y la inteligencia, dado que pueden afectar al grado de comprensión tanto del programa como de la forma en que se aplica. Al final nos encontramos con un conjunto de puntuaciones de cada uno de los 60 familiares, a partir de las cuales deseamos extraer conclusiones acerca de la eficacia del programa y de los efectos moduladores que puedan tener el nivel educativo y la inteligencia sobre esa eficacia.

EJEMPLO 3. Se centra en un aspecto de la atención selectiva. Se presentan al participante tres letras en la pantalla del ordenador y debe responder lo más rápidamente que pueda a la letra central, presionando una tecla si es una vocal y otra tecla si se trata de una consonante. Es habitual emplear dos condiciones experimentales (a veces se añaden otras): en la condición de flancos compatibles, las letras que acompañan a la letra central (llamadas flancos) son de la misma categoría que la letra central (vocales si la central es vocal y consonantes en caso contrario); en la condición de flancos incompatibles los flancos pertenecen a la categoría contraria. El resultado habitual, muchas veces replicado, es que se tarda más en responder a la letra central si los flancos son de la categoría contraria (flancos incompatibles) que si son de la misma categoría (flancos compatibles). Este resultado se interpreta en el sentido de que los observadores no consiguen que su selección atencional se sustraiga completamente a los flancos y no consiguen ignorarlos del todo. Supongamos que queremos probar este efecto con números y las categorías pares/impares. Administramos a una persona 30 ensayos de cada condición y registramos los tiempos de respuesta. Si nos centramos en la condición de flancos incompatibles, dispondremos de 30 mediciones de tiempos. No en todos los ensayos el participante tarda lo mismo. Hay una cierta variabilidad en las distintas ejecuciones de la tarea. Para hacernos una idea global de cómo realiza la tarea nuestro voluntario en cada condición tenemos que trabajar con los 30 valores procedentes de cada una.

CUADRO 1.1 *(continuación)*

EJEMPLO 4. Es un sondeo de opinión acerca de lo que la gente pensaría si se adoptasen medidas restrictivas en el consumo de tabaco. La idea es anticiparse a las reacciones, de forma que interesa conocer la opinión general a partir de las preguntas realizadas en el sondeo. Sin duda, lo más preciso sería preguntarles a todos y cada uno de los españoles por su opinión, pero por razones económicas esto no sería sensato. Decidimos, en consecuencia, seleccionar un grupo de 2.500 españoles de todas las comunidades autónomas y edades, consultándoles acerca de su opinión (a favor/en contra) sobre esta cuestión.

Sin embargo, las conclusiones extraídas se agotaban en el propio conjunto de datos observados; el objetivo consistía en hacerse una idea clara de lo que había, lo cual se contaba y medía. Lo que posibilitó el cálculo de probabilidades fue el desarrollo de un conjunto de métodos para extrapolar las conclusiones a entidades no observadas. Es decir, proporcionó la base conceptual para hacer inferencias acerca de potenciales observaciones a partir de unas pocas observaciones reales. Estas técnicas tuvieron su fundamento en el desarrollo de la curva normal por Gauss, en su aplicación por Galton a los problemas de herencia, etc. Sin embargo, los auténticos padres de estas técnicas fueron Karl Pearson (1857-1936) y Ronald Fisher (1890-1962); sobre la historia de la estadística en psicología, véase Cowles (1989) y Walker (1975).

Clásicamente, la estadística se ha dividido en dos partes: la *estadística descriptiva* y la *estadística inferencial*. Estas dos partes reflejan, como ya hemos dicho, las dos grandes épocas de su historia, pero también pueden reflejar la profundidad de los análisis que se realizan o, incluso, las fases de un estudio, puesto que para hacer un estudio inferencial primero hay que hacer un estudio descriptivo de los datos. Es decir, un estudio descriptivo se agota en la descripción, mientras que uno inferencial comienza por la descripción y luego aborda la inferencia.

Mientras que la estadística descriptiva se puede abordar sin conocimientos técnicos previos más allá del álgebra elemental, para el estudio de la estadística inferencial es imprescindible adquirir unas nociones básicas de probabilidad. Por ello es frecuente encontrar que los libros de estadística aplicada están organizados, al menos, en esos tres bloques. En este libro nosotros nos ocupamos sobre todo de los dos primeros, aunque también incluimos una introducción a la inferencia estadística, cuyo desarrollo pleno se puede seguir en otras obras (Pardo y San Martín, 2010).

Proponemos la siguiente definición de la estadística.

La *estadística* es la disciplina que se ocupa de la ordenación y análisis de datos procedentes de muestras y de la realización de inferencias acerca de las poblaciones de las que proceden.

Se puede decir que el sentido vulgar del término «estadística» al que nos referíamos al comenzar esta sección corresponde más o menos a la estadística des-

criptiva. El otro conjunto de técnicas estadísticas, que se utilizan para extraer conclusiones de poblaciones a partir de la observación de unos pocos casos, son las que integran la estadística inferencial. Por tanto, un trabajo en el que se aplica la estadística podría clasificarse como exclusivamente descriptivo o como inferencial (puesto que, como ya hemos dicho, la inferencia incluye la descripción). Si, por ejemplo, nos interesa conocer la opinión de un grupo docente de la facultad acerca de una serie de cuestiones que afectan a la organización del aula, podemos utilizar una pequeña encuesta. Con los datos recogidos podremos calcular promedios, porcentajes, etc., y con estos resúmenes numéricos podremos transmitir la información contenida en esos datos brutos utilizando los formatos compactos y de gran calidad informativa que nos proporciona la estadística descriptiva. El estudio se agota en esos mismos datos, por lo que es un estudio descriptivo.

Si, por el contrario, queremos hacernos una idea de las opiniones de los estudiantes de la universidad sobre esas mismas cuestiones, no podremos preguntarles a todos. Probablemente utilizaríamos la estrategia de seleccionar un grupo de estudiantes, administrarles la encuesta y, a partir de sus resultados, hacernos una idea de cuál es el estado de opinión del conjunto de los estudiantes de la universidad. En este caso se trata de hacer inferencias acerca de toda una universidad a partir de los datos observados en una pequeña parte de sus estudiantes; es un estudio inferencial. De la misma forma, podemos decir que nuestro ejemplo sobre la capacidad de liderazgo es descriptivo, mientras que los ejemplos sobre el programa para reducir el estrés de cuidadores, el tiempo de respuesta en la tarea de los flancos y el sondeo de opinión son inferenciales. En el cuadro 1.4 se señala el carácter descriptivo o inferencial de los trabajos descritos en el cuadro 1.1.

Una última idea que merece la pena destacar en esta introducción es la distinción entre estadística teórica y estadística aplicada. La primera se dedica al estudio y desarrollo de métodos formalmente válidos (sobre todo para hacer inferencias, pero también con otros objetivos), mientras que la segunda se dedica a la aplicación de esos métodos y modelos de actuación a disciplinas concretas. Según Kruskall (1974): «...estadística aplicada, al menos en principio, es la aplicación documentada de métodos que han sido teóricamente investigados, es decir, el salto real después de estudiar la teoría del salto» (p. 390).

De los modelos y métodos que proporciona la estadística teórica y las técnicas concretas y usos que desarrolla la estadística aplicada, no todos son utilizados en la misma medida por las distintas ciencias. Por eso, para referirse al conjunto de técnicas más utilizadas en cada una a veces se utilizan nombres tales como bioestadística, psicoestadística o socioestadística. Algunos autores han propuesto para los contenidos de la estadística aplicada otros términos, entre los que destaca el de *análisis de datos* (Tukey, 1962), que da nombre a este libro.

1.2. CONCEPTOS GENERALES

Cualquier aplicación de la estadística se refiere a un conjunto de entidades, conocido como «población», aunque casi siempre se desarrolle utilizando sólo una parte de ese conjunto, conocida como «muestra»; proponemos las siguientes definiciones:

Se llama *población* estadística al conjunto de todos los elementos que cumplen una o varias características.

Se llama *muestra* a cualquier subconjunto de los elementos de una población.

A los elementos que componen una *población* se les denomina «entidades estadísticas o individuos». Pueden ser personas, animales, objetos o, simplemente, números. En nuestro ejemplo 1, sobre la capacidad de liderazgo, son las personas que integran la plantilla de mandos intermedios de la empresa BSX; en el ejemplo 2, de los cuidadores, son todos los potencialmente cuidadores de un enfermo; en el ejemplo 3, sobre la selección atencional, serían todas las realizaciones de la tarea, en las condiciones de flancos compatibles e incompatibles, que potencialmente podría realizar la persona que realiza nuestro experimento; en el ejemplo 4, la población del sondeo son todos los ciudadanos españoles mayores de edad.

Dependiendo del número de elementos que la compongan, la población puede ser finita o infinita. Los niños que estudian la ESO en la Comunidad de Madrid, los niños invidentes españoles, las empresas de nuevas tecnologías con sede en Tres Cantos o las poblaciones de nuestros ejemplos sobre la capacidad de liderazgo, el estrés de los cuidadores y el sondeo son casos de poblaciones finitas, puesto que en ellas los elementos se podrían contar, obteniendo un número finito. El número de lanzamientos posibles de un dado, el conjunto de los números pares o la población de nuestro ejemplo sobre tiempo de respuesta son casos de poblaciones infinitas, puesto que teóricamente no tienen un límite: por muchas observaciones que realicemos, siempre podríamos recoger más.

Muchas poblaciones con las que trabajamos son finitas, pero tan numerosas que, a la hora de hacer inferencias acerca de ellas, se pueden considerar infinitas a efectos prácticos (en este caso estarían la poblaciones de nuestros ejemplos sobre estrés de cuidadores y sobre sondeos de opinión). En la estadística hay procedimientos de cálculo que varían dependiendo de que la población sea finita o infinita, pero a medida que se va incrementando el tamaño de las poblaciones finitas el uso de uno u otro procedimiento resulta indiferente; proporcionan resultados cada vez más parecidos. En consecuencia, la mayor parte de las veces trabajaremos con poblaciones infinitas, ya sea porque lo son de verdad o porque su tamaño es tan grande que tomarlas por tales no afecta prácticamente a los resultados.

Cuando un investigador aborda un trabajo empírico, debe definir la población correspondiente. La población ha de ser el marco o conjunto de referencia sobre el cual se van a realizar las conclusiones e interpretaciones; éstas no pueden exceder ese marco.

El hecho de que las poblaciones sean en general muy numerosas hace que la descripción de sus propiedades sea inaccesible. De ahí que se trabaje fundamentalmente con muestras.

La *muestra* nos va a proporcionar unos datos que podemos ordenar, simplificar y describir. Pero uno de los objetivos es el de poder describir la población

de partida mediante lo que encontremos en la muestra. Siguiendo con nuestros ejemplos del cuadro 1.1, podemos decir que lo que nos interesa no es la eficacia del programa de reducción de estrés en los 30 familiares concretos que participan en nuestro estudio del ejemplo 2, ni la forma de responder en los 30 ensayos de tiempo de reacción aplicados, ni la opinión de los 2.500 encuestados acerca de la pregunta. Lo que nos interesa realmente es extraer conclusiones generales acerca de la eficacia general de la técnica, la forma general de responder en la tarea y la opinión de toda la población. Y para poder extraer esas conclusiones lo más importante es que las muestras de observaciones sean representativas. Veámoslo con otro ejemplo. Supongamos que queremos estudiar la estatura de los españoles; para ello nos situamos en una calle de nuestra ciudad y nos disponemos a preguntar a los primeros cien transeúntes que pasen por aquel punto. Si por una casualidad nos hemos situado cerca de un polideportivo donde se practica el baloncesto, al cual suelen dedicarse individuos altos, los datos que recogeremos no serán representativos. Si lo que intentamos es hacernos una idea de cuál puede ser la estatura media de los españoles a partir de la estatura media de los integrantes de esa muestra, nuestras conclusiones serán incorrectas.

Existe todo un campo de la metodología, llamado muestreo, dedicado a estudiar procedimientos de extracción de muestras que maximicen la representatividad de las mismas. Sólo un adecuado muestreo asegurará la representatividad de la muestra. Remitimos al lector interesado a obras específicas sobre muestreo (Azorín y Sánchez-Crespo, 1986; Clairin y Brion, 2001).

Habitualmente, uno de los objetivos de cualquier investigación será la de alcanzar conclusiones acerca de la población a partir de la información obtenida en la muestra. Pero ese objetivo sólo se alcanzará plenamente en la medida en que esa información se aproveche adecuadamente. Por ello, un primer objetivo de la estadística descriptiva consiste en conseguir resúmenes de los datos, con índices compactos y muy informativos.

Las poblaciones se pueden caracterizar mediante unas constantes denominadas «parámetros». Una de las tareas de la estadística es hacer conjeturas acerca de esas cantidades. Para ello se utilizan magnitudes análogas obtenidas en las muestras, que se denominan «estadísticos». Podemos establecer las siguientes definiciones:

> Un *parámetro* es una propiedad descriptiva de una población.
> Un *estadístico* es una propiedad descriptiva de una muestra.

Por ejemplo, el estrés medio de la población de cuidadores o el tiempo medio que invertiría nuestro participante en todas sus hipotéticas realizaciones de la tarea de los flancos son ejemplos de parámetros. Como estas cantidades son desconocidas, haremos conjeturas sobre ellas a partir de cantidades similares obtenidas en las muestras. Así, es casi seguro que el estrés medio de los 60 familiares de nuestro estudio antes de comenzar con el programa no es idéntico al de la población, pero si la muestra seleccionada es realmente representativa, probablemente

la media muestral (o estadístico) no difiera mucho de la media poblacional (o parámetro).

Los parámetros y estadísticos no sólo son medias, sino que pueden ser otro tipo de cantidades, como por ejemplo porcentajes. Ejemplo de ello es nuestro sondeo, en el que el porcentaje de individuos de la población con opinión favorable se considera un parámetro. Veamos otro ejemplo tomado de la psicología. Supongamos que un investigador está estudiando la eficacia de un método terapéutico para la intervención en trastornos de alimentación. Ante la imposibilidad material y la dificultad económica que supone utilizar para la experiencia a todas las personas con trastornos de alimentación, decide tomar a cien personas que acuden a su consulta a lo largo de un año. Esta muestra es representativa de la población de personas con trastorno de alimentación. Utiliza el método con cada paciente, y tras el seguimiento correspondiente observa que hay 60 que no reinciden. Esto significa que se ha rehabilitado el 60 por 100. El valor 60 es un estadístico. Si al cabo de algún tiempo desea replicar la experiencia y toma otra muestra de personas con un trastorno de alimentación y se recuperan 58, tendremos el mismo estadístico en otra muestra. Repitiendo sucesivamente la experiencia con muestras de cien pacientes con estos trastornos, se encontrará con distintos porcentajes. Ninguno de ellos puede considerarse con seguridad el verdadero porcentaje de los que se rehabilitarían en la población con el método en cuestión, pero cada uno de ellos se puede utilizar para hacer conjeturas acerca de ese verdadero porcentaje o parámetro.

En la práctica no será preciso estar repitiendo el estudio; bastará con obtener una única muestra y, a partir de ella, tratar de estimar el parámetro. Para ello es fundamental que la muestra sea representativa de la población y que el estadístico calculado reúna la información necesaria y suficiente para que, a partir de él, podamos decir algo acerca de la verdadera eficacia del tratamiento, el verdadero porcentaje de los que se recuperarán con ese nuevo método, es decir, el parámetro.

Los parámetros se suelen representar por letras griegas (μ, σ, π), mientras que los estadísticos se suelen simbolizar por letras latinas ($\bar{X}$, S, P). En la primera fase de una investigación se obtienen los estadísticos y en la segunda se utilizan los valores obtenidos para hacer inferencias acerca de los parámetros.

1.3. MEDICIÓN

Cuando estudiamos las entidades que conforman una población, nos interesamos por alguna de las propiedades de sus elementos; esas propiedades adoptan distintas variedades (Amón, 1993):

> Una *característica* es una propiedad o cualidad de un elemento.
> Una *modalidad* es cada una de las maneras en que se puede presentar una característica.

Si trabajamos con la población española, sus elementos tienen las características sexo (que adopta dos modalidades: varón y mujer), estado civil (soltero, casado, viudo...), estatura (cada una de las estaturas diferentes que adoptan los españoles), inteligencia (cada uno de los posibles valores diferentes que adoptarían los españoles en inteligencia, según el instrumento que utilicemos para evaluarla), etc.

Por supuesto, la psicología se centra en aquellas características que son propias de su objeto de estudio, como la inteligencia, la memoria, la personalidad, etc. En el primero de nuestros ejemplos nos interesábamos por un rasgo de personalidad, en el segundo por el nivel de estrés, en el tercero por el tiempo de respuesta en una tarea y en el cuarto por la opinión. Cada una de estas características puede mostrar distintas modalidades (en nuestros ejemplos, los grados en que se tiene capacidad de liderazgo, los niveles de estrés, las distintas cantidades de tiempo y las opiniones sobre una cuestión particular).

Las técnicas estadísticas no se aplican directamente a las modalidades observadas. Las modalidades se representan por números y la estadística se aplica a esos números. La medición no es otra cosa que el proceso de atribuir números a las modalidades de las características. Los números se asignan a las características siguiendo las reglas que se derivan de algún modelo de medición; del estudio de los modelos de medición se ocupa la «teoría de la medida».

Ya hemos visto que las características permiten calificar a los elementos. Algunos de ellos adoptan la misma modalidad de una característica, mientras que otros adoptan modalidades diferentes. De algunas características incluso podemos decir que unos individuos las exhiben en mayor medida que otros. Es decir, a partir de una característica se puede establecer un sistema relacional empírico (porque se refiere a entidades y relaciones reales). Igualmente, el sistema numérico está formado por un conjunto de entidades (números) y unas relaciones entre ellos; es decir, se trata de un sistema relacional numérico.

Asumiremos la siguiente definición de medición:

> Se llama *medición* de una característica a la conexión entre un sistema relacional empírico y un sistema relacional numérico, de tal forma que las relaciones entre las entidades se reflejen en las relaciones entre los números que los simbolizan.

Sólo si se consigue el objetivo implicado en esta definición, ocurrirá que de las relaciones entre los números se podrán hacer inferencias válidas acerca de las relaciones entre las entidades.

Esta conexión es lo que Stevens (1946) llamaba *schemapiric union*. Por ejemplo, las modalidades que adopta la variable estatura son tales que se podría decir que una determinada modalidad es una estatura superior a otra determinada modalidad. Pues bien, los números que se atribuyan a esas modalidades en el proceso de medición deben reflejar esa superioridad. Por el contrario, lo único que podemos decir al comparar las modalidades de dos individuos en la variable sexo es

si esas modalidades son la misma o no, pero nada respecto a magnitudes. Los números asignados a las modalidades del sexo deben reflejar simplemente ese hecho diferencial; de la comparación de los números no se podrá deducir conclusión alguna distinta a la de si esos individuos tienen o no el mismo sexo. Es habitual asignar los valores 0 y 1 a las modalidades de variables como el sexo, pero se trata de un mero etiquetado, y desde luego que un 1 no implica «más sexo» que un 0.

1.3.1. Las escalas de medida

Como ya hemos avanzado, la medición estudia las condiciones de construcción de representaciones numéricas. Los modelos desarrollados para la medición se llaman *escalas de medida*. Aunque no podemos entrar aquí en profundidad en el complejo campo de la medición (para una exposición más detallada véase Jáñez, 1989), vamos a exponer las características fundamentales del sistema de clasificación de escalas propuesto por Stevens (1946, 1975) y que es todavía la clasificación más utilizada: escalas nominales, ordinales, cuantitativas de intervalo y cuantitativas de razón. Esta clasificación se ilustra en el cuadro 1.2 con ejemplos de cada tipo de escala.

El científico se centra en aquellas características que considera relevantes para su trabajo de investigación. Aplica a esas características un esquema de clasificación, sin el cual no podría realizar su trabajo de registrar, ordenar y comunicar lo observado. En su forma más simple y primitiva, un esquema no es más que una regla que permite organizar las observaciones en clases de equivalencia, de manera que las observaciones que son incluidas en la misma clase son consideradas como cualitativamente iguales, y las que son incluidas en clases diferentes son consideradas como cualitativamente diferentes. Se utiliza una clase para cada una de las modalidades que adopta la característica que se está estudiando. Las clases han de ser mutuamente exclusivas y exhaustivas, es decir, cada observación es incluida en una clase y sólo en una. Supongamos que tenemos un conjunto de n elementos $\{e_1, e_2, ..., e_n\}$ que tienen una característica cuyo estudio nos interesa. Esa característica adopta un número k de modalidades distintas; representamos por $m(e_i)$ a la modalidad del elemento e_i. Asignamos números a los elementos en función de la modalidad que presentan en esa característica; representamos por $n(e_i)$ al número asignado al elemento e_i. Establecemos una regla de asignación de números a los objetos, de tal forma que se cumplan las siguientes condiciones:

$$\text{Si} \quad n(e_i) = n(e_j), \qquad \text{entonces} \quad m(e_i) = m(e_j)$$

$$\text{Si} \quad n(e_i) \neq n(e_j), \qquad \text{entonces} \quad m(e_i) \neq m(e_j)$$

Al sencillo tipo de medición que cumple estas condiciones se le llama *escalamiento cualitativo o nominal;* al conjunto de clases que la integran se le llama *escala nominal*.

En realidad, no haría falta asignar números a las clases formadas en una escala nominal. Puesto que los números asignados no se van a utilizar como tales, sino como simples códigos de identificación, se podrían utilizar otros símbolos, como letras, palabras, etc. Veamos algunos ejemplos de escalas nominales. El más sencillo y utilizado para diferenciar a las personas es el sexo. Podríamos tomar una muestra representativa y clasificar a sus elementos según esta característica, que adopta sólo dos modalidades, asignando el valor 1 a los varones y el valor 0 a las mujeres (o al revés, dado que son números arbitrarios). Tras realizar esa operación tendremos a los elementos de la muestra clasificados en dos clases de equivalencia, una por cada modalidad, que son mutuamente exclusivas (ninguno de los elementos es incluido simultáneamente en más de una clase) y exhaustivas (todos los elementos han sido asignados a alguna de las clases utilizadas). Ahora podemos emplear los números utilizados para realizar operaciones estadísticas como las que vamos a exponer en los próximos capítulos.

Otros ejemplos de características que se miden a nivel nominal son el estado civil (soltero, casado, viudo, etc.), la comunidad autónoma de nacimiento (Andalucía, Aragón, etc.), el tipo de sangre (*A*, *B*, *AB* o 0) o la asignatura preferida por los estudiantes de bachillerato (matemáticas, biología, etc.). Ejemplos de la psicología son los diagnósticos psicopatológicos (neurosis, psicosis, psicopatías, etcétera) o el patrón de apego que muestra un niño (seguro, resistente, evitativo). La clave de estas escalas de medida es que sólo reflejan la igualdad o desigualdad de los elementos en una característica, pero no de posibles ordenaciones, puesto que la característica a la que se refieren no se tiene en mayor o menor medida, sino que simplemente adopta formas cualitativamente distintas. No se puede decir que las mujeres tengan «más sexo» que los varones, que los «neuróticos» tengan más psicopatología que los psicópatas, que los castellanos tengan más comunidad de origen que los madrileños, ni tiene sentido extraer conclusiones comparativas de las asignaturas a partir de los códigos con los que se representan en los impresos de matrícula.

Debemos advertir que algunos autores consideran que las escalas nominales no son casos de auténtica medición, puesto que consideran que, para que haya medición, la característica con la que se está trabajando debe existir en alguna cantidad, y las modalidades que adopte no son más que las diferentes magnitudes con las que se presenta esa característica. En los tres tipos de escala que vamos a ver a continuación se da esta circunstancia, de forma que los números asignados a las diferentes modalidades no sólo van a reflejar similitudes o diferencias, sino que van a permitir concluir acerca de las magnitudes relativas con las que se presenta la característica.

Supongamos que contamos de nuevo con un conjunto de elementos (e_1, e_2, ..., e_n) que difieren en una característica que cada uno posee en una cierta cantidad [$c(e_1)$, $c(e_2)$, ..., $c(e_n)$]. De nuevo, el proceso de medición debe consistir en la aplicación de una regla de asignación de números a las diferentes cantidades, pero ahora de tal forma que los números asignados a los elementos [$n(e_1)$, $n(e_2)$, ..., $n(e_n)$] reflejen esos distintos grados en los que se presenta la característica. De esta forma, los números asignados sí que nos permitirán alcanzar conclusiones acerca

de las magnitudes. Sin embargo, a veces lo único que esos números nos permiten inferir son relaciones del tipo «mayor que» o «menor que». En concreto, a veces la escala con la que estamos trabajando cumple sólo, para todo par de elementos, e_i y e_j, las dos condiciones siguientes:

$$\text{Si} \quad n(e_i) = n(e_j), \qquad \text{entonces} \quad c(e_i) = c(e_j)$$

$$\text{Si} \quad n(e_i) > n(e_j), \qquad \text{entonces} \quad c(e_i) > c(e_j)$$

A las escalas de medida que cumplen estas características se les llama *escalas ordinales*; también se dice que se está haciendo una medición a nivel ordinal. Estas condiciones implican un paso más allá de lo que suponían las escalas nominales. Al igual que en estas últimas, si dos elementos comparten el mismo número podemos concluir que presentan la misma modalidad (en este caso tienen la misma cantidad de esa propiedad), pero de dos elementos a los que se han asignado números diferentes no sólo se puede decir que son diferentes en esa característica, sino que se pueden establecer relaciones del tipo «mayor que» o «menor que». Se puede decir cuál de esos elementos presenta una mayor magnitud en la característica. Dicho de otro modo, los elementos se pueden ordenar; de ahí el nombre de la escala.

Un ejemplo tradicionalmente utilizado para ilustrar este tipo de escalas es la medición de la dureza de los minerales. Supongamos que tomamos cuatro minerales (e_1, e_2, e_3 y e_4) y tratamos de rayar unos con otros, haciendo todas las combinaciones posibles. Cuando un mineral raya a otro se dice que el primero es más duro que el segundo. Dos minerales con distinta dureza no sólo son diferentes en esa característica, sino que se puede decir que la poseen en distinta magnitud. El proceso de medición, o asignación de números, debe ser tal que refleje esas distintas magnitudes. Supongamos que e_3 ha rayado a todos los demás, mientras que e_1 ha rayado a e_2 y e_4; por último, e_4 ha rayado a e_2. La ordenación de los objetos según su dureza sería la siguiente: e_3, e_1, e_4, e_2. Pues bien, en una escala ordinal los números asignados deben respetar esa ordenación. Por ejemplo, podríamos hacer la siguiente asignación: $n(e_3) = 4$, $n(e_1) = 3$, $n(e_4) = 2$ y $n(e_2) = 1$. La comunicación de esta información permite al receptor extraer conclusiones del tipo «el mineral 3 tiene una mayor dureza que el mineral 4» o «el mineral 2 tiene una menor dureza que el mineral 3».

En psicología son muchas las características cuya medición se considera de nivel ordinal, pues son muchos los casos en los que lo único que se puede decir es que un individuo es más extravertido que otro, que un niño es más hiperactivo que otro o que el aprendizaje es más rápido con el método *A* que con el método *B*. Igualmente, si tomamos las calificaciones como un índice de los conocimientos de un estudiante, entonces lo único que podemos decir de un estudiante con sobresaliente es que tiene mejores conocimientos que otro con notable, y éste que otro con aprobado.

Un ejemplo de escalas ordinales en ciencias sociales es el nivel educativo formal alcanzado. Supongamos que asignamos a los individuos los siguientes valores: 1, si no tiene estudios; 2, si ha completado los estudios primarios; 3, si ha

completado la enseñanza secundaria; 4, si ha terminado algún estudio universitario. Estos números se pueden utilizar para hacer inferencias del tipo «igual que» o «mayor que».

La limitación de las escalas ordinales es que, aunque nos informan de que un elemento representa la característica en cuestión en una mayor magnitud que otro elemento, no nos dice en cuánto más. Para poder alcanzar conclusiones más precisas, como la de en cuánto más presenta la característica un elemento sobre otro, hay que contar con una unidad de medida. Pero en ese caso ya estaríamos hablando de otros tipos de escala, que expondremos a continuación.

Supongamos ahora una escala en la que, además de las dos condiciones expresadas para las escalas ordinales se cumple una tercera, según la cual, para cualquier elemento e_i:

$$n(e_i) = a + b \cdot c(e_i) \qquad \text{(siendo} \quad b \neq 0)$$

A este tipo de escala se le llama *escala de intervalo.* La tercera condición añadida a las exigidas para una escala ordinal impone que el número asignado al elemento e_i, que representamos por $n(e_i)$, sea una función lineal (véase el capítulo 7) de la magnitud real que ese objeto presenta en la característica en cuestión. La clave de esta tercera condición (que supone una mejora sustancial con respecto a las escalas ordinales) es que se cuenta con una unidad de medida, sin importar que tanto esta unidad de medida como el origen de la escala sean arbitrarios. Lo cierto es que si se cumple esta tercera condición podemos alcanzar consecuencias acerca de la igualdad o desigualdad de diferencias. Es decir, que para todo cuarteto de elementos, e_i, e_j, e_k, e_l, se cumplen las siguientes condiciones:

$$\text{Si} \qquad n(e_i) - n(e_j) = n(e_k) - n(e_l)$$

$$\text{entonces} \quad c(e_i) - c(e_j) = c(e_k) - c(e_l)$$

y

$$\text{Si} \qquad n(e_i) - n(e_j) > n(e_k) - n(e_l)$$

$$\text{entonces} \quad c(e_i) - c(e_j) > c(e_k) - c(e_l)$$

Es decir, que si la diferencia entre los números asignados a dos elementos es igual a la diferencia entre los números asignados a otros dos, entonces también son iguales las diferencias de magnitudes entre estos dos pares. Igualmente, una mayor diferencia entre los números asignados implica una mayor diferencia entre las magnitudes representadas.

El ejemplo clásico de este tipo de escalas es el de las temperaturas. Para construir la escala centígrada se enfría el agua hasta la temperatura de congelación y se pone un cero en la altura que alcanza la columna de mercurio. Después se calienta el agua hasta el punto de ebullición, y donde se encuentre la altura de la columna de mercurio se marca un cien. Posteriormente se divide el espacio entre

esas dos marcas en cien partes iguales, a las que se llaman grados centígrados. Si utilizamos este instrumento para medir la temperatura de los objetos, podemos extraer conclusiones como las siguientes: «si los elementos e_1 y e_2 están a 25 y 20 grados, respectivamente, mientras que los elementos e_3 y e_4 están a 17 y 12 grados, respectivamente, podemos decir que la diferencia de temperaturas entre los objetos e_1 y e_2 es igual a la diferencia de temperaturas entre los elementos e_3 y e_4. Por otra parte, si los elementos e_5 y e_6 están a 10 y 7 grados, respectivamente, entonces podemos decir que la diferencia entre los dos primeros pares es mayor que la diferencia entre este último par. En las escalas ordinales no podíamos hacer este tipo de afirmaciones acerca de las diferencias de magnitudes.

La principal limitación de este tipo de escalas es que, aunque cuenta con una unidad de medida, no tiene un cero absoluto; el número cero no representa realmente la ausencia de esa característica. En el caso de la temperatura en grados centígrados es claro que el valor cero no significa temperatura nula, puesto que se pueden observar temperaturas inferiores (valores negativos).

Efectivamente, el hecho de no contar con un cero absoluto impide alcanzar conclusiones todavía más precisas. La superación de esta limitación nos llevará al cuarto tipo de escalas, llamadas *escalas de razón*. En ellas se sustituye la tercera condición de las escalas de intervalo por otra más restrictiva. En concreto, se requiere que para todo elemento, e_i:

$$n(e_i) = a \cdot c(e_i) \qquad (\text{siendo} \quad a > 0)$$

Esta tercera condición cumple la función de preservar el significado del valor cero, de forma que siempre represente la ausencia de esa característica. Por ejemplo, la medición de distancias se puede hacer con el sistema métrico decimal, en metros y centímetros, o con el sistema que se emplea en Estados Unidos, en yardas, pies y pulgadas; en ambos casos, cuando se dice que algo mide cero significa lo mismo. Esto no ocurría con la temperatura, en la que el cero de la escala Farenheit no se corresponde con el cero en la escala centígrada. La consecuencia fundamental de la presencia de un origen absoluto, en lugar de un origen arbitrario, es que, además de poder alcanzar conclusiones acerca de la igualdad o desigualdad de diferencias, también se puede hablar de la igualdad o desigualdad de razones. Así, para todo cuarteto de elementos, e_i, e_j, e_k, e_l, se cumplen las siguientes condiciones:

$$\text{Si} \quad \frac{n(e_i)}{n(e_j)} = \frac{n(e_k)}{n(e_l)}, \qquad \text{entonces} \quad \frac{c(e_i)}{c(e_j)} = \frac{c(e_k)}{c(e_l)}$$

y

$$\text{Si} \quad \frac{n(e_i)}{n(e_j)} > \frac{n(e_k)}{n(e_l)}, \qquad \text{entonces} \quad \frac{c(e_i)}{c(e_j)} > \frac{c(e_k)}{c(e_l)}$$

Dicho en palabras, si al medir distancias decimos que el elemento e_1 mide 10 y el elemento e_2 mide 5, entonces podemos decir que el elemento e_1 mide el doble que el elemento e_2, al igual que siempre que el cociente entre dos mediciones de

distancias nos dé igual a 2, significará que una mide el doble que la otra. Por otra parte, un mayor cociente entre dos mediciones siempre indicará una mayor razón de las magnitudes de los elementos, en este ejemplo distancias.

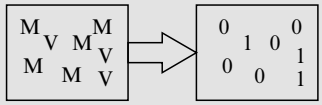
Los dos últimos tipos de escalas, las de intervalo y las de razón, reciben el nombre colectivo de escalas cuantitativas; en concreto, a veces se les llama *cuantitativas de intervalo* y *cuantitativas de razón*, respectivamente.

El cuadro 1.3 resume las características de cada tipo de escala y algunos ejemplos prototípicos. Igualmente, en el cuadro 1.4 se indica el tipo de escala de cada variable involucrada en los ejemplos del cuadro 1.1.

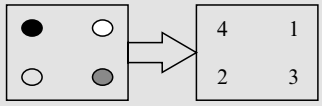
CUADRO 1.2

Las escalas de medida

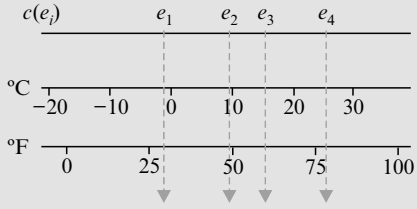
NOMINAL: El sexo de los individuos se clasifica simbolizando con un 0 «mujer» y con un 1 «varón».



ORDINAL: La dureza de los elementos se ordena, asignándoles números que representen esa ordenación. En este ejemplo, representamos la dureza con la oscuridad de su color; el más oscuro (negro) es el más duro, y el más claro (blanco) es el menos duro.

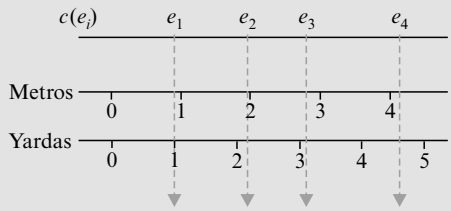


INTERVALO: Las cantidades de calor de los elementos, $c(e_i)$, se pueden representar (medir) por distintos conjuntos de números, siempre y cuando entre ellos se mantenga que la diferencia de temperatura entre, por ejemplo, los elementos 1 y 2, sea la misma que la diferencia entre los elementos 3 y 4, y que ambas diferencias sean mayores que la diferencia entre los objetos 2 y 3. Estas condiciones las cumplen tanto la escala centígrada como la escala Farenheit. Cada una tiene su propia unidad de medida y su propio origen (0).



CUADRO 1.2 *(continuación)*

RAZÓN: Las longitudes, $c(e_i)$, también se pueden representar por distintos conjuntos de números, siempre y cuando en ellos se mantenga que, por ejemplo, el elemento 2 tenga el doble que el elemento 1, y que el cociente entre los números asignados a los elementos 3 y 1 sea mayor que el cociente entre los números asignados a los elementos 2 y l. Estas condiciones se cumplen tanto al medir en metros como al medir en yardas. Aunque cada una tiene su unidad de medida, ambas respetan el cero absoluto, que coincide en las dos; el valor 0 representa la ausencia de esa característica.



CUADRO 1.3

Resumen de las escalas de medida

Tipo	Relaciones	Ejemplos
Nominal	Relaciones «igual que» o «distinto que»	Sexo, estado civil, diagnóstico clínico
Ordinal	Relaciones «mayor que» o «igual que»	Dureza, nivel socioeconómico, nivel educativo
Intervalo	Igualdad o desigualdad de diferencias	Temperatura, calendario, inteligencia
Razón	Igualdad o desigualdad de razones	Longitud, peso

1.3.2. Las variables: clasificación y notación

Ya hemos visto que en el proceso de medición se asignan números a los elementos de la muestra según unas reglas. Pues bien, el conjunto de valores numéricos atribuidos a las modalidades de una característica constituyen las variables estadísticas:

Una *variable* es una representación numérica de una característica.

El término variable refleja el hecho de que los valores atribuidos a las modalidades de una característica permiten diferenciar a los elementos, que varían entre sí en esa característica. En el caso de la psicología, ésta se centra en el estudio de las variables que le son propias y que constituyen su objeto de estudio. Por el contrario, a veces una característica tiene una única modalidad. Estrictamente hablando, no se trataría de una característica, ya que todas las entidades estudiadas mostrarían necesariamente el mismo valor numérico y se trataría más bien de una *constante*.

Las variables se pueden clasificar de varias formas. Ya hemos visto que, según el tipo de escala a la que pertenecen las variables, se pueden clasificar en nominales, ordinales, de intervalo y de razón. En el cuadro 1.3 aparecen algunos ejemplos de cada tipo.

A su vez, las variables cuantitativas (sean de intervalo o de razón) se pueden clasificar en discretas y continuas, en función del número de valores asumibles por ellas.

Una variable *discreta* es aquella que adopta valores aislados. Por tanto, fijados dos consecutivos, no puede adoptar ningún valor intermedio. Ejemplos de este tipo de variables son el número de hijos, el número de piezas dentales que conservan los internos de una residencia de ancianos, el número de libros leídos el pasado verano, el número de aciertos en un test, el número de asignaturas aprobadas el curso pasado o el número de accidentes de carretera durante el pasado fin de semana. Todas estas variables sólo pueden adoptar valores discretos.

En una variable *continua* ocurre que, entre dos valores cualesquiera, por próximos que sean, siempre es posible observar valores intermedios. Ejemplos de estas variables son la longitud, la duración de los sucesos o el peso. Por parecidas que sean las longitudes de dos objetos, mientras no midan lo mismo siempre es posible encontrar un objeto con una longitud intermedia entre ambos.

Aunque la distinción entre variables discretas y continuas es importante desde un punto de vista teórico, en la práctica las variables continuas no se pueden representar numéricamente como tales. Los instrumentos de medida son imprecisos y sólo permiten atribuir números discretos. La medición en la práctica supone una discretización artificial de las variables. Cuando decimos que un suceso ha durado 20 segundos, lo que queremos decir es que el número de segundos más cercano a su duración es 20; es decir, que su duración está en el intervalo $20 \pm 0{,}5$. En este ejemplo el valor 20 recibe el nombre de «valor informado». Los valores 19,5 y 20,5 se llaman límites exactos de la medida, y se obtienen sumando y restando al valor informado la mitad de la unidad de medida utilizada, que puede consistir en unidades, décimas, centésimas, etc.

No hay que confundir los valores discretos con los valores enteros, aunque en muchos casos coincidan, como en los ejemplos que hemos mencionado unas líneas más arriba. Por el contrario, un ejemplo en el que no se da esta coincidencia sería la proporción de aciertos en un test de cien preguntas. Si asignamos a cada individuo, como representación de sus conocimientos, el número que se obtiene al dividir el número de respuestas correctas por el número de preguntas del test, estos valores pueden ser 0, 0,01, 0,02, etc. Entre los valores 0,58 y 0,59, por ejemplo, no se puede observar ninguno intermedio.

Otro aspecto que suele ser fuente de confusiones es la idea de que las variables son características intrínsecas de los individuos u objetos, pero no siempre es así. A veces, las modalidades de las características son asignadas por el investigador. Por ejemplo, si queremos estudiar la supuesta ventaja del método de enseñanza A sobre el B podemos tomar dos grupos escolares y emplear un método en cada grupo. Una variable será, por supuesto, el aprendizaje mostrado, que nos indicará qué método es mejor, pero otra variable es en este caso el propio método de enseñanza aplicado. Sería una variable nominal con dos modalidades (A y B), pero el que a un individuo se le asocie una u otra dependerá de la asignación que haga el investigador, en lugar de ser algo propio del individuo.

CUADRO 1.4

Clasificación de los ejemplos del cuadro 1.1 y sus variables

Ejemplo	Tipo de estudio	Variables	Tipo de escala
1	Descriptivo	Capacidad de liderazgo	Intervalo
2	Inferencial	Grupo Nivel de estrés Nivel educativo Inteligencia	Nominal Intervalo Ordinal Intervalo
3	Inferencial	Condición de flancos Tiempo de respuesta	Nominal Razón
4	Inferencial	Opinión	Nominal

Las variables estadísticas se simbolizan por letras mayúsculas latinas, generalmente con un subíndice, para distinguirlas de las constantes (por ejemplo, U_i, V_i, X_i, Y_i). El subíndice i sirve, además, para indicar la posición que ocupa un determinado valor en el conjunto de valores de una variable. Por ejemplo, si la variable X_i (tiempo que tardan 4 personas en responder a una pregunta) adopta los valores 8; 5,2; 3,1 y 4,6, el símbolo X_1 representará al valor 8, el X_2 al valor 5,2, el X_3 al valor 3,1 y el X_4 al valor 4,6. El subíndice es un número que nada tiene que ver con la magnitud del valor al que se está refiriendo, sino simplemente al lugar que dicho valor ocupa dentro de una serie de valores. Se suele emplear el símbolo X_i para hacer referencia a los valores en general y X_N para el último valor de la serie. En el ejemplo mencionado, tendremos que $X_N = X_4 = 4{,}6$.

Es frecuente que una variable se mida en varios grupos, por lo que para simbolizar un valor cualquiera de dicha variable sería preciso utilizar dos subíndices, por ejemplo X_{ij}, donde la i se refiere a un orden de los valores y la j al grupo. Así, X_{23} simbolizará al segundo valor del grupo tercero.

Supongamos que para comprobar el efecto del consumo de una sustancia sobre la forma en que se percibe un estímulo, el investigador organiza tres grupos de individuos. Al primer grupo, formado por tres personas, le da un poco de agua;

al segundo grupo, con seis personas, le da un miligramo de la sustancia en estudio; al tercero, con cuatro personas, le da dos miligramos. Éstos han sido los resultados:

Grupo 1: 3,2; 5; 5,2

Grupo 2: 2,9; 4,5; 3,9; 4,7; 4,3; 3,7

Grupo 3: 8,9; 9; 9,3; 7,2

En tal caso, $X_{12} = 2{,}9$; $X_{22} = 4{,}5$; $X_{43} = 7{,}2$. El símbolo X_{41} no se refiere a ningún valor concreto de los observados, dado que en el grupo 1 sólo hay tres valores.

Con los valores observados haremos con frecuencia operaciones aritméticas para obtener estadísticos o para transformar esos valores. Una de las operaciones más frecuentes es la de sumar los valores, puesto que en ello se basan muchos estadísticos que iremos exponiendo en los próximos capítulos. También la operación de sumar se representa con un símbolo específico, el signo del sumatorio, del que exponemos algunas relaciones básicas en el apéndice del presente capítulo. El lector no familiarizado con su uso debería estudiar dicho apéndice antes de continuar, dado que en los próximos capítulos emplearemos con frecuencia las reglas del sumatorio.

PROBLEMAS Y EJERCICIOS

1. Diga a qué tipo de escala de medida pertenecen las siguientes variables:

a) Lugar de la residencia de verano.
b) Grado de aceptación (totalmente de acuerdo; de acuerdo; en desacuerdo; totalmente en desacuerdo) de la regulación de la muerte asistida.
c) Diagnóstico clínico.
d) Numero de estímulos visuales detectados.
e) Inteligencia medida con el Test de Inteligencia WAIS III para adultos.
f) Distancia tolerada a un estímulo fóbico.
g) Nivel de adecuación (adecuado, neutro, inadecuado) de una conducta en un estadio de fútbol.
h) Orientación terapéutica seguida por un psicólogo.
i) Gravedad del diagnóstico clínico.
j) Años de ejercicio profesional de un psicólogo clínico.
k) Sexo del terapeuta.

2. Se está diseñando una investigación acerca del efecto que tiene el cambio de residencia en los adolescentes varones españoles sobre el rendimiento académico y la socialización al final de su primer año de estancia. ¿Cuáles de las siguientes características serían constantes y cuáles variables?:

a) Rendimiento académico.
b) Nacionalidad.
c) Sexo.
d) Tiempo total en la nueva residencia.
e) Edad.
f) Socialización.

3. Atendiendo a la clasificación de variables expuesta en este tema, diga a qué tipo pertenecen las siguientes:

a) Número de experimentos en los que ha participado un estudiante en un año académico.
b) Tiempo que se tarda en responder a un ensayo de discriminación de tonos de estímulos auditivos.
c) Proporción de menores infractores con respecto a la población de menores.
d) Adscripción ideológica.
e) Selección de un color (nada adecuado; adecuado; totalmente adecuado) asociado a una emoción.
f) Número de elementos recordados de una lista de palabras.
g) Ingresos familiares.
h) Estación del año en la que se prefiere tomar las vacaciones.
i) Precisión en la bisección de una recta.

4. Se está realizando una investigación que tiene como objetivo estudiar el efecto del entrenamiento previo en una tarea de recuerdo de una lista de palabras. Se seleccionan dos grupos de participantes. Al grupo 1 se le aplica el entrenamiento y después la tarea de recuerdo, mientras que el grupo 2 realiza la tarea sin entrenamiento previo. Para cada participante se obtiene el número de palabras recordadas. Los resultados fueron:

	Grupo 1	Grupo 2
Participante	**Número de palabras**	**Número de palabras**
1	9	7
2	8	5
3	7	6
4	6	5
5	9	4

Responda a las siguientes cuestiones:

a) ¿Cuántas variables hay en el experimento?
b) Atendiendo a la clasificación de variables, diga a qué tipo pertenecen las variables implicadas en el experimento.
c) Utilizando la notación adecuada, indique el número de palabras recordadas por el tercer participante del segundo grupo.
d) Diga cuáles son los valores de X_{12} y X_{21}.

5. Se han medido en una muestra de estudiantes de psicología las variables *extraversión* (X) y *apertura a nuevas experiencias* (Y). Atendiendo a los resultados obtenidos, calcule las expresiones que aparecen más abajo.

X	*Y*
5	7
3	2
6	5
8	10
4	6

a) ΣX_i
b) ΣY_i
c) $\Sigma(X_i \cdot Y_i)$
d) $\Sigma X_i \cdot \Sigma Y_i$
e) $\Sigma(4 \cdot X_i)$
f) $\Sigma\left(\frac{Y_i}{5}\right)$
g) $\Sigma(X_i - 2)$
h) $\Sigma(X_i + Y_i)$
i) $(\Sigma Y_i)^2$
j) ΣY_i^2
k) $\Sigma\left(\frac{X_i}{3} - 2 \cdot Y_i + 10\right)$

6. A partir de los valores obtenidos para cinco sujetos en las variables T, X e V, calcule las expresiones que se indican más abajo.

T	*X*	*V*
4	5	2
9	2	4
7	9	7
5	7	5
1	2	1

a) $\sum V_i$
b) $\sum 2 \cdot X_i$
c) $\sum (T_i + 3)$
d) $\sum T_i \cdot \sum V_i \cdot \sum X_i$
e) $\sum (V_i \cdot X_i)$
f) $(\sum T_i)^2$ y $\sum T_i^2$
g) $\sum (T_i + X_i + V_i)$
h) $\sum (3 \cdot X_i - 4 \cdot V_i)$
i) $\sum (V_i - 5)^2$
j) $\sum T_i + 3$
k) $\dfrac{\sum X_i^2}{5} - \sum \left(\dfrac{\sum X_i}{5} \right)^2$

7. Atendiendo al sumatorio, escriba de forma abreviada las siguientes expresiones:

a) $X_1 + X_2 + X_3 + X_4 + X_5$
b) $X_1 + X_2 + X_3 + ... + X_N$
c) $7 \cdot X_1 + 7 \cdot X_2 + 7 \cdot X_3 + ... + 7 \cdot X_N$
d) $X_1 + X_2 + X_3 + ... + X_N + (4 \cdot N)$
e) $X_1 + X_2 + X_3 + X_4 + X_5 - (k \cdot N)$
f) $(X_1 + X_2 + X_3 + ... + X_N)^2$
g) $X_1^2 + X_2^2 + X_3^2 + ... + X_N^2$
h) $(X_1 - Y_1) + (X_2 - Y_2) + (X_3 - Y_3) + ... + (X_N - Y_N)$
i) $(X_1 + X_2 + X_3 + ... + X_N) \cdot (Y_1 + Y_2 + Y_3 + ... + Y_N)$
j) $(X_1 \cdot Y_1) + (X_2 \cdot Y_2) + (X_3 \cdot Y_3) + ... + (X_N \cdot Y_N)$
k) $(0{,}5 \cdot X_1 + Y_1 - 3 \cdot U_1) + (0{,}5 \cdot X_2 + Y_2 - 3 \cdot U_2) + ... + (0{,}5 \cdot X_N + Y_N - 3 \cdot U_N)$
l) $\dfrac{X_1}{N} + \dfrac{X_2}{N} + \dfrac{X_3}{N} + ... + \dfrac{X_N}{N}$

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*) Nominal.
b) Ordinal.
c) Nominal o cualitativa.
d) De razón.

e) De intervalo.
f) De razón.
g) Ordinal.
h) Nominal.
i) Ordinal.
j) De razón.
k) Nominal.

2. *a*) Variable.
b) Constante.
c) Constante.
d) Constante.
e) Variable.
f) Variable.

3. *a*) Cuantitativa discreta.
b) Cuantitativa continua.
c) Cuantitativa discreta.
d) Nominal o cualitativa.
e) Ordinal.
f) Cuantitativa discreta.
g) Cuantitativa discreta.
h) Nominal o cualitativa.
i) Cuantitativa continua.

4. *a*) Dos: entrenamiento previo con dos niveles (ausencia, presencia) y el número de palabras recordadas.
b) Ausencia − Presencia de entrenamiento previo: variable nominal. Número de palabras recordadas: cuantitativa discreta.
c) $X_{32} = 6$.
d) $X_{12} = 7$; $X_{21} = 8$.

5. *a*) $\sum X_i = 26$.
b) $\sum Y_i = 30$.
c) $\sum(X_i \cdot Y_i) = 175$.
d) $\sum X_i \cdot \sum Y_i = 26 \cdot 30 = 780$.
e) $\sum(4 \cdot X_i) = 4 \cdot \sum X_i = 4 \cdot 26 = 104$.
f) $\sum\left(\frac{Y_i}{5}\right) = \frac{1}{5} \cdot \sum Y_i = \frac{1}{5} \cdot 30 = 6$.
g) $\sum(X_i - 2) = \sum X_i - \sum 2 = 26 - (5 \cdot 2) = 16$.
h) $\sum(X_i + Y_i) = \sum X_i + \sum Y_i = 26 + 30 = 56$.
i) $(\sum Y_i)^2 = 30^2 = 900$.

j) $\sum Y_i^2 = 214.$

k) $\sum\left(\frac{X_i}{3} - 2 \cdot Y_i + 10\right) = \frac{1}{3} \cdot \sum X_i - 2 \cdot \sum Y_i + \sum 10 = \frac{1}{3} \cdot 26 - 2 \cdot 30 + 5 \cdot 10 = -1{,}333.$

6. *a*) $\sum V_i = 19.$

b) $\sum 2 \cdot X_i = 50.$

c) $\sum(T_i + 3) = 41.$

d) $\sum T_i \cdot \sum V_i \cdot \sum X_i = 12.350.$

e) $\sum(V_i \cdot X_i) = 118.$

f) $(\sum T_i)^2 = 676; \quad \sum T_i^2 = 172.$

g) $\sum(T_i + X_i + V_i) = 70.$

h) $\sum(3 \cdot X_i - 4 \cdot V_i) = -1.$

i) $\sum(V_i - 5)^2 = 30.$

j) $\sum T_i + 3 = 29.$

k) $\frac{\sum X_i^2}{5} - \sum\left(\frac{\sum X_i}{5}\right)^2 = 7{,}6.$

7. *a*) $\sum_{i=1}^{5} X_i.$

b) $\sum X_i$

c) $\sum(7 \cdot X_i).$

d) $\sum(X_i + 4).$

e) $\sum(X_i - k).$

f) $(\sum X_i)^2.$

g) $\sum X_i^2.$

h) $\sum(X_i - Y_i).$

i) $\sum X_i \cdot \sum Y_i.$

j) $\sum(X_i \cdot Y_i).$

k) $\sum(0{,}5 \cdot X_i + Y_i - 3 \cdot U_i).$

l) $\frac{\sum X_i}{N}.$

APÉNDICE

El sumatorio

La operación que con más frecuencia se hace en estadística es sumar un conjunto de valores. Por ello se ha acordado la utilización de un signo especial para representarla en las fórmulas y demostraciones: Σ. Así, cuando se quiere expresar la suma de las N puntuaciones X_1, X_2, ..., X_N, se puede hacer de una forma muy compacta y sencilla de la siguiente forma:

$$\sum_{i=1}^{N} X_i$$

La anterior expresión se lee «sumatorio de las X_i», comenzando por el primer valor y terminando por el enésimo. A veces no se quiere sumar las N puntuaciones, sino sólo parte de ellas. En esos casos se especifica en los índices superior e inferior cuáles son los valores por los que hay que comenzar y terminar la suma. Sin embargo, como en la casi totalidad de las ocasiones el signo se refiere a la suma de todos los valores, se suelen omitir esos indicadores. A lo largo de este libro utilizaremos la expresión:

$$\Sigma X_i$$

para referirnos a la suma de los N valores observados en X.

El trabajo con el sumatorio se simplifica bastante si se aplican algunas sencillas reglas, de las que vamos a exponer las más utilizadas:

a) *Regla del producto de la constante.* Si los valores de una variable se multiplican por una constante, el sumatorio de tales valores queda multiplicado por dicha constante. Dicho de otra forma, cuando nos encontremos con el sumatorio de una constante por una variable, podemos sacar la constante del sumatorio:

$$\Sigma(c \cdot X_i) = c \cdot X_1 + c \cdot X_2 + ... + c \cdot X_N =$$
$$= c \cdot (X_1 + X_2 + ... + X_N) = c \cdot \Sigma X_i$$

b) *Regla del sumatorio de la constante.* El sumatorio de una constante es igual a N multiplicado por la constante. Es decir:

$$\Sigma c = c + c + ... + c \ (N \text{ veces}) = N \cdot c$$

c) *Regla de distribución del sumatorio.* El sumatorio de una suma es igual a la suma de los sumatorios. Es decir:

$$\Sigma(V_i + X_i + Y_i) = (V_1 + X_1 + Y_1) + (V_2 + X_2 + Y_2) + ... +$$
$$+ (V_N + X_N + Y_N) = (V_1 + V_2 + ... + V_N) + (X_1 + X_2 + ... + X_N) +$$
$$+ (Y_1 + Y_2 + ... + Y_N) = \Sigma V_i + \Sigma X_i + \Sigma Y_i$$

En los próximos capítulos estas sencillas reglas nos serán de gran ayuda para las demostraciones. Así, aplicando las reglas *b*) y *c*), cuando aparezca el sumatorio de una variable más una constante pondremos:

$$\sum(c + X_i) = \sum c + \sum X_i = N \cdot c + \sum X_i$$

Cuando se trate del sumatorio de un binomio al cuadrado, con una constante y una variable, haremos el siguiente desarrollo:

$$\sum(c + X_i)^2 = \sum(c^2 + X_i^2 + 2 \cdot c \cdot X_i) = N \cdot c^2 + \sum X_i^2 + 2 \cdot c \cdot \sum X_i$$

Mientras que si se trata de dos variables, tendremos:

$$\sum(X_i + Y_i)^2 = \sum(X_i^2 + Y_i^2 + 2 \cdot X_i \cdot Y_i) = \sum X_i^2 + \sum Y_i^2 + 2 \cdot \sum(X_i \cdot Y_i)$$

La última expresión de esta fórmula es un caso especial, en el que cada sumando es un producto. En concreto:

$$\sum(X_i \cdot Y_i) = X_1 \cdot Y_1 + X_2 \cdot Y_2 + ... + X_N \cdot Y_N$$

Queremos advertir a aquellos lectores que no estén familiarizados con el uso de los sumatorios que hay algunos errores que típicamente se cometen cuando se comienza. Resaltaremos algunos, precisamente para estar atentos a no incurrir en ellos:

$$\sum X^2 \neq (\sum X_i)^2$$

$$\sum(X_i \cdot Y_i) \neq \sum X_i \cdot \sum Y_i$$

En el primero de estos errores destacamos que la suma de unos valores elevados al cuadrado no es igual al cuadrado de su suma:

$$X_1^2 + X_2^2 + ... + X_N^2 \neq (X_1 + X_2 + ... + X_N)^2$$

mientras que en el segundo señalamos que la suma de unos productos no es igual al producto de esas sumas:

$$X_1 \cdot Y_1 + X_2 \cdot Y_2 + ... + X_N \cdot Y_N \neq (X_1 + X_2 + ... + X_N) \cdot (Y_1 + Y_2 + ... + Y_N)$$

En algunas ocasiones, el sumatorio se referirá a unos valores organizados en función de dos subíndices, como por ejemplo el referido al grupo y el referido a los individuos. En esos casos se empleará el doble signo de sumatorio para hacer referencia a la suma de todos los casos de todos los grupos. El primer subíndice

se referirá al primer sumatorio (el más externo) y el segundo subíndice al segundo sumatorio (el más interno). Así, cuando queremos indicar la suma de los valores de los *i* individuos que hay dentro de cada uno de los *j* grupos, lo haremos de la siguiente manera:

$$\sum_{j=1}^{J}\sum_{i=1}^{I} X_{ij} = (X_{11} + X_{21} + \ldots + X_{I1}) + (X_{12} + X_{22} + \ldots + X_{I2}) + \ldots + \\ + (X_{1J} + X_{2J} + \ldots + X_{IJ})$$

PARTE PRIMERA

Estadística descriptiva con una variable

Organización y representación de datos. Medidas de posición

2

2.1. INTRODUCCIÓN

Tras la exposición de conceptos y definiciones del capítulo anterior, podemos imaginarnos en la situación de haber recogido un conjunto de valores tomados en una o varias variables. Nos disponemos a extraer, a partir de ellos, conclusiones relacionadas con los objetivos de la investigación. El primer paso debe ser siempre hacer una inspección cuidadosa de los datos. A veces esta primera inspección nos aporta ya alguna información sobre circunstancias llamativas interesantes. Si la muestra es pequeña, la simple inspección visual de dichos datos puede darnos una idea cabal de lo observado. Sin embargo, cuando la muestra es demasiado grande, que es el caso más frecuente, es difícil que una simple inspección pueda ser suficientemente comprensiva. Por eso, el primer paso suele consistir en organizar los datos utilizando un formato más inteligible que la simple yuxtaposición de números. Un instrumento para conseguir esa ordenación es la denominada *distribución de frecuencias,* y a partir de ella es habitual construir también *representaciones gráficas*. En este capítulo comenzaremos por describir estos dos instrumentos (en el apéndice de este capítulo se describe otro instrumento para organizar los datos, llamado *diagrama de tallo y hojas*). Además, expondremos las llamadas *medidas de posición*, que permiten valorar en términos relativos las puntuaciones individuales.

2.2. DISTRIBUCIÓN DE FRECUENCIAS

La distribución de frecuencias es un instrumento estadístico al que tradicionalmente se le han atribuido tres funciones: *a*) proporcionar una reorganización y ordenación de los datos; *b*) ofrecer la información necesaria para confeccionar representaciones gráficas, y *c*) facilitar los cálculos necesarios para obtener los estadísticos muestrales. Sin embargo, esta última función ha quedado obsoleta desde que los cálculos se realizan mediante ordenadores, que permiten que dichos cálculos sean mucho más rápidos y precisos. Vamos a definir algunos elementos que aparecen en una distribución de frecuencias, utilizando la terminología siguiente: representaremos por X a la variable con la que trabajamos, y que puede

adoptar distintos valores, X_1, X_2, X_3, ..., X_N; pero cada uno de esos valores puede aparecer más de una vez en los N elementos que componen la muestra. Con esta terminología, podemos ofrecer las siguientes definiciones:

Se llama *frecuencia absoluta* de un valor X_i, y se simboliza por n_i, al número de veces que ese valor se repite en la muestra.

Se llama *frecuencia relativa* de un valor X_i, y se simboliza por p_i, al cociente entre su frecuencia absoluta y el tamaño de la muestra. Es decir, $p_i = n_i/N$.

Se llama *frecuencia absoluta acumulada* de un valor X_i, y se simboliza por n_a, al número de veces que se repite en la muestra ese valor o cualquiera inferior a él.

Se llama *frecuencia relativa acumulada* de un valor X_i, y se simboliza por p_a, al cociente entre su frecuencia absoluta acumulada y el tamaño de la muestra. Es decir, $p_a = n_a/N$.

A veces, las frecuencias relativas, ya sean simples o acumuladas, se expresan en términos de porcentajes. En esos casos se suelen representar con mayúsculas; para obtenerlas basta con multiplicar por 100 las frecuencias relativas:

$$P_i = p_i \cdot 100 \qquad \text{y} \qquad P_a = p_a \cdot 100$$

Una distribución de frecuencias se organiza en forma de tabla, la cual puede contener todos o algunos de los elementos que hemos definido. Por supuesto, en variables cualitativas los elementos acumulados (tanto la frecuencia absoluta acumulada como la relativa acumulada) no existen, por carecer de sentido. Su definición sólo se puede aplicar a variables en las que los números empleados para representar las modalidades implican magnitudes, pero no cuando esos números son meras etiquetas.

En una distribución de frecuencias completa aparece en primer lugar una columna con los valores que adopta la variable, ordenados de forma creciente de arriba hacia abajo. En las siguientes columnas aparecen los cuatro elementos que hemos definido si se trata de una variable cuantitativa, mientras que si se trata de una cualitativa o nominal sólo pueden aparecer los dos primeros (n_i y p_i).

Veamos un ejemplo de cada uno de estos tipos de variables. La tabla 2.1 muestra un ejemplo con la siguiente variable cuantitativa. Supongamos que seleccionamos una muestra de veinte familias de una determinada población y anotamos el número de hijos (X) de cada familia. Encontramos los siguientes valores: 1, 1, 0, 1, 2, 2, 3, 1, 0, 0, 1, 2, 1, 1, 0, 2, 4, 2, 3 y 1. A partir de estos datos, construimos la distribución de frecuencias siguiendo los pasos descritos:

a) La variable es cuantitativa discreta, y en esta muestra adopta valores entre 0 y 4; por tanto, ponemos en la primera columna esos valores (X_i), creciendo de arriba hacia abajo.

b) Para la columna de frecuencias absolutas (n_i) contamos el número de veces que se repite cada valor. Una forma de comprobar que no hemos cometido ciertos tipos de errores es asegurarnos de que la columna de las n_i suma N (en este caso, 20).

c) Para la columna de frecuencias relativas (p_i) dividimos cada frecuencia absoluta por N. Para detectar errores podemos hacer la comprobación de que la suma de las p_i es igual a 1 (a veces esta columna no suma exactamente 1; por ejemplo, debido a la necesidad de redondear las p_i no es raro encontrar que la suma de esta columna sea igual a 0,999).

d) Para obtener las frecuencias absolutas acumuladas (n_a) sumamos para cada valor su frecuencia absoluta más la frecuencia absoluta acumulada del valor anterior. Aquí comprobamos que la frecuencia absoluta acumulada del valor mayor es igual a N.

e) Para las frecuencias relativas acumuladas (p_a) dividimos cada frecuencia absoluta acumulada por N. La frecuencia relativa acumulada del valor mayor debe ser igual a 1.

TABLA 2.1

Distribución de frecuencias del número de hijos de 20 familias

X_i	n_i	p_i	n_a	p_a
0	4	0,20	4	0,20
1	8	0,40	12	0,60
2	5	0,25	17	0,85
3	2	0,10	19	0,95
4	1	0,05	20	1,00
	20	1,00		

Una simple inspección de la distribución de frecuencias de la tabla 2.1 nos permite extraer de forma inmediata algunas informaciones. Por ejemplo, en la columna de frecuencias absolutas (n_i) comprobamos que el tamaño de familia más frecuente en la muestra es el de un hijo, seguido por el de dos hijos. Sólo una familia tiene más de tres hijos. De la columna de frecuencias relativas acumuladas deducimos que más de la mitad de las familias (proporción de 0,60) no alcanza la barrera de la autoreproducción; esta barrera se alcanza cuando de cada pareja de adultos nace una pareja de hijos.

A veces se organiza la distribución de frecuencias de una forma distinta. En concreto, si el número de distintos valores observados es muy grande, la distribución será larga y engorrosa (por ejemplo, si la variable es ingresos mensuales de una muestra de 500 personas, no es fácil encontrar a dos personas que ingresen exactamente lo mismo). Para hacerla más manejable, tradicionalmente se han agrupado los valores observados en conjuntos, llamados «intervalos» o «clases».

Sin embargo, este procedimiento está cayendo en desuso debido a la creciente automatización en el tratamiento de los datos.

Como ejemplo de una distribución de frecuencias con una variable cualitativa vamos a tomar las respuestas de 133 licenciados en psicología por la Universidad Autónoma de Madrid que terminaron sus estudios entre 1999 y 2000. Esos 133 constituyen una muestra de licenciados que respondieron al cuestionario telefónico y, además, dijeron que estaban trabajando como psicólogos. Una de las preguntas se refería al principal campo profesional de la psicología en el que estaban trabajando. Con los cuestionarios recogidos confeccionamos la distribución de frecuencias de la tabla 2.2, en la que hemos incluido las frecuencias absolutas (n_i) y relativas (p_i). En ella observamos que la modalidad más frecuente, aquella a la que se dedican más licenciados, es la de Recursos Humanos y Organizaciones, aunque las otras dos especialidades clásicas (Clínica y Educativa) están bastante bien representadas. Hay 20 de los 133 (el 15 por 100) que indican un campo fuera de estos tres más clásicos, lo que indica que se empieza a producir una mayor diversificación.

TABLA 2.2

Distribución de frecuencias del ejemplo de los campos profesionales de los licenciados en psicología de la UAM

Campo profesional	n_i	p_i
Psicología clínica	42	0,316
Psicología educativa	20	0,150
Recursos humanos y organizaciones	51	0,383
Servicios sociales	7	0,053
Programas psicosociales	4	0,030
Docencia	1	0,007
Otros	8	0,060
	133	1,000

En el apéndice de este capítulo se describe un instrumento alternativo a las distribuciones de frecuencias, llamado *diagrama de tallo y hojas*.

2.3. REPRESENTACIONES GRÁFICAS

Las representaciones gráficas son instrumentos ideados para proporcionar informaciones globales mediante un solo golpe de vista. Como su variedad es enorme, aquí nos centraremos sólo en las de uso más frecuente en el ámbito de la psicología. Aunque cada tipo se puede a su vez confeccionar de muy distintas formas, es conveniente adoptar algunas convenciones que faciliten su

interpretación. Cada disciplina tiene algunas de esas convenciones, con frecuencia derivadas de la simple tradición. En psicología la mayoría de las revistas científicas adoptan el estilo de la guía de publicaciones de la APA (*American Psychological Association*, 2010), por lo que remitimos al lector a la misma (véase también León y Montero, 2003). Sólo haremos un par de indicaciones orientadas a la homogeneización. La primera es que los valores de la variable se pongan en el eje de abscisas, creciendo de izquierda a derecha, y las frecuencias en el eje de ordenadas, creciendo de abajo hacia arriba. La segunda es que para facilitar su interpretación es aconsejable incluir siempre las etiquetas identificativas de ambos ejes y la leyenda que identifica cada figura, cuando hay más de una.

2.3.1. Representaciones gráficas de uso frecuente

a) *Diagrama de barras*. Se emplea tanto con variables nominales como con variables cuantitativas discretas. En el eje de abscisas se sitúan las modalidades (o los números que las representan) y en el eje de ordenadas las frecuencias (pueden ser absolutas o relativas; si es una variable cuantitativa también pueden ser acumuladas). En la figura 2.1 de la izquierda aparece el diagrama de barras de la variable *número de hijos* de la tabla 2.1, mientras que en la de la derecha aparece el de la variable *campo profesional* de la tabla 2.2.

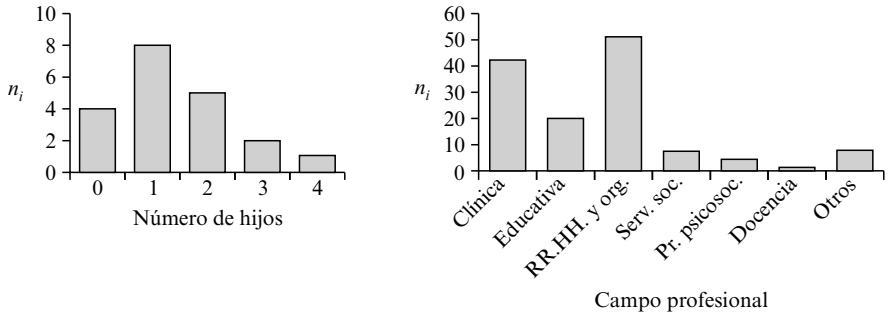


Figura 2.1.—Diagramas de barras de las variables «número de hijos» de una muestra de familias (izquierda) y «campo profesional» de una muestra de psicólogos licenciados por la UAM (derecha).

b) *Diagrama de pastel*. También denominado diagrama de sectores, se emplea con variables cualitativas. Son representaciones en forma de círculos en las que éstos son divididos, mediante radios, en secciones con superficies proporcionales a las frecuencias de las modalidades que representan. En la figura 2.2 aparece el diagrama de pastel de la variable campo profesional (en realidad, los diferentes tonos de grises aparecerían como diferentes colores).

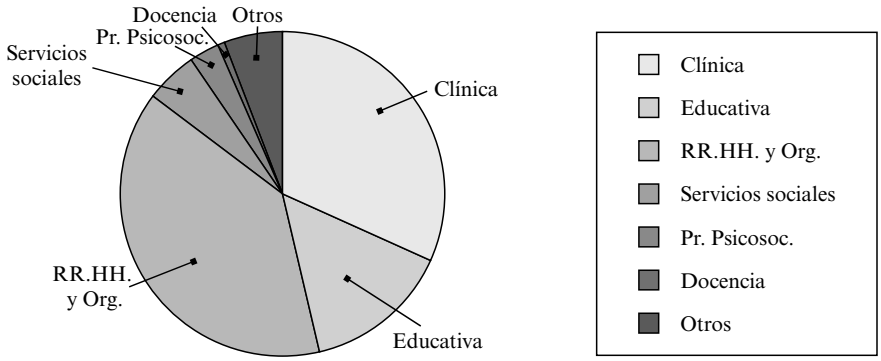


Figura 2.2.—Diagrama de pastel de la variable «campo profesional» de una muestra de psicólogos licenciados por la UAM.

c) *Histograma*. Se utiliza con variables cuantitativas continuas. Es muy parecido al diagrama de barras, pero en éste se levantan rectángulos yuxtapuestos en el eje de abscisas, con alturas proporcionales a las frecuencias. El hecho de que los rectángulos estén yuxtapuestos, en lugar de separados (como las barras), pretende recordar que se trata de una variable continua (a diferencia de los diagramas de barras, que se emplean con variables discretas); conviene recordar (véase el capítulo anterior) que en una variable continua como la del ejemplo siguiente, un valor como 24 representa a todos los valores entre 23,5 y 24,5. Se puede aplicar a frecuencias simples o acumuladas. En la figura 2.3 aparece un ejemplo.

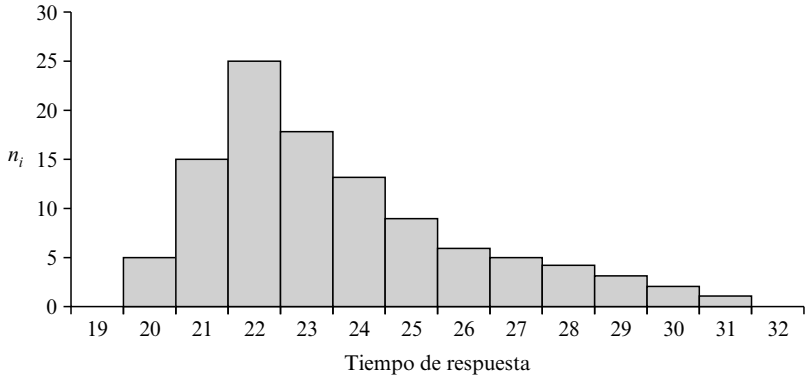


Figura 2.3.—Histograma de la variable «tiempo empleado en dar la respuesta», en décimas de segundo.

d) *Polígono de frecuencias*. Para variables discretas, el polígono de frecuencias es la figura que resulta de unir los puntos centrales de los extremos superiores de las que hubieran sido las barras si se hubiera hecho un diagrama de barras con los datos de la tabla 2.1 (figura 2.4, izquierda).

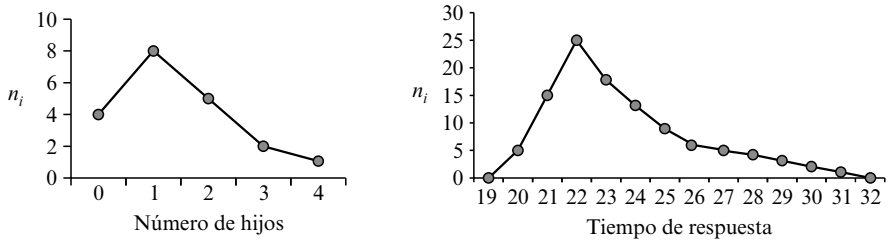


Figura 2.4.—Polígonos de frecuencias de una variable cuantitativa discreta (izquierda: número de hijos) y otra continua (derecha: tiempo de respuesta).

Si se trata de una variable continua se suele añadir un valor superior y otro inferior con frecuencia cero (figura 2.4, derecha). Es frecuente que se prefiera representar conjuntamente los datos de dos o más grupos en una misma variable. Para facilitar la comparación se puede dibujar en una misma gráfica un polígono de frecuencias por cada grupo, aunque hay otros procedimientos (véase el capítulo 8). Es importante resaltar que si los grupos son de tamaños marcadamente distintos no se deben utilizar las frecuencias absolutas sino las relativas. En la figura 2.5 presentamos un ejemplo.

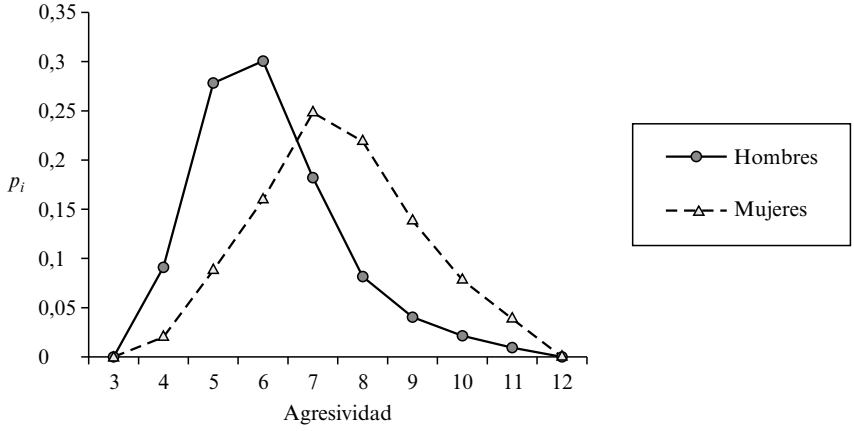


Figura 2.5.—Representación gráfica conjunta, mediante polígonos de frecuencias, de los datos de una muestra de mujeres y otra de hombres en la variable «agresividad».

2.3.2. Convenciones sobre las representaciones gráficas

No queremos dejar de hacer un comentario final sobre la flexibilidad y el sentido común aplicados a la representación gráfica, aunque pueda parecer contraria a la idea de las convenciones. En este aspecto de la estadística se ha ob-

servado como en pocos la aparición de ideas creativas e innovadoras. En realidad, cualquier representación gráfica es bienvenida, siempre y cuando sirva a los objetivos planteados y lo haga honestamente (véase el apartado siguiente). En este sentido, debemos tener presentes los hallazgos de la psicología cognitiva sobre nuestras limitaciones en cuanto al número de elementos distintos que podemos manejar simultáneamente a la hora de hacernos una idea global de un problema. Los estudios indican que este límite suele estar en torno a 7 ± 2, y, por tanto, un buen consejo sería evitar, siempre que ello no fuerce demasiado los datos, un número de intervalos, columnas, barras, grupos, secciones, etc., mayor de 9. Por otra parte, hay que tener siempre presente que las representaciones sirven para comunicar información de un solo golpe de vista, y por ello en su construcción se debe tener en cuenta el público al que van dirigidas, sus necesidades de informaciones más bien globales y generales o específicas y precisas y cualquier otra consideración que permita una comunicación ágil y precisa (Henry, 1995).

2.3.3. Tendenciosidad en las representaciones gráficas

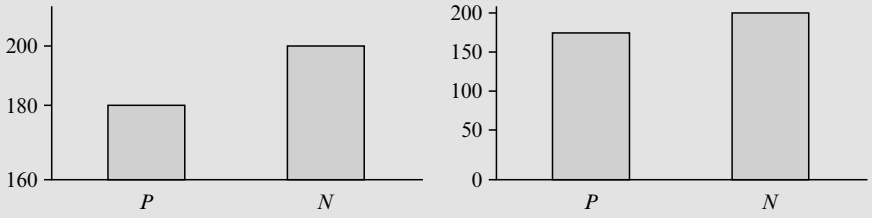
Las representaciones gráficas se pueden utilizar de manera tendenciosa para inducir impresiones engañosas e interesadas. El pequeño y divertido libro de Huff (1954) recogió por primera vez los métodos más frecuentemente utilizados para ello; posteriormente se han publicado muchos otros. Aunque su libro se podría considerar una estupenda introducción a los métodos de engaño para desaprensivos y desahogados, el autor justifica su publicación de la siguiente manera: «Quizá podría justificarla de la misma forma que lo haría el ladrón retirado cuyas memorias publicadas equivalían a un curso superior de cómo manipular cerraduras y moverse sin ser oído. Los delincuentes ya conocen esos trucos; la gente honrada debe conocerlos para su autodefensa» (Huff, 1954, p. 9). Por ese mismo argumento, hemos recogido aquí dos de los procedimientos más utilizados.

El primero consiste en recortar el eje de ordenadas (y, por tanto, las barras, los histogramas o la figura que se haya utilizado), eliminando el tramo inferior con la excusa de que no hay ninguna observación que muestre esas frecuencias. Esto tiene como consecuencia que las diferencias parezcan mayores. El segundo tipo de distorsión se produce cuando se utilizan figuras representativas de aquello que se está midiendo. Como en los diagramas de barras, estas figuras se suelen confeccionar con alturas proporcionales a las frecuencias correspondientes. Sin embargo, en este caso el incremento en la altura conlleva también un incremento en el ancho y, por tanto, en la superficie. La consecuencia es que la superficie de las figuras ya no es proporcional a las frecuencias observadas, dando la impresión de que las diferencias son mayores que las realmente registradas. Una alternativa no tendenciosa son los diagramas de barras, en los que no se produce esta distorsión porque el ancho de las barras permanece constante. En el cuadro 2.1 presentamos ejemplos de gráficas hechas con estos procedimientos, acompañadas de sus alternativas no tendenciosas.

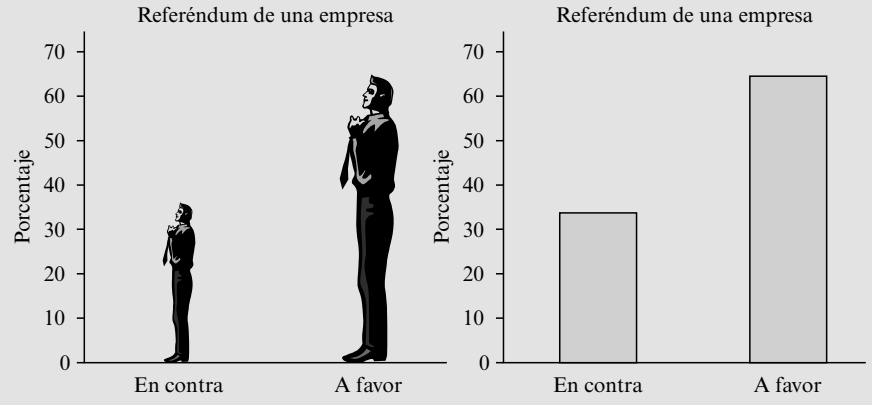
CUADRO 2.1

Ejemplos de gráficas tendenciosas, acompañadas de su alternativa correcta

a) Al confeccionar el diagrama de barras de la izquierda de la variable «diagnóstico psiquiátrico» parece que hay muchos más neuróticos que psicóticos. Esta impresión se debe a que el eje de ordenadas ha sido recortado; en realidad hay 180 psicóticos y 200 neuróticos. Al rehacer el diagrama con el eje completo (diagrama derecho), se aprecia que la diferencia real entre ambos tipos es menor de lo que parecía.



b) Para representar las frecuencias de las intenciones de voto registradas en una encuesta utilizamos siluetas de figuras humanas. Dos tercios han indicado que votarán SÍ y un tercio que votará NO (la frecuencia de respuestas afirmativas dobla la de las negativas). Las alturas de las figuras respetan estas proporciones, pero como el ancho de las figuras se establece proporcionalmente a su altura, la superficie abarcada por la silueta de la derecha es más del doble que la de la silueta de la izquierda. La impresión visual de la figura de la izquierda es que la diferencia es mayor de la real. Esto no ocurre en la representación en forma de diagrama de barras de la derecha.



2.3.4. Propiedades de las distribuciones de frecuencias

Los conjuntos de datos de variables cuantitativas tienen algunas características que iremos exponiendo detalladamente a lo largo de los próximos capítulos. Sin embargo, con objeto de preparar al lector a ubicar esos contenidos, vamos a

describir sucintamente en este punto cuáles son esas características o propiedades. Para ilustrarlas utilizaremos polígonos de frecuencias idealizadas como curvas. Describiremos los conjuntos de datos mediante cuatro propiedades:

a) *Tendencia central*. Se refiere a la magnitud general de las observaciones hechas. Esta magnitud general se puede cuantificar mediante unos índices conocidos como índices de tendencia central o promedios, y que reciben ese nombre porque pretenden ser síntesis de los valores de la variable. Así, en la figura 2.6 se puede observar que los valores del grupo A tienen una tendencia central en torno al valor 90, mientras que la tendencia central de los del grupo B está en torno al valor 100.

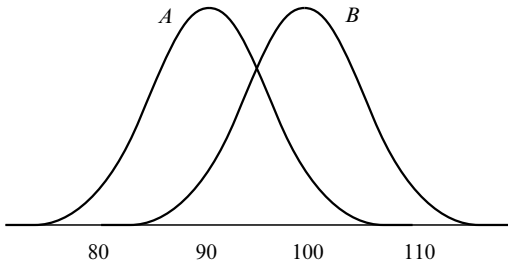


Figura 2.6.—Ejemplo de dos distribuciones con tendencias centrales distintas.

b) *Variabilidad o dispersión*. Se refiere al grado de concentración de las observaciones en torno al promedio. Una muestra de valores será homogénea o poco dispersa si los datos difieren poco entre sí y, por tanto, se concentran en torno a su promedio. Será heterogénea o muy dispersa si los datos se alejan de su promedio. Esta propiedad es independiente de la anterior, es decir, dos grupos que tengan distinta variabilidad pueden tener tendencias centrales muy distintas o similares. Así, en la figura 2.7 aparecen las representaciones gráficas de tres muestras, *A*, *B* y *C*. Las muestras *A* y *B* tienen la misma tendencia central, pero la *B* tiene mayor variabilidad o dispersión que la *A*, mientras que la muestra *C* es igual de dispersa que la *A,* pero tiene un mayor promedio.

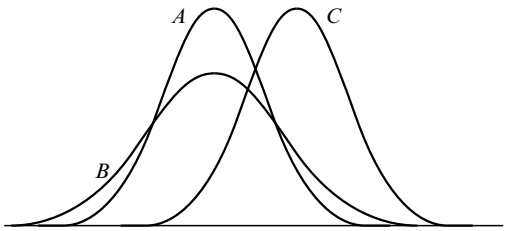


Figura 2.7.—Ejemplo de tres distribuciones en las que *A* y *B* tienen tendencias centrales similares, y menores que la de *C*, mientras que la variabilidad de *B* es mayor que la de las otras dos.

c) *Asimetría*. En la figura 2.8 aparecen las gráficas de las muestras *A*, *B* y *C*. La de la muestra *A* indica que, en general, la mayoría de los individuos han obtenido puntuaciones centrales, en torno a la media, mientras que unos pocos han obtenido puntuaciones relativamente altas y otros pocos han obtenido puntuaciones relativamente bajas: se dice que la distribución *A* es simétrica. Esto no ocurre en las distribuciones de las muestras *B* y *C*. En la primera de ellas hay muchas observaciones con puntuaciones bajas y pocas con puntuaciones altas, mientras que en la segunda ocurre lo contrario: se dice que la distribución *B* tiene asimetría positiva y la *C* asimetría negativa. Esta propiedad se refiere, por tanto, al grado en que los datos tienden a concentrarse en los valores centrales, en los valores inferiores al promedio o en los valores superiores a éste. Existe simetría perfecta cuando, en caso de doblar la representación gráfica por una línea vertical trazada sobre el valor correspondiente a la tendencia central, las dos mitades se superponen perfectamente. Las distribuciones con asimetría negativa son propias de las pruebas, tareas o tests fáciles, en las que la mayoría de los individuos puntúan alto. Las distribuciones asimétricas positivas son típicas de pruebas, tareas o tests difíciles, en las que la mayoría de los individuos puntúan bajo. Las pruebas, tareas o tests de dificultad media suelen producir distribuciones más o menos simétricas.

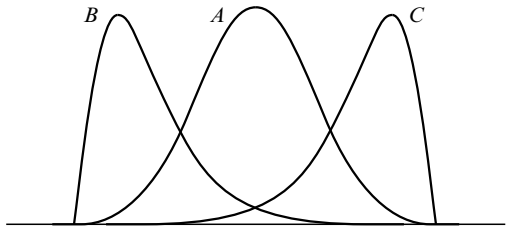


Figura 2.8.—Ejemplo de distribuciones con distintos tipos de sesgo. La distribución *A* es simétrica, la *B* es asimétrica positiva y la *C* asimétrica negativa.

d) *Curtosis*. Se refiere al grado de apuntamiento de la distribución de frecuencias. Si es muy apuntada, se llama *leptocúrtica;* si es muy aplastada, se llama *platicúrtica*. Generalmente, el grado de curtosis de una distribución se compara con un modelo de distribución llamado «distribución normal», que expondremos en capítulos posteriores, y que respecto a la curtosis se llama distribución *mesocúrtica*, pues está entre los otros dos tipos de curtosis. En la figura 2.9 aparecen las representaciones de tres muestras, *A*, *B* y *C*; la del grupo *A* es leptocúrtica, la del *B* mesocúrtica y la del *C* platicúrtica.

En este apartado sólo hemos pretendido indicar, de una forma muy intuitiva, las propiedades más importantes de las distribuciones de frecuencias. Sin embargo, no nos vamos a contentar con hacer las comparaciones visuales que, de forma algo grosera, hemos hecho en los ejemplos anteriores, sino que en los capítulos siguientes vamos a exponer procedimientos para cuantificar esas propiedades,

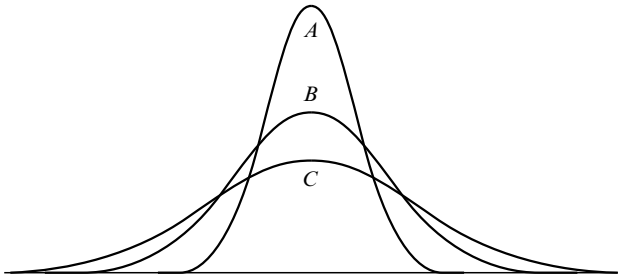


Figura 2.9.—Ejemplo de distribuciones con distintos tipos de curtosis. La distribución *A* es leptocúrtica, la *B* mesocúrtica y la *C* platicúrtica.

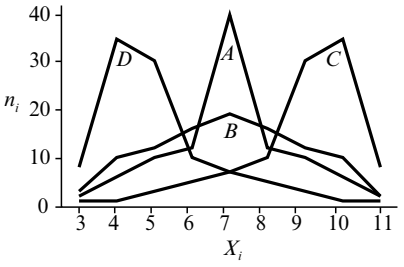
para así poder comparar muestras de valores de forma precisa en estas características. Podremos así hacer afirmaciones del tipo «la muestra *A* tiene un mayor promedio que la *B*», «la muestra *C* es menos homogénea que la *D*» o «la muestra *E* es más simétrica que la *F*». Veámoslo con un último ejemplo.

En la distribución de frecuencias de la tabla 2.3 hemos dispuesto, junto con la columna con los valores de la variable, cuatro columnas con las frecuencias absolutas obtenidas en cuatro grupos de cien personas cada uno, que representaremos por las letras *A*, *B*, *C* y *D*. Estas distribuciones de frecuencias se pueden comparar, de forma simplemente visual, en las propiedades que hemos descrito en el texto. Así, las distribuciones *A* y *B* tienen tendencias centrales parecidas, pero la distribución *A* es más homogénea que la *B*; sin embargo, ambas distribuciones son simétricas. Por el contrario, las distribuciones *C* y *D* tienen el mismo grado de variabilidad, aunque la primera tiene asimetría negativa y la segunda asimetría positiva.

TABLA 2.3

Distribuciones de frecuencias absolutas de cuatro muestras en una misma variable

X_i	*A*	*B*	*C*	*D*
3	2	3	1	8
4	6	10	1	35
5	10	12	3	30
6	12	16	5	10
7	40	19	7	7
8	12	16	10	5
9	10	12	30	3
10	6	10	35	1
11	2	2	8	1
	100	100	100	100



2.4. MEDIDAS DE POSICIÓN

Si nos informan de que una persona ha obtenido la puntuación 84 en el cuestionario de fallos cognitivos (CFC) y no sabemos nada más sobre este test, ese valor no nos resultará de utilidad. Necesitamos algo más para poder hacer una valoración de esa puntuación, dado que una valoración útil sólo se puede hacer en términos relativos. La puntuación 84 puede ser estadísticamente muy frecuente y representativa de personas sin dificultades especiales, pero también podría ser una puntuación alta, representativa de personas con algún deterioro cognitivo, ya sea el asociado a la edad o a algún tipo de trastorno. En general, para poder interpretar el significado de una puntuación es necesario hacerlo en términos relativos, y con respecto a un marco de referencia. Así, sabemos que una persona que mida 1,98 se puede calificar como de estatura alta, puesto que en el contexto de las estaturas de la especie humana ésta sería superada por muy pocos individuos. Por el contrario, la información de que un habitante de otro planeta mide 1,47 no nos dice nada, en términos relativos, acerca de si ésta es una estatura normal en su grupo de referencia o más bien se trata de una estatura extrema.

Para hacer estas valoraciones relativas se pueden utilizar las llamadas medidas de posición, que son índices diseñados especialmente para indicar la situación de una puntuación con respecto a un grupo, utilizando a éste como marco de referencia. Un tipo concreto de medida de posición son las llamadas medidas de tendencia central, que expondremos en detalle en el próximo capítulo. En esta sección vamos a describir unas medidas de posición más generales, que reciben el nombre genérico de cuantiles. Los cuantiles más utilizados, con mucha diferencia, son los *centiles* o *percentiles*; otros de menor uso son los *deciles* y los *cuartiles*.

2.4.1. Centiles o percentiles

Son 99 valores de la variable que dividen a la distribución en 100 secciones, cada una conteniendo el 1 por 100 de las observaciones. Se pueden representar por la inicial de cada uno más el subíndice correspondiente, C_k o P_k (k = 1, 2, ..., 99), aunque nosotros utilizaremos en este libro sólo la primera. Así, se simboliza por C_{28} a aquella puntuación que deja por debajo de sí al 28 por 100 de las observaciones (iguales o inferiores) y que es superada por el 72 por 100. Si disponemos de esos 99 valores, podremos hacer valoraciones relativas de las puntuaciones individuales. Por ejemplo, si un individuo obtiene la puntuación 110 en el CFC y sabemos que C_{90} = 110, quiere decir que su puntuación coincide con la del centil noventa. Es decir, el 90 por 100 de las observaciones del grupo de referencia son iguales o inferiores a la suya, mientras que es superada solamente por el 10 por 100 restante.

Aunque por definición son sólo 99 valores, por extensión a veces se utilizan posiciones intermedias, como, por ejemplo, el centil 88,5 o $C_{88,5}$, que sería aquel valor de la variable por debajo del cual se encuentra el 88,5 por 100 de las obser-

vaciones. Esta flexibilidad parecía estar presente en la mente de Galton cuando definió por primera vez los percentiles en el encabezamiento de una de las tablas incluidas en uno de los artículos en los que describió los resultados de su laboratorio antropométrico: «El valor al que no llegan el *n* por 100 de un grupo grande de mediciones, y es superado por el otro (100 − *n*) por 100, se dice que es su *n* percentil» (Galton, 1885, p. 277).

Dado que los valores correspondientes a los centiles se determinan en función de los porcentajes de observaciones, normalmente las distancias entre ellos, en términos de puntuación, no serán constantes. Generalmente, las distancias entre los centiles intermedios serán menores que las distancias entre los centiles extremos. La razón es que se suelen obtener más valores intermedios que extremos, de forma que las puntuaciones correspondientes a los centiles 55 y 56 serán más cercanas entre sí que las puntuaciones correspondientes a los centiles 98 y 99 o las de los centiles 2 y 3. Esto se dará, sobre todo, en distribuciones simétricas, mientras que, a medida que las distribuciones se van haciendo más asimétricas, esta relación hay que matizarla (véase la figura 2.10a).

Para identificar el centil correspondiente a un determinado valor de la variable basta con acudir a la columna de frecuencias relativas acumuladas (p_a) de la distribución de frecuencias, y calcular la expresión $100 \cdot p_a$ correspondiente a ese valor. Si no se ha calculado previamente p_a pero se dispone de n_a, entonces resulta más cómodo obtenerlo calculando directamente $100 \cdot (n_a/N)$. En el apartado *a*) del ejemplo expuesto unas líneas más adelante presentamos algunos ejercicios prácticos.

Aunque en este punto pueda resultar algo extraño, en cursos posteriores se verá cómo a veces interesa obtener el centil correspondiente a una puntuación imposible de observar o que, sencillamente, no ha sido observada. Así, podría interesarnos comparar los centiles de dos distribuciones distintas que se corresponden a un mismo valor (por ejemplo, qué centil ocupa la estatura 170 cm en las distribuciones de estaturas de sendas muestras de hombres y mujeres, respectivamente). Si en alguna de esas distribuciones ese valor no ha sido observado, para poder hacer la comparación hay que asignarle alguna posición relativa mediante la aplicación de algún criterio objetivo de estimación. Esto se puede hacer de varias formas, dependiendo de los supuestos que se asuman respecto a la forma y el modelo de medición (Palmer, 1999). En concreto, la pregunta que se pretende responder es, más o menos, ¿qué centil estimamos que le hubiera correspondido a ese valor en caso de haber sido observado? La manera de responder a esta pregunta no es otra que aplicar una fórmula de interpolación entre los valores inmediatamente anterior y posterior a aquel cuyo centil deseamos obtener y que sí hayan sido observados. Como este procedimiento se aplica sólo en ámbitos muy específicos no vamos a exponerlo aquí, pero se puede consultar en diversas fuentes (e.g., Botella, León, San Martín y Barriopedro, 2001; Palmer, 1999; Solanas, Salafranca, Fauquet y Núñez, 2005) (véase también el apéndice de este capítulo).

En muchas ocasiones, lo que nos interesará no es determinar el centil correspondiente a una determinada puntuación, sino lo contrario; es decir, qué puntuación es aquella a la que le corresponde un determinado centil, *k*. Para obte-

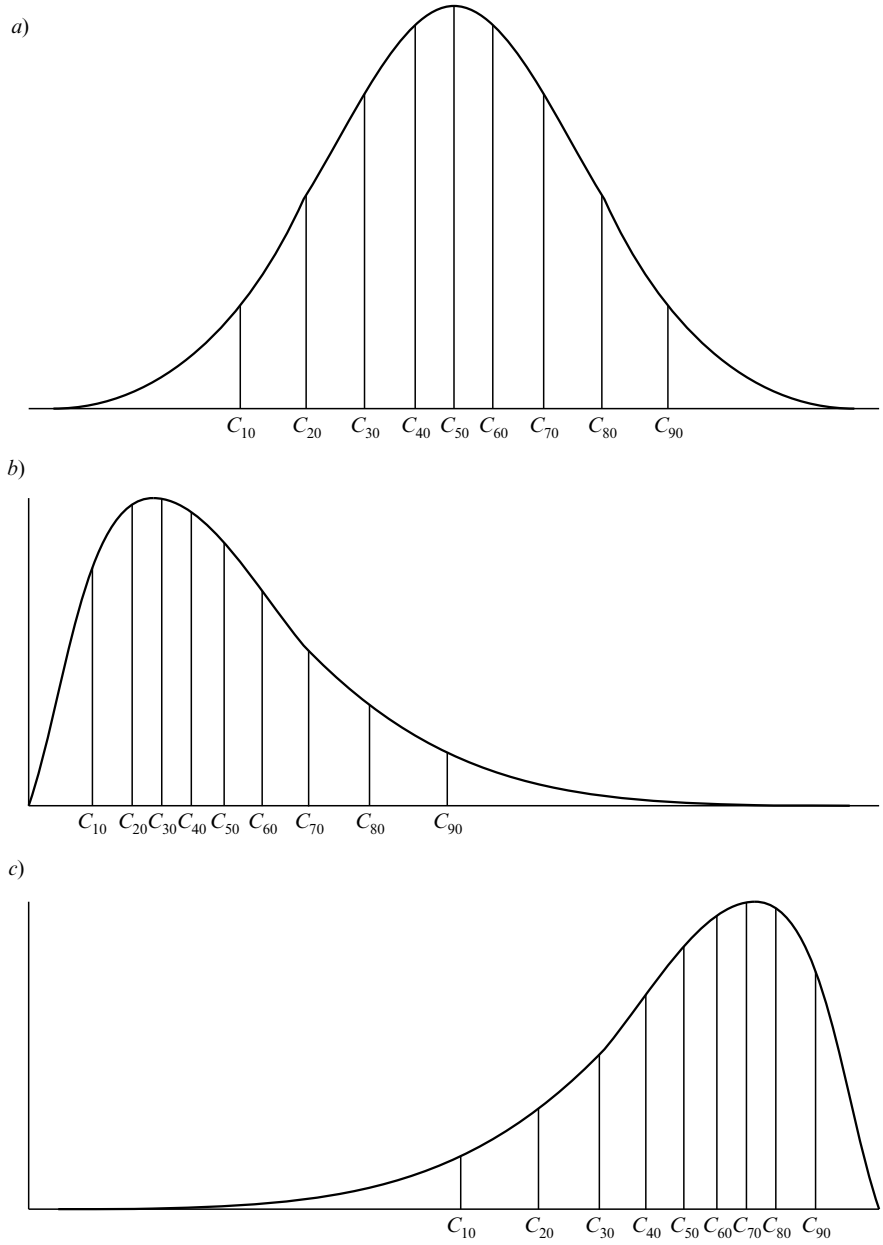


Figura 2.10.—En esta representación gráfica, sección *a*), podemos observar cómo en una distribución simétrica la distancia entre los centiles centrales (por ejemplo, entre C_{60} y C_{50}) es menor que entre los centiles extremos (por ejemplo, entre C_{90} y C_{80}). Lo que es idéntico son las áreas comprendidas entre cada dos centiles consecutivos; como la curva es más alta en la zona central, se necesita una menor base que en las zonas laterales para que las áreas sean similares. En las secciones *b*) y *c*) observamos cómo se alteran esas distancias en condiciones de asimetría.

nerla hay que acudir a la distribución de frecuencias para comprobar si hay algún valor de entre los observados para el que la expresión $100 \cdot p_a$ coincida exactamente con el valor de k buscado. En caso afirmativo, el valor correspondiente es el buscado, C_k (véase el apartado *b* del ejemplo que sigue). Pero muchas veces esto no ocurrirá. Habrá dos valores consecutivos tales que, para el menor de ellos, la expresión $100 \cdot p_a$ es menor que k, mientras que para el superior esa expresión es mayor que k. En estos casos, la puntuación buscada se puede obtener, de nuevo, mediante interpolación entre esos valores.

Ejemplo. Supongamos que hemos administrado un test de ansiedad a una muestra de 200 personas. Sabiendo que con las puntuaciones obtenidas hemos confeccionado la distribución de frecuencias siguiente: *a*) determine el centil que le correspondería a las puntuaciones 9, 13 y 17, y *b*) obtenga las puntuaciones correspondientes a los centiles 6, 48 y 70.

X_i	n_i	n_a	p_i	p_a
3	1	1	0,005	0,005
5	2	3	0,010	0,015
6	4	7	0,020	0,035
7	5	12	0,025	0,060
8	12	24	0,060	0,120
9	26	50	0,130	0,250
10	46	96	0,230	0,480
11	32	128	0,160	0,640
12	22	150	0,110	0,750
13	16	166	0,080	0,830
14	12	178	0,060	0,890
15	8	186	0,040	0,930
16	6	192	0,030	0,960
18	4	196	0,020	0,980
19	2	198	0,010	0,990
20	2	200	0,010	1,000
	200		1,000	

a) Como las puntuaciones 9 y 13 son dos de los valores empíricamente observados, sus centiles correspondientes se obtienen calculando la expresión $100 \cdot p_a$ para cada uno de ellos:

Para la puntuación 9 $\quad 100 \cdot 0{,}25 = 25 \quad \rightarrow \quad C_{25} = 9$

Para la puntuación 13 $\quad 100 \cdot 0{,}83 = 83 \quad \rightarrow \quad C_{83} = 13$

Por el contrario, la puntuación 17 no es un valor empíricamente observado, por lo que para asociarle un centil tendríamos que asumir algu-

nos supuestos y aplicar un procedimiento de interpolación (véase apéndice de este capítulo). Si no queremos hacer esa interpolación, al menos podemos responder a esta pregunta diciendo que el centil correspondiente a la puntuación 17 estará comprendido entre el 96 y el 98, dado que éstos son los centiles correspondientes a las puntuaciones inmediatamente anterior y posterior a ella ($C_{96} = 16$ y $C_{98} = 18$).

b) Empleando la terminología expuesta, estamos buscando las tres puntuaciones correspondientes a C_6, C_{48} y C_{70}. Comprobamos que en los dos primeros casos k coincide con la expresión $100 \cdot p_a$ de alguno de los valores de la distribución. En concreto, las de los valores 7 y 10 son iguales a 6 y 48, respectivamente. Por tanto, concluimos que $C_6 = 7$ y $C_{48} = 10$. Por el contrario, siguiendo con la misma lógica del apartado anterior, C_{70} es alguna puntuación intermedia entre 11 y 12, dado que los porcentajes acumulados de esos valores son los más cercanos, por arriba y por abajo, al 70 por 100 buscado. En concreto, $C_{64} = 11$ y $C_{75} = 12$. De nuevo podríamos hacer una estimación del valor buscado haciendo una interpolación entre esos valores, pero aquí nos conformamos con decir que C_{70} es algún valor comprendido entre 11 y 12 (véase el apéndice del capítulo presente).

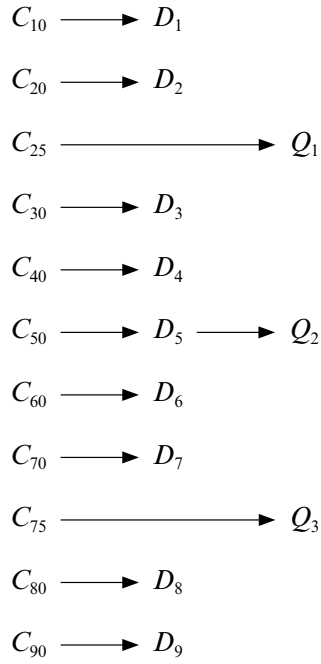
2.4.2. Otras medidas de posición. Equivalencias

A veces se utilizan otras particiones de la distribución distintas a los centiles, aunque conceptualmente son muy similares. Vamos a describir brevemente las más frecuentes: los deciles y los cuartiles.

Los *deciles* son nueve puntuaciones que dividen a la distribución en diez partes, cada una conteniendo al 10 por 100 de las observaciones. Se representan por D_k, donde k indica el número del decil al que se refiere. Así, el decil cuarto, o D_4, es la puntuación que deja por debajo de sí al 40 por 100 de las observaciones y por encima de sí al 60 por 100. Por supuesto, existe una equivalencia directa entre los deciles y los centiles, de forma que el decil primero es equivalente al C_{10}, el segundo es equivalente al C_{20}, etc. El procedimiento para calcular los deciles es el mismo que el de sus centiles correspondientes.

Los *cuartiles* son tres puntuaciones que dividen a la distribución en cuatro partes, cada una conteniendo al 25 por 100 de las observaciones. Se representan por Q_k, donde k indica el número del cuartil al que se refiere. Así, el cuartil primero, o Q_1, es la puntuación que deja por debajo de sí al 25 por 100 de las observaciones y por encima de sí al 75 por 100. Por supuesto, existe una equivalencia directa entre los cuartiles y los centiles, de forma que el cuartil primero es equivalente al C_{25}, el segundo al C_{50} y el tercero al C_{75}. El procedimiento para calcular los cuartiles es el mismo que el de sus centiles correspondientes.

Podemos resumir las equivalencias entre los cuantiles expuestos con el siguiente esquema:



Uno de los usos más frecuentes de los centiles consiste en la elaboración de baremos de pruebas de evaluación. En el apéndice de este capítulo se explica con más detalle.

PROBLEMAS Y EJERCICIOS

1. Una universidad madrileña está estudiando la implantación del plan de estudios de grado en psicología. Para ello se han consultado los expedientes académicos de 60 estudiantes que lo cursaron durante la primera promoción. En concreto, de cada estudiante anotamos la asignatura en la que ha obtenido la calificación más alta utilizando los siguientes códigos: AD (análisis de datos I), HP (historia de la psicología), IP (introducción a la psicología I), NC (neurociencia y conducta I), FPC (fundamentos psicosociales del comportamiento) y PD (psicología del desarrollo). Los resultados aparecen a continuación. Confeccione con ellos una distribución de frecuencias y una representación gráfica apropiada, comentando los resultados.

FPC	HP	PD	AD	NC	HP	PD	FPC	HP	FPC	IP	FPC
HP	IP	AD	IP	HP	IP	IP	FPC	PD	AD	IP	FPC
AD	NC	IP	HP	FPC	FPC	PD	IP	AD	NC	PD	IP
IP	NC	IP	PD	NC	HP	IP	HP	HP	FPC	IP	IP
PD	NC	IP	HP	HP	IP	PD	FPC	AD	IP	HP	IP

2. Queremos saber cuál es el tamaño más generalizado de las familias de una ciudad. Para ello seleccionamos una muestra representativa de 40 familias y anotamos el número de miembros de la unidad familiar, obteniendo los datos que aparecen más abajo. Confeccione con ellos una distribución de frecuencias y una representación gráfica apropiada, comentando los resultados.

5	7	4	2	4	5	5	5	3	3
4	3	3	3	3	2	5	3	6	4
5	4	4	3	4	4	4	3	4	6
3	4	2	2	3	5	6	5	3	4

3. Los datos que se presentan a continuación corresponden a las puntuaciones obtenidas por 50 sujetos en un test de razonamiento espacial compuesto por 22 ítems con cuatro opciones de respuesta de las que sólo una es correcta. A partir de estos datos, obtenga la distribución de frecuencias del número de aciertos en el mencionado test.

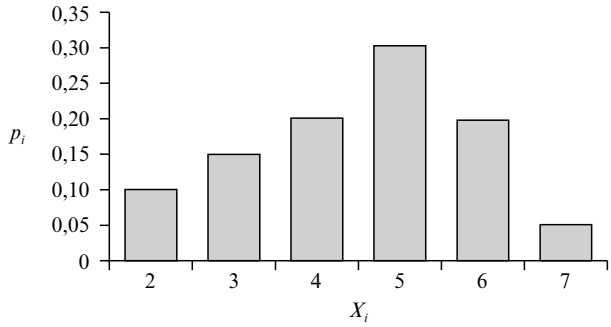
6	15	10	8	22	5	10	6	6	19
12	20	16	10	21	7	7	9	7	11
10	10	9	20	20	20	8	13	9	11
8	14	8	8	18	19	9	9	11	22
9	14	13	12	11	11	6	10	17	5

4. Con los datos del ejercicio 3, confeccione un polígono de frecuencias absolutas acumuladas.

5. En la tabla siguiente se presentan las distribuciones de frecuencias de las puntuaciones obtenidas por un grupo de chicos y otro de chicas en una prueba de fluidez verbal que se aplicó a estudiantes de bachillerato de un instituto de la ciudad de Sevilla. Complete las distribuciones de frecuencias de la tabla inferior y confeccione una representación gráfica conjunta apropiada para los datos, comentando los resultados obtenidos.

X_i	Chicas	Chicos
	n_i	n_i
0	4	1
1	6	2
2	7	5
3	12	6
4	48	15
5	100	42
6	45	14
7	45	7
8	12	4
9	16	3
10	5	1
	300	100

6. Reconstruya la distribución de frecuencias absolutas que generó el diagrama de barras de la siguiente figura, sabiendo que se confeccionó con 300 puntuaciones.



7. Reconstruya la distribución de frecuencias absolutas que generó la columna de frecuencias relativas acumuladas de la siguiente tabla, sabiendo que se confeccionó con 500 puntuaciones.

X_i	p_a
8	0,04
9	0,16
10	0,32
11	0,62
12	0,74
13	0,95
14	1,00

8. Para estudiar los hábitos de lectura de un grupo de estudiantes de Bachillerato les preguntamos al principio del curso el número de libros que habían leído durante el año anterior, recogiendo los datos que aparecen a continuación. Construya una distribución de frecuencias con los datos y dibuje una representación gráfica apropiada para ello.

3	1	1	4	1	2	0	2	1	3	1	0
1	0	2	0	1	1	1	0	3	2	3	1
0	1	1	1	0	1	4	3	3	3	2	1
5	2	2	1	0	1	1	2	1	1	2	1
0	1	2	1	3	1	2	2	3	3	1	0

9. Tras aplicar un test de extroversión separamos a un grupo de claros extrovertidos y otro de claros introvertidos y les pasamos un *autoinforme de conducta antinormativa* (*ACA*). Las puntuaciones obtenidas aparecen a continuación:

Grupo de extravertidos:	27	31	31	29	32	26	28	25	27	31
	31	24	25	33	24	25	28	28	24	31
Grupo de introvertidos:	26	24	25	27	27	33	29	31	31	27
	24	28	33	32	25	26	33	27	28	26

Confeccione las distribuciones de frecuencias correspondientes para cada grupo. A continuación, elabore una representación gráfica conjunta apropiada y extraiga a partir de ella las conclusiones oportunas, comentando la forma de la distribución en cada uno de los grupos.

10. Indique qué centiles les corresponden a los individuos que han obtenido, respectivamente, las puntuaciones 3 y 5 en la muestra que ha dado lugar a la siguiente distribución de frecuencias. ¿Entre qué centiles se encontraría una puntuación igual a 4,5?

X_i	n_i
1	5
2	20
3	50
4	15
5	10
6	25

11. Calcule C_{16}, C_{31}, $C_{38,5}$ y C_{80} en la siguiente distribución de frecuencias:

X_i	n_i
5	20
9	12
16	10
26	13
28	7
34	5
35	10
36	15
37	20
42	8
43	10
44	30
45	11
46	10
47	10
50	9

12. A partir de la distribución de frecuencias:

X_i	n_i
10,50	2
10,75	3
11,00	1
11,25	2
11,50	3
11,75	2
12,00	2
12,25	1
12,50	1
12,75	2
13,00	1

Complete la siguiente tabla:

Centil	Puntuación
( )	10,50
55	( )
( )	11,75
85	( )

13. Calcule los tres cuartiles de la siguiente distribución de frecuencias.

X_i	n_i
0	1
1	2
2	1
3	1
4	3
5	1
6	2
7	1
8	1
9	2
10	1

14. En un proceso de selección de personal se admitirán a aquellos candidatos que superen el D_7 y no hayan superado el D_9 en una prueba de atención sostenida. Si a la prueba se han presentado 400 candidatos, ¿cuántos candidatos serán seleccionados?

15. Se ha realizado un estudio sobre la capacidad de cálculo matemático en una población de niños con TGD (trastorno generalizado del desarrollo) con respecto a una población de niños con desarrollo normal (DN). Para ello se tomaron dos muestras, una de 20 niños con TGD y otra de 20 niños con DN, y se les pasó una prueba de cálculo matemático. Las distribuciones de frecuencias obtenidas fueron las siguientes:

X_i	n_{TGD}	n_{DN}
1	1	1
2	0	2
3	1	2
4	2	0
5	2	5
6	1	3
7	3	4
8	5	2
9	3	1
10	2	0

a) Calcule el Q_2 en ambos grupos.
b) Interprete el resultado a partir de los cuartiles.

16. En una investigación sobre miedo a volar se ha tomado una muestra de 50 personas y se les ha aplicado un cuestionario que indica el grado de miedo a volar. Los resultados se presentan en la siguiente distribución de frecuencias:

X_i	n_i
0	5
1	10
2	6
3	4
4	5
5	5
6	4
7	4
8	3
9	2
10	2

Se considera que el miedo a volar es moderado si se obtienen puntuaciones comprendidas entre el centil 50 y el centil 70; y severo, si una persona obtiene puntuaciones por encima del centil 92. Calcule la puntuación que ha de obtener una persona en la prueba para considerar que:

a) Su miedo a volar es moderado.
b) Su miedo a volar es severo.

17. A partir de los datos del ejercicio 3 del presente capítulo, se considera que el 64 por 100 central de las puntuaciones de la distribución de frecuencias indican un nivel de razonamiento espacial medio. ¿Qué rango de puntuaciones indican que un sujeto tiene dicho nivel de razonamiento espacial?

18. Se está diseñando una prueba para evaluar la capacidad psicomotriz y viso-espacial en niños. Ésta consiste en recorrer de forma correcta un laberinto virtual (presentado en un monitor) mediante un puntero. Los datos del ejercicio 12 del presente capítulo muestran la distribución de frecuencias de los tiempos que tardan 20 sujetos. Se considera que una persona tiene un nivel bajo de las capacidades a evaluar si el tiempo es superior a 12,25 segundos; un nivel correcto si el tiempo es superior a 10,75 segundos e igual o inferior a 12,25 segundos, y un nivel óptimo si el tiempo es igual o inferior a 10,75 segundos. Calcule:

a) Los centiles correspondientes a cada uno de los niveles.
b) El porcentaje de sujetos que hay en cada categoría.

19. Siguiendo con los datos del ejercicio anterior, supongamos ahora que una persona se encontrase en el D_6. ¿Qué tiempos cabría esperar que tardará? ¿En qué nivel se encontraría? ¿Y si estuviera en el C_{15}?

20. Se están evaluando dos procedimientos de mejora de la memoria. Para cada uno de los procedimientos, se mide el número de elementos que recuerda una persona antes y después de aplicar el procedimiento, y se calcula como índice de eficacia, que toma valores de 0 a 7, la diferencia entre elementos recordados después y antes. Los resultados obtenidos quedan resumidos en la siguiente tabla (distribución de frecuencias):

	P_a	
Eficacia	**Procedimento I**	**Procedimiento II**
0	10	5
1	40	20
2	50	40
3	70	50
4	90	70
5	95	80
6	99	95
7	100	100

Conteste a las siguientes cuestiones:

a) ¿Qué porcentaje de personas tienen en cada tratamiento un nivel de eficacia superior a 4?
b) Si se utiliza el segundo cuartil como medida de comparación entre procedimientos, ¿cuál de los dos procedimientos es más eficaz? ¿Por qué?
c) Si se considera el D_7 como el umbral al partir del cual la mejora del recuerdo es adecuado, ¿qué valor tomaría en cada procedimiento?, ¿qué procedimiento es más exigente?
d) Si una persona tiene un índice de eficacia igual a 1, ¿en qué posición se encontraría si se le hubiera aplicado el procedimiento I?, ¿y si se le hubiera aplicado el II?
e) Calcular para cada uno de los procedimientos el porcentaje de personas que obtienen un nivel de eficacia igual a 5.

21. En un centro de psicología clínica se está siguiendo el desarrollo psicomotriz de tres niños de 12 meses: CR, GH y KB. Para ello se aplica una prueba estandarizada cada tres meses, ya que cabe esperar que en cada uno de esos períodos de tiempo se produzca una maduración del niño. Para interpretar las puntuaciones se utiliza el siguiente baremo, en el que aparecen los centiles y las puntuaciones que se obtienen para poblaciones de niños con 12, 15 y 18 meses.

	Puntuaciones		
Centil	**A 12 meses**	**A 15 meses**	**A 18 meses**
10	80	90	98
15	84	94	102
20	87	97	105
25	89	99	107
30	92	102	110
35	94	104	112
40	96	106	114
45	98	108	116
50	100	110	118
55	101	111	119
60	103	113	121
65	105	115	123
70	107	117	125
75	110	120	128
80	112	122	130
85	115	125	133
90	119	129	137
95	124	134	142
99	134	144	152

Se considera que si un niño obtiene una puntuación inferior al C_{20} es recomendable una exploración más detallada, ya que puede sufrir algún tipo de problema de desarrollo psicomotriz. Por otra parte, si un niño presenta una puntuación superior al C_{90} en desarrollo psicomotriz, se puede interpretar como un indicador de una capacidad intelectual superior.

Las puntuaciones en desarrollo psicomotriz obtenidas para cada niño han sido:

	Puntuación		
Niño	**A 12 meses**	**A 15 meses**	**A 18 meses**
CR	87	94	98
GH	100	110	118
KB	119	134	152

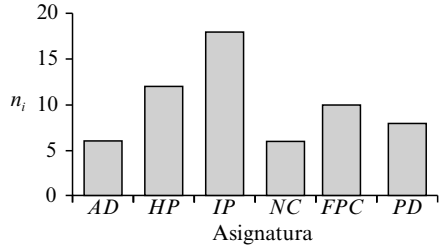
Obtenga:

a) Los centiles correspondientes a las tres puntuaciones de cada niño.
b) Diga, justificadamente, si alguno de los tres niños requiere una exploración más detallada.
c) ¿Se puede esperar que alguno de los tres niños tenga una capacidad intelectual superior?
d) Interprete el desarrollo del niño CR.
e) Interprete el desarrollo del niño GH.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

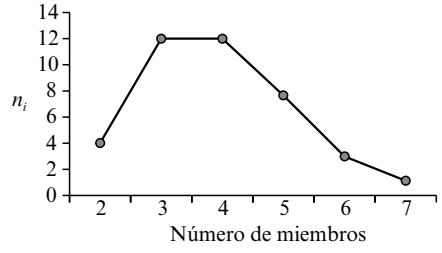
1. Optamos por elaborar un diagrama de barras:

X_i	n_i	p_i
AD	6	0,100
HP	12	0,200
IP	18	0,300
NC	6	0,100
FPC	10	0,167
PD	8	0,133
	60	1,000



2. En este caso confeccionamos un polígono de frecuencias. Como puede apreciarse, la mayor parte de las familias están compuestas como mucho por 4 miembros y sólo un 30 por 100 tienen más de 2 hijos.

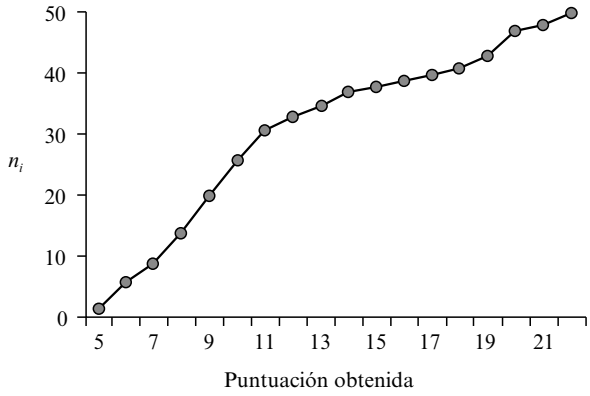
X_i	n_i	n_a	p_i	p_a
2	4	4	0,100	0,100
3	12	16	0,300	0,400
4	12	28	0,300	0,700
5	8	36	0,200	0,900
6	3	39	0,075	0,975
7	1	40	0,025	1,000
	40		1,000	



3.

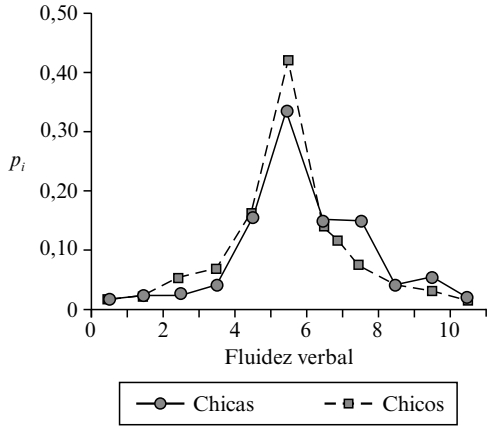
X_i	n_i	n_a	p_i	p_a
5	2	2	0,04	0,04
6	4	6	0,08	0,12
7	3	9	0,06	0,18
8	5	14	0,10	0,28
9	6	20	0,12	0,40
10	6	26	0,12	0,52
11	5	31	0,10	0,62
12	2	33	0,04	0,66
13	2	35	0,04	0,70
14	2	37	0,04	0,74
15	1	38	0,02	0,76
16	1	39	0,02	0,78
17	1	40	0,02	0,80
18	1	41	0,02	0,82
19	2	43	0,04	0,86
20	4	47	0,08	0,94
21	1	48	0,02	0,96
22	2	50	0,04	1,00
	50		1,000	

4.



5. Como los tamaños de las muestras son marcadamente distintos, utilizamos las frecuencias relativas para la representación gráfica.

X_i	Chicas				Chicos			
	n_i	n_a	p_i	p_a	n_i	n_a	p_i	p_a
0	4	4	0,013	0,013	1	1	0,010	0,010
1	6	10	0,020	0,033	2	3	0,020	0,030
2	7	17	0,023	0,057	5	8	0,050	0,080
3	12	29	0,040	0,097	6	14	0,060	0,140
4	48	77	0,160	0,257	15	29	0,150	0,290
5	100	177	0,333	0,590	42	71	0,420	0,710
6	45	222	0,150	0,740	14	85	0,140	0,850
7	45	267	0,150	0,890	7	92	0,070	0,920
8	12	279	0,040	0,930	4	96	0,040	0,960
9	16	295	0,053	0,983	3	99	0,030	0,990
10	5	300	0,017	1,000	1	100	0,010	1,000
	300		1,000		100		1,000	



Como se puede apreciar, los resultados indican que las chicas, comparadas con los chicos, tienen mayor fluidez verbal.

6.

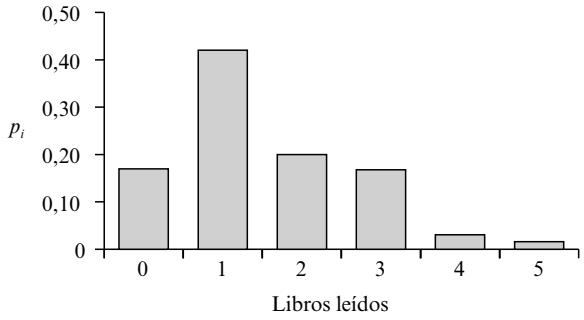
X_i	n_i
2	30
3	45
4	60
5	90
6	60
7	15
	300

7.

X_i	n_i
8	20
9	60
10	80
11	150
12	60
13	105
14	25

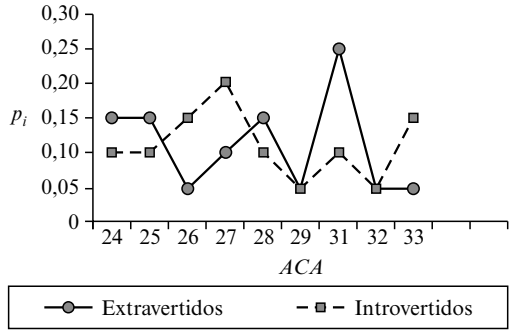
8. Tal y como se aprecia en los resultados, la mayoría de los estudiantes de esta muestra leen como máximo 2 libros al año y muy pocos leen más de 3, aunque hay un 16,7 por 100 que no han leído ningún libro.

X_i	n_i	n_a	p_i	p_a
0	10	10	0,167	0,167
1	25	35	0,417	0,584
2	12	47	0,200	0,784
3	10	57	0,167	0,951
4	2	59	0,033	0,984
5	1	60	0,016	1,000
	60		1,000	



9. Tal y como se aprecia en el diagrama, los extravertidos presentan una tendencia a mostrar valores superiores a los introvertidos en *conducta antinormativa*. Además, las gráficas muestran que la forma de la distribución de los extravertidos es asimétrica negativa y la de los introvertidos es asimétrica positiva.

X_i	Extravertidos				Introvertidos			
	n_i	n_a	p_i	p_a	n_i	n_a	p_i	p_a
24	3	3	0,150	0,150	2	2	0,10	0,10
25	3	6	0,150	0,300	2	4	0,10	0,20
26	1	7	0,050	0,350	3	7	0,15	0,35
27	2	9	0,100	0,450	4	11	0,20	0,55
28	3	12	0,150	0,600	2	13	0,10	0,65
29	1	13	0,050	0,650	1	14	0,05	0,70
31	5	18	0,250	0,900	2	16	0,10	0,80
32	1	19	0,050	0,950	1	17	0,05	0,85
33	1	20	0,050	1,000	3	20	0,15	1,00
	20		1,00		20		1,00	



10. Calculando la distribución de frecuencias:

X_i	n_i	p_i	n_a	p_a	Centiles
1	5	0,04	5	0,04	4
2	20	0,16	25	0,20	20
3	50	0,40	75	0,60	60
4	15	0,12	90	0,72	72
5	10	0,08	100	0,80	80
6	25	0,20	125	1,00	100
	125	1,00			

Se obtiene: $3 = C_{60}$; $5 = C_{80}$. Además, el valor 4,5 se encontraría entre el C_{72} y el C_{80}.

11. Calculando la distribución de frecuencias:

X_i	n_i	p_i	n_a	p_a	Centiles
5	20	0,100	20	0,100	10
9	12	0,060	32	0,160	16
16	10	0,050	42	0,210	21
26	13	0,065	55	0,275	27,5
28	7	0,035	62	0,310	31
34	5	0,025	67	0,335	33,5
35	10	0,050	77	0,385	38,5
36	15	0,075	92	0,460	46
37	20	0,100	112	0,560	56
42	8	0,040	120	0,600	60
43	10	0,050	130	0,650	65
44	30	0,150	160	0,800	80
45	11	0,055	171	0,855	85,5
46	10	0,050	181	0,905	90,5
47	10	0,050	191	0,955	95,5
50	9	0,045	200	1,000	100
	200	1,000			

Se obtiene: $C_{16} = 9$; $C_{31} = 28$; $C_{38,5} = 35$; $C_{80} = 44$.

12. Calculando la distribución de frecuencias:

X_i	n_i	p_i	n_a	p_a	Centiles
10,50	2	0,10	2	0,10	10
10,75	3	0,15	5	0,25	25
11,00	1	0,05	6	0,30	30
11,25	2	0,10	8	0,40	40
11,50	3	0,15	11	0,55	55
11,75	2	0,10	13	0,65	65
12,00	2	0,10	15	0,75	75
12,25	1	0,05	16	0,80	80
12,50	1	0,05	17	0,85	85
12,75	2	0,10	19	0,95	95
13,00	1	0,05	20	1,00	100
	20	1			

Se obtiene:

Centil	Puntuación
10	10,50
55	11,50
65	11,75
85	12,50

13. $Q_1 = 2$; $Q_2 = 4$; $Q_3 = 7$.

14. Han sido seleccionados 80 candidatos.

15. *a*) $Q_2 = 7$ en la muestra de niños con TGD y $Q_2 = 5$ en la muestra de niños con DN.
b) La diferencia de dos puntos en la prueba puede indicar una mayor capacidad de los niños con TGD en cálculo matemático.

16. *a*) Superior al $C_{50} = 3$ e igual o inferior al $C_{70} = 5$.
b) Una puntuación superior al $C_{92} = 8$.

17. Puntuaciones superiores al $C_{18} = 7$ e iguales o inferiores al $C_{82} = 18$.

18. *a*)

Nivel	Centil
Bajo	Tiempo > C_{80}
Correcto	C_{25} < Tiempo ≤ C_{80}
Óptimo	Tiempo ≤ C_{25}

b)

Nivel	Porcentaje de sujetos
Bajo	20
Correcto	55
Óptimo	25

19.

Puntuación del sujeto	Tiempo estimado	Nivel
C_{15}	Tiempo < 10,75	Óptimo
D_6	10,75 < Tiempo < 12,25	Correcto

20. *a*) En el procedimiento I: 10 por 100. En el procedimiento II: 30 por 100.
b) Para el procedimiento I: $Q_2 = 2$. En el procedimiento II: $Q_2 = 3$. Cabe esperar que el procedimiento II sea más eficaz, ya que su Q_2 es mayor.

c) Para el procedimiento I: $D_7 = 3$. Para el procedimiento II: $D_7 = 4$. Es más exigente el procedimiento II, ya que es mayor el D_7.
d) Procedimiento I: $C_{40} = 1$. Procedimiento II: $C_{20} = 1$.
e) En el procedimiento I: 5 por 100. En el procedimiento II: 10 por 100.

21. *a*)

Niño	Centil		
	A 12 meses	**A 15 meses**	**A 18 meses**
CR	C_{20}	C_{15}	C_{10}
GH	C_{50}	C_{50}	C_{50}
KB	C_{90}	C_{95}	C_{99}

b) El niño CR tenía a los 15 meses una puntuación igual al $C_{15,}$ requiriendo una exploración más detallada. A los 18 meses tuvo una puntuación igual al C_{10}, por lo que sigue requiriendo una exploración más detallada.
c) El niño KB, ya que las tres medidas se sitúan en centiles iguales o superiores al C_{90}.
d) Si sólo se interpreta la puntuación de forma absoluta, se podría concluir que el niño CR tiene una mejora de la capacidad psicomotriz. Ahora bien, en términos de baremos, con respecto a la posición relativa que ocupa en la población de niños de su edad se observa que hay un retroceso en su desarrollo.
e) Aunque en términos absolutos el niño GH tiene un incremento de la puntuación, si se interpreta su posición relativa (los centiles) se puede concluir que ocupa la misma posición, concretamente en el centro de la distribución.

APÉNDICE

Interpolación lineal

Como se ha indicado en el texto, a veces se necesita obtener algún valor intermedio, no observado, entre dos que sí se han observado. Veamos un ejemplo continuando con el del apartado 2.4.1. En el apartado *a*) de aquel ejemplo nos preguntábamos cuál sería el centil correspondiente a la puntuación 17, concluyendo de manera difusa porque el valor 17 no había sido observado. La interpolación lineal asume que los valores se distribuyen de manera homogénea entre los valores observados, por lo que la estimación del valor buscado se puede hacer por una simple regla de tres. En concreto, entre el valor 18 y el 16 (un intervalo de 2 puntos) hay un 2 por 100 de los casos (98-96). El valor 17 está en el punto medio entre esos dos valores, por lo que le corresponde también la mitad del porcentaje de ese intervalo (1 por 100). Por tanto, estimamos que el valor 17 acumula el 96 + 1 = 97 por 100 de las observaciones, así como que $C_{97} = 17$.

En el apartado *b*) del mismo ejemplo nos preguntábamos por el valor correspondiente al C_{70}. No hay en la distribución un valor que tenga como frecuencia relativamente acumulada exactamente 0,70. En cambio, los valores 11 y 12 tienen asociados valores de 0,64 y 0,75. El 70 por 100 supone añadir 6 puntos porcentuales de los 11 (75 − 64) que hay entre los valores 11 y 12. Esto supone el 54,5 por 100. Por tanto, debemos buscar el valor entre 11 y 12 que supera ese límite inferior en el 54,5 por 100 del intervalo. Como se trata exactamente de un punto, el valor pedido es 11 + 0,545 = 11,545. En consecuencia, estimamos que $C_{70} = 11{,}545$.

Diagrama de tallo y hojas

La distribución de frecuencias no es la única herramienta disponible para resumir y exponer conjuntos de datos; una alternativa a ellas es el llamado *diagrama de tallo y hojas*, ideado por Tukey (1977) en el contexto del enfoque denominado *análisis exploratorio de datos*.

Su confección requiere separar cada puntuación en dos partes: el primer o primeros dígitos por la izquierda, que reciben el nombre de tallo, y el dígito o dígitos restantes, que reciben el nombre de hojas. Por ejemplo, $X = 56$ se puede separar en 5 (tallo) y 6 (hoja). Sin embargo, como veremos más adelante, estos diagramas tienen la suficiente flexibilidad como para admitir otras posibilidades. Los pasos que hay que seguir para construir un diagrama de tallo y hojas son los siguientes:

a) Se identifican los valores máximo y mínimo observados.
b) Se toma una decisión acerca del número más apropiado de tallos distintos.
c) Se ubican en columna los tallos distintos, ordenados de forma creciente de arriba hacia abajo.
d) Se escribe cada hoja junto al tallo que le corresponda, preferiblemente ordenados según su valor.

Uno de los pasos que más dificultades suele presentar al aprendiz es la decisión acerca del número más apropiado de tallos. No hay normas estrictas al respecto; las directrices que podemos dar son muy generales. Un número de tallos superior a cinco y que no pase de veinte suele ser apropiado. Veamos un ejemplo.

Supongamos que hemos obtenido las puntuaciones de 30 personas en una variable (el grupo de datos de la izquierda), y los hemos ordenado de menor a mayor (grupo de datos de la derecha):

37, 72, 71, 65, 54, 78	32, 33, 37, 42, 46, 49
85, 42, 49, 63, 61, 32	51, 54, 55, 57, 58, 61
51, 33, 77, 93, 85, 83	63, 63, 65, 68, 71, 72
63, 55, 58, 46, 57, 73	73, 73, 73, 75, 77, 77
73, 68, 73, 91, 75, 77	78, 83, 85, 85, 91, 93

a) Los valores mayor y menor son 93 y 32, respectivamente.

b) Si tomamos la decena como tallo tendremos siete tallos distintos, que parece un número apropiado. Por tanto, separamos las puntuaciones en dos partes, con un dígito cada una. Los tallos distintos ordenados de menor a mayor son 3, 4, 5, 6, 7, 8 y 9.

c) y *d*) Colocamos en columna los tallos y escribimos cada hoja junto a su tallo correspondiente:

3	237
4	269
5	14578
6	13358
7	123335778
8	355
9	13

Aparte de ser más fácil de construir que una distribución de frecuencias, el diagrama de tallo y hojas tiene varias ventajas sobre aquella, aunque también algún inconveniente. Entre los últimos señalamos que no facilita el cálculo de estadísticos. Entre las primeras podemos señalar las siguientes:

a) Ofrece simultáneamente tanto un listado de las puntuaciones como un dibujo de la distribución; si giramos el diagrama 90º obtenemos una especie de histograma.

b) Al contener los valores de cada observación, es más fácil de modificar para obtener un dibujo con un nivel de detalle distinto, mayor o menor, de la distribución. Por ejemplo, supongamos que decidimos rehacer el diagrama anterior con un grado mayor de detalle; podemos rehacerlo dividiendo cada tallo en dos partes (hojas «altas» y hojas «bajas»):

3 −	23
3 +	7
4 −	2
4 +	69
5 −	14
5 +	578
6 −	133
6 +	58
7 −	12333
7 +	5778
8 −	3
8 +	55
9 −	13

− = La hoja toma valores entre 0 y 4
+ = La hoja toma valores entre 5 y 9

c) Otra ventaja de esta técnica es que se pueden representar dos conjuntos de datos simultáneamente en el mismo diagrama, con lo que se facilita la comparación. Veámoslo también con un ejemplo. Disponemos de los datos de un grupo de control y otro experimental, con 25 personas en cada uno. Confeccionamos un diagrama de tallo y hojas en el que los tallos son comunes y las hojas de cada grupo aparecen por separado.

Control	Experimental
23, 21, 22, 30, 17	30, 27, 21, 19, 28
15, 15, 24, 27, 30	29, 33, 35, 22, 30
25, 28, 21, 22, 16	33, 28, 24, 26, 30
18, 31, 30, 24, 22	34, 35, 35, 25, 26
20, 31, 26, 25, 26	32, 29, 28, 27, 34

Control		Experimental
87655	1 +	9
443222110	2 −	124
876655	2 +	5667788899
11000	3 −	00023344
	3 +	555

En este diagrama de tallo y hojas conjunto se aprecia de forma inmediata que, en general, los datos del grupo experimental tienden a concentrarse más en los valores altos que los del grupo de control (los tallos más repetidos son el de los veintes altos para el grupo experimental y el de los veintes bajos para el de control).

Los centiles y los baremos de los tests

Un baremo es una tabla que facilita la interpretación de las puntuaciones obtenidas en un test estandarizado. Para ello se realizan transformaciones de las

puntuaciones obtenidas tras la aplicación del test en muestras representativas de diferentes poblaciones de referencia. Una de las transformaciones más utilizadas es la de centiles, que asocia a cada puntuación del test el centil correspondiente en una determinada población de referencia.

Vemos un ejemplo con un test que mide la capacidad viso-espacial. Para crear el baremo se aplicó la prueba a muestras representativas de poblaciones que difieren en sus estudios superiores. Con las puntuaciones obtenidas en cada muestra se calcularon sus centiles. Así, para cada muestra se puede establecer una correspondencia entre las puntuaciones y los centiles. El baremo aparece en la siguiente tabla, que presenta la correspondencia entre la puntuación del test que mide capacidad viso-espacial y los centiles. En la primera columna aparecen las puntuaciones de la prueba; en las siguientes, los centiles correspondientes a las puntuaciones obtenidas por muestras representativas de titulados en estudios de Ciencias de la Salud, Ciencias Sociales e Ingeniería, respectivamente.

	Centiles		
X	**CC. Salud**	**CC. Sociales**	**Ingeniería**
1	8	23	2
2	10	25	2
3	12	28	3
4	14	31	4
5	16	34	5
6	18	37	6
7	21	40	7
8	25	43	9
9	27	47	11
10	31	50	13
11	34	53	16
12	38	57	19
13	42	60	22
14	46	63	25
15	50	66	29
16	54	69	33
17	58	72	37
18	62	75	41
19	66	77	46
20	69	80	50
21	73	82	54
22	75	84	59
23	79	86	63
24	82	88	67
25	84	89	71
26	86	91	75
27	88	92	78
28	90	93	81
29	92	94	84
30	93	95	87

Veamos cómo se utiliza. Supongamos una persona que tiene estudios superiores en Ciencias Sociales a la cual se le ha aplicado el test, obteniendo una puntuación $X = 8$; para obtener su centil, el primer paso es localizar la puntuación en la primera columna; después, siguiendo la fila hasta llegar a la columna de Ciencias Sociales se obtiene el centil respectivo, que es 43; por tanto, $C_{43} = 8$. Supongamos ahora que se ha aplicado la prueba a una persona que terminó los estudios de Telecomunicaciones, obteniendo una puntuación $X = 13$; se sigue ahora la fila hasta llegar a la columna de Ingeniería, obteniéndose el valor 22, que es el que se corresponde con el centil; de un modo más resumido: $C_{22} = 13$.

El baremo también permite conocer la puntuación, dado un centil determinado. En nuestro ejemplo, para una persona que tiene una titulación en Ciencias de la Salud y se sitúa en el C_{62} en su capacidad viso-espacial, se puede conocer cuál es su puntuación en el test mediante el baremo. Hay que situarse en la columna de Ciencias de la Salud, hasta llegar al centil informado, y se recorre la fila hasta llegar a la columna de las puntuaciones. En el ejemplo, la puntuación solicitada es 18; por tanto, $C_{62} = 18$.

Los baremos nos permiten interpretar las puntuaciones de un test estandarizado en términos relativos. Una puntuación se puede interpretar como media, alta o baja en función de la posición que ocupe ésta en la población de referencia. Siguiendo con nuestro ejemplo, una puntuación $X = 15$ puede tener diferente interpretación atendiendo a su posición relativa en las tres poblaciones de referencia. Utilizando el baremo podemos obtener el centil respectivo en cada población:

Población	Puntuación	Centil
CC. de la Salud	15	50
CC. Sociales	15	66
Ingeniería	15	29

Se observa que la misma puntuación, $X = 15$, es: *a*) media si se refiere a la población de titulados en Ciencias de la Salud, ya que corresponde al C_{50}; *b*) alta si se sitúa en la población de titulados en Ciencias Sociales, ya que corresponde al C_{66}, y *c*) baja si se sitúa en la población de titulados en Ingeniería, ya que corresponde al C_{29}.

Estadísticos univariados: tendencia central, variabilidad, asimetría y curtosis 3

En el capítulo anterior hemos presentado cuatro propiedades que permiten caracterizar las distribuciones de frecuencias y, por tanto, compararlas entre sí. En éste vamos a exponerlas con más detalle. Dedicaremos más atención a las dos más importantes, la tendencia central y la variabilidad, mientras que las dos últimas, la asimetría y la curtosis, serán expuestas de forma más escueta. Para todas ellas señalaremos los estadísticos de uso más frecuente.

3.1. MEDIDAS DE TENDENCIA CENTRAL

Ya sabemos que las medidas de posición nos permiten comparar una puntuación con los valores que ocupan ciertas posiciones especiales en el grupo de referencia. De todas esas posiciones hay una, la posición central, que suele suscitar un mayor interés; precisamente por eso se llaman *medidas de tendencia central.* Estos índices se interpretan como los representantes de la magnitud general de los valores observados. Así, si en un experimento ponemos a prueba si en condiciones de carga perceptiva alta los individuos tardan más en responder en una tarea de discriminación que si la carga perceptiva es baja, podemos decir que esperamos que los tiempos de respuesta (*TR*) serán en general más largos en la primera condición que en la segunda. Podríamos responder a la cuestión enumerando los *TR* observados, o aportando las distribuciones de frecuencias construidas bajo cada condición experimental. Sin embargo, ese procedimiento es laborioso y poco clarificador. Por el contrario, para comparar globalmente los datos éstos se suelen resumir mediante algunos estadísticos que expresan la magnitud general de las observaciones. Una de sus utilidades es la de comparar conjuntos de valores. En términos de uno de estos estadísticos, diríamos que con carga perceptiva baja esperamos un *TR* medio menor que con carga perceptiva alta.

Las medidas o índices de tendencia central deben ser valores únicos que capturen y comuniquen bien la distribución de magnitudes, como un todo. Veámoslo con un ejemplo. Supongamos que hemos administrado una prueba de memoria inmediata y hemos anotado el número de elementos correctamente recordados por diez voluntarios. Los valores registrados han sido 6, 5, 4, 7, 5, 7, 8, 6, 7 y 8. Hay varias estrategias que nos permitirían resumir en un solo indicador la mag-

nitud general observada. Una de ellas podría ser hallar el promedio de los valores observados, estadístico muy popular y sencillo de calcular, que en este ejemplo nos daría 6,3. La segunda podría consistir en tomar como indicador un valor que sea superado por la mitad de las observaciones, pero no por la otra mitad, que en este ejemplo podría ser 6,5. La tercera y última opción podría consistir en tomar el valor más frecuentemente observado, que en este caso sería el 7. Obsérvese que estos índices no necesariamente coinciden con valores realmente observados; el número de elementos correctamente recordados es un número entero y, por tanto, no puede haber observaciones iguales a 6,3 o 6,5.

Estas tres opciones son las soluciones más frecuentes. Son el fundamento, respectivamente, de los tres índices de tendencia central más conocidos y utilizados: la media aritmética, la mediana y la moda. En el resto de esta sección exponemos estos tres estadísticos.

3.1.1. Media aritmética. Puntuaciones diferenciales

El índice de tendencia central más utilizado es la media. Aunque su nombre completo es *media aritmética* (para distinguirla de otros promedios, como la media geométrica o la media armónica), por simplicidad nos referiremos a ella simplemente como «la media». Es conocida y ampliamente utilizada desde la antigua Grecia (Walker, 1975). Se define como la suma de los valores observados, dividida por el número de observaciones. Se representa con la misma letra que representa a la variable, en mayúsculas, con una barra horizontal encima. Por tanto, si recogemos N observaciones de la variable X, entonces la media de los valores observados es:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad [3.1]$$

de donde se deduce que la suma de los valores es igual a N veces la media:

$$\sum_{i=1}^{N} X_i = N \cdot \bar{X} \qquad [3.2]$$

En el ejemplo numérico anterior teníamos diez observaciones; su media será:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{6 + 5 + 4 + 7 + 5 + 7 + 8 + 6 + 7 + 8}{10} = 6{,}3$$

Algunos autores (Amón, 1993; Hays, 1988) proponen una interpretación geométrica de la media. Supongamos que tomamos una regla ideal (sin peso) sobre la que ponemos unas piezas, todas de igual peso. Colocamos una pieza

sobre el valor que ocuparía en ese eje cada una de las observaciones hechas. En caso de repetirse algún valor, se ponen tantas piezas como veces se repite el valor. Supongamos las diez observaciones siguientes: 3, 10, 8, 4, 7, 6, 9, 12, 2 y 4; su media es igual a 65/10 = 6,5. Pongamos esas piezas imaginarias en el supuesto eje, tal y como aparece en la figura 3.l; pues bien, la media es un valor tal que, si apoyamos la regla en un fulcro situado en el valor correspondiente a la media, el conjunto quedará en equilibrio. La media se comporta como si fuera el centro de gravedad de la distribución.

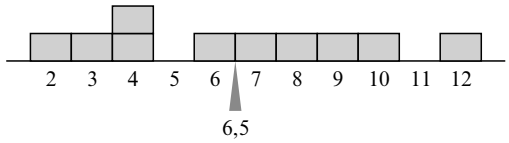


Figura 3.1.—Interpretación geométrica de la media.

Una vez definida la media aritmética estamos en condiciones de definir un tipo especial de puntuaciones que permiten dar una información más interesante que las puntuaciones que hemos visto hasta ahora (puntuaciones directas, X_i); reciben el nombre de *puntuaciones diferenciales.* Supongamos que a un individuo se le ha asignado una puntuación igual a 73 en un test de extraversión. ¿Qué podríamos decir acerca de la extraversión de este individuo? Esa puntuación aislada no resulta útil para extraer conclusiones; ella sola no nos permite hacernos una idea de si el grado de extraversión de ese individuo es alto, bajo o está entre los valores intermedios, que son los más habituales. Ya sabemos que las medidas de posición nos permiten valorar puntuaciones individuales en términos relativos, comparándolos con los de una distribución de referencia. Pero los centiles no son la única herramienta disponible para alcanzar este objetivo. Una alternativa consiste en informar de la distancia que separa a ese individuo de la media del grupo de referencia. Por ejemplo, podríamos indicar que, dado que la media del grupo de referencia es 68, el individuo con puntuación 73 se separa de la media en cinco puntos por encima de ella. Esto nos da una información que el mero valor del sujeto no nos proporciona. Es todavía una información muy rudimentaria, pero al menos nos indica si el sujeto obtuvo una puntuación superior, inferior o igual a la media del grupo. A las puntuaciones que hemos venido empleando hasta aquí las denominaremos a partir de ahora *puntuaciones directas* y las representaremos con letras mayúsculas. Por el contrario, llamaremos *puntuaciones diferenciales* a las diferencias entre cada puntuación directa y la media del grupo, y las representaremos con letras minúsculas. Por tanto:

$$x_i = X_i - \bar{X} \qquad [3.3]$$

Cada individuo u observación se puede describir tanto por su puntuación directa como por su puntuación diferencial. En el ejemplo siguiente presentamos

una muestra de cuatro puntuaciones directas, así como la puntuación diferencial asociada a cada una:

$$\begin{array}{l} X_1 = 8 \\ X_2 = 12 \\ X_3 = 10 \\ X_4 = 8 \end{array} \qquad \bar{X} = \frac{8 + 12 + 10 + 8}{4} = 9{,}5 \qquad \begin{array}{l} x_1 = 8 - 9{,}5 = -1{,}5 \\ x_2 = 12 - 9{,}5 = 2{,}5 \\ x_3 = 10 - 9{,}5 = 0{,}5 \\ x_4 = 8 - 9{,}5 = -1{,}5 \end{array}$$

Con las puntuaciones diferenciales podemos proporcionar una información más precisa que con las directas, indicando, por ejemplo, que el primer individuo se separa de la media en 1,5 puntos por debajo de ella. Una inspección cuidadosa de las puntuaciones diferenciales de este ejemplo nos revela que su suma es igual a cero. Esto no es una coincidencia. En realidad, es una propiedad fundamental de las puntuaciones diferenciales, como vamos a demostrar inmediatamente:

$$\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - \sum \bar{X} = \sum X_i - N \cdot \bar{X}$$

sustituyendo la expresión $\sum X_i$ por su equivalente, según la fórmula [3.2]:

$$\sum x_i = N \cdot \bar{X} - N \cdot \bar{X} = 0$$

En resumen, podemos expresar de la siguiente forma la principal propiedad de las puntuaciones diferenciales: la suma de las diferencias de N puntuaciones directas con respecto a su media, o puntuaciones diferenciales, es igual a cero:

$$\sum x_i = 0 \qquad [3.4]$$

De donde se deduce que la media de las puntuaciones diferenciales es igual a cero:

$$\bar{x}_i = 0$$

Esto ocurre porque algunas puntuaciones diferenciales son positivas y otras negativas (las que superan la media y las que quedan por debajo de ella, respectivamente), y se compensan unas con otras. En cambio, la suma no sería necesariamente igual a 0 si tomáramos esas diferencias en valor absoluto, si las elevásemos al cuadrado, o si fueran diferencias con respecto a cualquier otro valor distinto de la media.

3.1.2. Mediana

Como ya hemos avanzado, una opción alternativa a la media para representar la tendencia central de un conjunto de valores consiste en tomar aquella puntuación que sea superada por la mitad de las observaciones, pero no por la otra mitad. Este estadístico se denomina *mediana* y se suele representar por *Mdn*. Galton había uti-

lizado el concepto, aunque no el término, en 1869. Utilizó por primera vez el término en el capítulo sobre métodos estadísticos de uno de sus libros sobre las facultades humanas (Galton, 1883). También Fechner desarrolló de forma independiente este concepto y propuso una fórmula de interpolación muy empleada todavía.

Es muy frecuente que se quiera obtener la mediana de un conjunto de valores de los que se dispone de la correspondiente distribución de frecuencias. En estos casos la solución se reduce a la obtención del C_{50}, según los procedimientos descritos en el capítulo anterior. Como el lector habrá advertido ya, la mediana se corresponde con C_{50}, D_5 y Q_2.

Para su cálculo podemos encontrarnos en dos casos generales, aquel en el que contamos con un número impar de observaciones y aquel en que nos encontramos con un número par de ellas. En el primero se toma como mediana el valor central; en el segundo se da la circunstancia de que cualquier valor comprendido entre los dos centrales cumple con la definición de la mediana. Por ello, Fechner propuso tomar la media aritmética de los dos valores centrales. Veámoslo con dos ejemplos:

a) *Número impar de valores* ($N = 11$): 7, 11, 6, 5, 7, 12, 9, 8, 10, 6, 9.

Ordenamos estos once valores de menor a mayor: 5, 6, 6, 7, 7, 8, 9, 9, 10, 11, 12.

Como se trata de un número impar de valores, tomamos como mediana el valor central [el que ocupa el orden $(N + 1)/2 = 6$], concluyendo que se trata del valor 8:

$$Mdn = 8$$

b) *Número par de valores* ($N = 10$): 23, 35, 43, 29, 34, 41, 33, 38, 38, 32.

Valores ordenados: 23, 29, 32, 33, 34, 35, 38, 38, 41, 43. Como se trata de un número par de valores, tomamos como mediana la media aritmética de los valores centrales (5.º y 6.º): $(34 + 35)/2 = 34{,}5$; concluimos que éste es el valor de la mediana:

$$Mdn = 34{,}5$$

Algunos autores proponen fórmulas especiales para aquellos casos en los que el valor central y/o los adyacentes a él se repite(n) varias veces (Amón, 1993; Palmer, 1999). Se trata de casos particulares de uso infrecuente para los que, en caso necesario, remitimos al lector a las obras citadas. De hecho, en los paquetes estadísticos informatizados no se utilizan estas fórmulas de interpolación por defecto, sino que se proporciona la puntuación nominal del valor central, a no ser que se especifique otra cosa.

3.1.3. Moda

La tercera estrategia para definir índices de tendencia central se basa en las propias frecuencias. El procedimiento principal consiste, sencillamente, en tomar el valor más frecuentemente observado. Se trata de la *moda*, que se representa por

Mo, y se define sencillamente como el valor de la variable con mayor frecuencia absoluta. Aunque esta idea fue adelantada y definida por Fechner, fue Pearson quien dio a este estadístico su nombre actual. En su cálculo hay algunos casos particulares destacables, por lo que vamos a detenernos en ellos. Como norma, para obtener la moda ordenaremos los valores de menor a mayor, para así facilitar la identificación del de mayor frecuencia. Veamos algunos ejemplos de conjuntos de datos, ya ordenados de menor a mayor:

a) 8, 8, 11, 11, 11, 15, 15, 15, 15, 15, 17, 17, 17, 19, 19.

Es el caso más directo y sencillo; se trata de una distribución *unimodal*. El valor que más veces se repite es el 15; por tanto:

$$Mo = 15$$

b) 8, 8, 8, 11, 11, 11, 15, 15, 15, 17, 17, 17, 19, 19, 19.

Todos los valores tienen la misma frecuencia; por tanto, es un caso en el que la moda no se puede calcular. Se dice que es una distribución *amodal*.

c) 8, 9, 9, 10, 10, 10, 10, 11, 11, 13, 13, 13, 13, 15, 15.

Hay dos valores con la misma (y máxima) frecuencia, el 10 y el 13; en este caso se dice que la distribución tiene dos modas, o que es una distribución *bimodal*, donde:

$$Mo_1 = 10 \qquad \text{y} \qquad Mo_2 = 13$$

d) 8, 8, 9, 9, 9, 11, 11, 11, 11, 12, 12, 12, 12, 14, 15, 15.

Al igual que en el caso anterior, hay dos valores que comparten la máxima frecuencia (11 y 12), pero en este caso esos dos valores son adyacentes. Cuando se da esta circunstancia, se toma como moda la media aritmética de esos dos valores:

$$Mo = \frac{11 + 12}{2} = 11{,}5$$

3.1.4. Cómo elegir una medida de tendencia central

Habida cuenta de que hemos expuesto tres índices de tendencia central, cada uno basado en una estrategia diferente para representar la magnitud general, el estudiante se estará preguntando cuál debe elegir en cada caso y con qué criterio hacerlo. Vamos a exponer algunos criterios razonados para tomar decisiones.

Si no hay ningún argumento de peso en contra, se preferirá siempre la media. Esta norma general se basa en dos argumentos. El primero es que en él se fundamentan otros estadísticos que expondremos posteriormente. El segundo es que es mejor estimador de su parámetro que la mediana y la moda. Este segundo argumento significa que, en términos generales, las medias halladas sobre mues-

tras representativas se parecen más a la media poblacional de lo que las medianas y modas muestrales se parecen a la mediana y la moda poblacional.

Pero entonces, ¿qué razones pueden hacernos preferir otro índice, como la mediana? Hay al menos dos situaciones en las que se preferirá la mediana a la media: *a*) cuando la variable esté medida en una escala ordinal (véase capítulo 1), y *b*) cuando haya valores extremos que puedan distorsionar la interpretación de la media. Veámoslo en el siguiente conjunto de puntuaciones: 3, 4, 8, 5, 6, 124. Aunque la media de estos valores es 25, no parece que este valor sea una buena representación de la magnitud general de los valores. La media se ve muy influida por un valor extremo (el 124). La media es muy sensible a las puntuaciones. Un cambio grande en sólo una de ellas puede suponer un cambio importante en la media aritmética, mientras que la mediana sólo se vería alterada por cambios en los valores centrales. Supongamos, por ejemplo, que el valor 124 de la muestra anterior es un error tipográfico y que en realidad debería ser un 14. Esa alteración tiene un efecto muy importante sobre la media, que debería haber sido 6,7, mientras que la mediana no se vería alterada por el error. No podemos ofrecer reglas rigurosas para decidir cuándo un valor es lo suficientemente extremo como para que se pueda considerar mermada la representatividad de la media. En el punto en el que nos encontramos de exposición de la estadística no tenemos más remedio que dejarlo en manos del sentido común.

Pero no siempre se puede utilizar la mediana. A veces se presentan casos en los que es más apropiado utilizar la moda. No obstante, podemos de nuevo establecer una regla general en los siguientes términos: la mediana será la segunda candidata para representar la tendencia central; si no hay argumentos de peso en contra, se preferirá la mediana a la moda. Entonces, ¿qué razones pueden hacernos preferir la moda sobre la mediana? La principal se produce cuando se trate de una variable medida en una escala nominal. En esos casos no tiene sentido hacer operaciones algebraicas con los números que representan a las modalidades, pues no son más que etiquetas identificativas y no se pueden ordenar; el único índice de tendencia central apropiado será la moda.

En algunos casos, los tres índices de tendencia central dan valores parecidos, o incluso pueden coincidir exactamente, pero no necesariamente ha de ser así. En distribuciones unimodales simétricas coinciden exactamente, pero cuanto más asimétricas son las distribuciones, más diferencia suele haber entre ellos. Por eso, cuando hay valores extremos es preferible la mediana a la media: los valores extremos hacen más asimétrica a la distribución. En la figura 3.2 aparecen representaciones gráficas de tres distribuciones con los tres índices de tendencia central. En cualquier caso, cuando estos estadísticos proporcionan valores marcadamente distintos, es conveniente informar de más de uno, para comunicar una idea más completa de los datos.

3.2. MEDIDAS DE VARIACIÓN

Tal y como vimos en el apartado 2.3.4, las muestras de datos no se deben describir sólo mediante medidas de tendencia central. Dos conjuntos de puntuaciones pueden tener la misma media y ser, sin embargo, muy distintos. Para con-

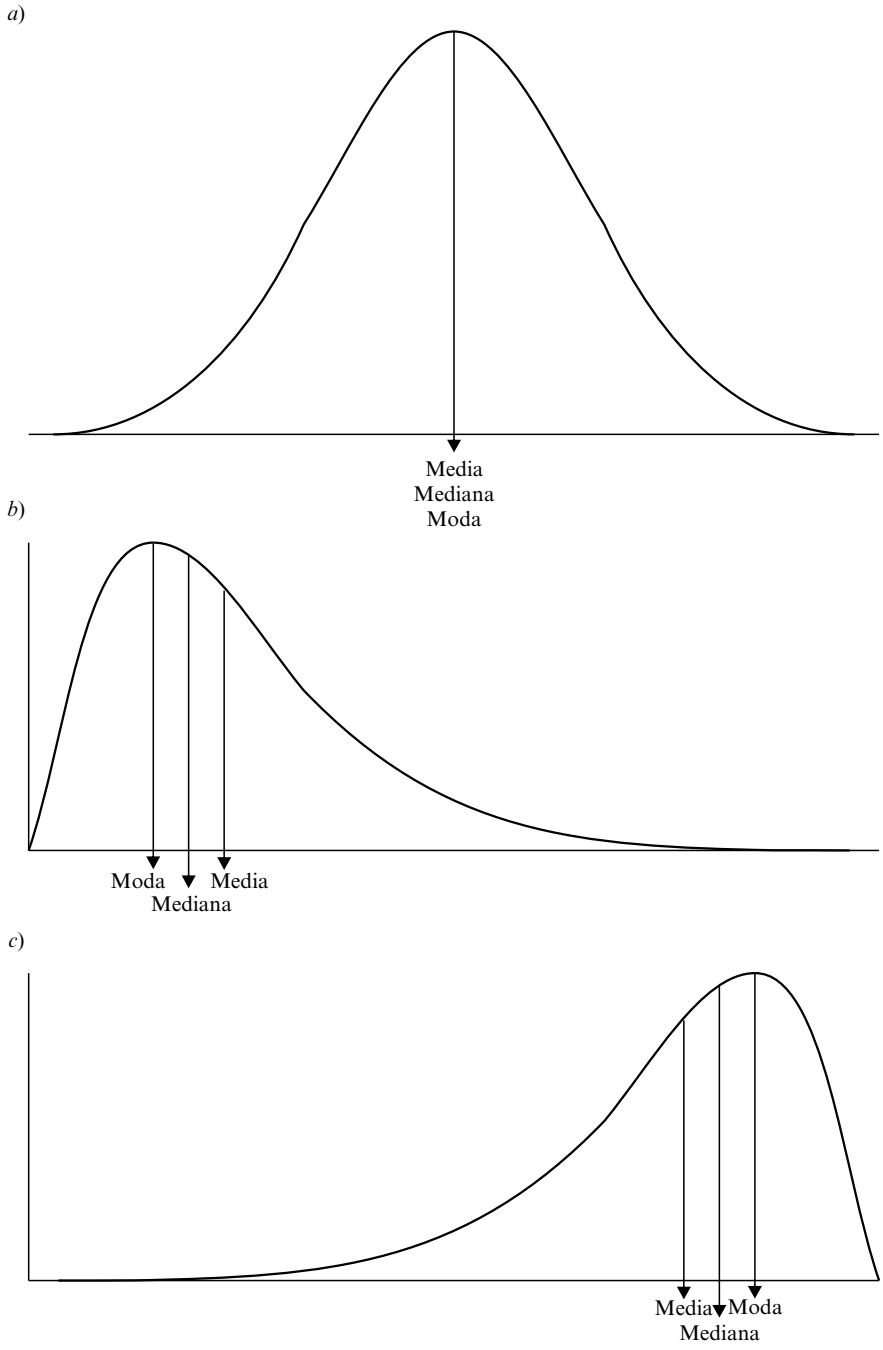


Figura 3.2.—En distribuciones simétricas y unimodales coinciden la media, la mediana y la moda [*a*)], pero a medida que se van haciendo más asimétricas, esos estadísticos tienden a separarse [*b*) y *c*)].

seguir una visión completa y comprensiva de los datos hay que complementar las medidas de tendencia central con las de otras propiedades de los mismos. Una de las más importantes es el grado en que los datos se parecen o se diferencian entre sí. Esta propiedad se denomina variabilidad o dispersión, y hay que distinguirla con claridad de la tendencia central. Supongamos, por ejemplo, que dos hermanos deciden repartir sus tierras, en ambos casos de 40.000 m^2, a los cuatro hijos que cada uno tiene, pero no siguen los mismos criterios. Mientras que el primero deja en su testamento instrucciones para legárselas a partes iguales (10.000 m^2 a cada uno), el otro decide repartirlas proporcionalmente a la productividad que cree sabrá darle cada hijo. En la figura 3.3 se representan las particiones hechas por cada uno. Aunque puede decirse que en ambos grupos de hermanos se han recibido en promedio las mismas extensiones (la media es igual a 10.000 en ambos casos), este único dato no describe por igual ambas situaciones. Seguramente el hermano del segundo grupo que recibió la parcela *A* será el primero en notar la apreciable diferencia que hay entre los dos criterios de reparto.

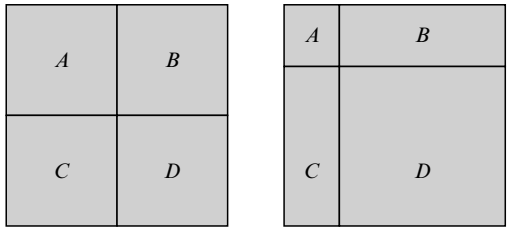


Figura 3.3.—Representación gráfica del ejemplo del texto. El hermano mayor divide sus tierras en cuatro parcelas de igual tamaño, mientras que el menor lo hace en parcelas desiguales. En ambos casos la extensión media recibida por los hermanos es la misma (ambas medias son iguales a 10.000), pero mientras que en el primero todos reciben lo mismo (las cuatro extensiones son idénticas), en el segundo hay una parcela mucho mayor (parcela *D*) y otra mucho menor (parcela *A*).

Algo parecido ocurre con las valoraciones que se hacen del nivel de vida de un estado o región basándose sólo en la renta per cápita. Una renta per cápita alta puede estar escondiendo grandes diferencias. Un país integrado por unos pocos multimillonarios y muchos pobres podría tener una renta media parecida a la de otro en el que las riquezas están mejor repartidas. Parece que hay una dimensión de los datos, diferente a la simple media, que merece la pena tener en cuenta para hacerse una idea cabal de la situación, a partir de los datos. Esta dimensión se puede apreciar numéricamente en las tres muestras de datos siguientes:

A:	4,	10,	12,	14,	20	$\bar{X}_A = 12$
B:	10,	11,	12,	13,	14	$\bar{X}_B = 12$
C:	104,	110,	112,	114,	120	$\bar{X}_C = 112$

Las tres muestras están formadas por cinco valores, pero difieren en sus características. Las muestras *A* y *B* tienen la misma tendencia central (la media es 12 en ambas), pero ya a simple vista se observa que los datos del grupo *A* son

más diferentes entre sí que los del grupo *B*; los segundos están más concentrados en torno a su media que los primeros. Por el contrario, la muestra *C* está compuesta por cinco valores cuyas diferencias entre sí son idénticas a las de los valores del grupo *A*, aunque su media sea bien distinta. En este ejemplo queda ilustrado el hecho destacado anteriormente de que la tendencia central y la variabilidad son propiedades diferentes; puede haber grupos de datos con la misma tendencia central y diferente variabilidad, y viceversa.

Naturalmente, la comparación de la variabilidad no se debe hacer mediante apreciaciones subjetivas del grado de dispersión, sino que, como en el caso de la tendencia central, vamos a exponer algunos procedimientos para cuantificar esta propiedad. Se trata de medir el grado de variación que hay en un conjunto de datos. Vamos a exponer los índices de la variabilidad más utilizados y sus propiedades.

3.2.1. Varianza y desviación típica

Aunque hay muchos procedimientos para cuantificar la variabilidad o dispersión, las medidas más importantes son la *varianza* y la *desviación típica* (en el apéndice de este capítulo se mencionan algunas más). Veamos en qué consisten y cuál es su fundamento.

Una idea que se ha demostrado útil a la hora de cuantificar la variabilidad es la de trabajar con las distancias o diferencias entre los valores y algún valor central, que podría ser la media aritmética. De esta forma, el indicador de la variabilidad o dispersión se basará en algo así como la «separación promedio» hasta ese valor. La figura 3.4 ofrece una representación de los valores de los grupos *A* y *B* del apartado anterior en un eje horizontal.

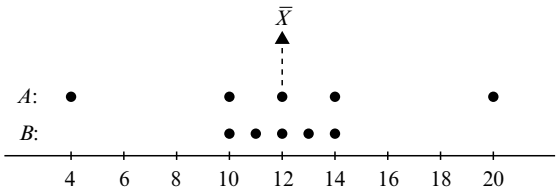


Figura 3.4.—Representación en un eje de los valores que forman los grupos *A* y *B* del ejemplo del texto.

Como se puede observar, en promedio los datos del grupo *A* están más alejados de su media que los del grupo *B* con respecto a la suya. Este hecho se puede concretar mejor calculando esas distancias; para ello obtenemos las diferencias entre cada puntuación y su media (o puntuaciones diferenciales). En el ejemplo serían las siguientes:

$$\begin{array}{lccccc} A: & -8, & -2, & 0, & 2, & 8 \\ B: & -2, & -1, & 0, & 1, & 2 \end{array}$$

El mayor grado de concentración de los valores del grupo *B* alrededor de su media se manifiesta en que estas distancias son en general menores. Por tanto, la medición de la variabilidad podría basarse en ellas hallando, por ejemplo, la media aritmética de estas distancias. Sin embargo, vimos en el apartado 3.1.1 que la suma de las puntuaciones diferenciales es necesariamente igual a cero, y, por tanto, también lo es su media. Para solventar esta dificultad se pueden adoptar varias soluciones; una consiste en tomar esas distancias en valor absoluto, pero la más utilizada consiste en tomar los cuadrados de esas distancias. En cualquiera de esos casos, el sumatorio de las cantidades así transformadas ya no es necesariamente cero, pero mantiene su propiedad de ser sensible al grado de concentración de los valores en torno a la media. La solución basada en los valores absolutos no ha demostrado ser muy útil; vamos a centrarnos en la segunda solución.

Tal y como ya hemos avanzado, una solución al problema de que las distancias con respecto a la media sumen cero consiste en elevar al cuadrado esas distancias antes de hallar su promedio, dado que los cuadrados son siempre positivos. El estadístico basado en esta idea se llama *varianza,* y se representa por la expresión S_X^2, donde el subíndice recoge la letra con la que se representa a la variable. La fórmula se reduce, por tanto, al cálculo del promedio de las desviaciones cuadráticas con respecto a la media:

$$S_X^2 = \frac{\sum(X_i - \bar{X})^2}{N} \qquad [3.5]$$

Dado que lo que aparece en el numerador son puntuaciones diferenciales al cuadrado, esta fórmula se puede expresar como el promedio de los cuadrados de las diferenciales:

$$S_X^2 = \frac{\sum x_i^2}{N} \qquad [3.6]$$

Si bien las fórmulas 3.5 y 3.6 nos proporcionan el valor de la varianza, cuando hay que calcularla a mano o con una calculadora y es un número moderadamente grande de datos es más cómodo emplear una fórmula equivalente, pero más fácil de manejar (en el apéndice del capítulo incluimos su demostración):

$$S_X^2 = \frac{\sum X_i^2}{N} - \bar{X}^2 \qquad [3.7]$$

Cuando se quiere valorar el grado de variabilidad de una muestra de valores, basta con obtener el estadístico con una de estas fórmulas. Así, en el ejemplo anterior los valores del grupo *A* eran 4, 10, 12, 14 y 20; su media aritmética era igual a 12. Las distancias con respecto a la media, o puntuaciones diferenciales, eran –8, –2, 0, 2 y 8. El promedio de los cuadrados de estas cantidades es la varianza de las puntuaciones:

$$S_{X_A}^2 = \frac{(-8)^2 + (-2)^2 + (0)^2 + 2^2 + 8^2}{5} = \frac{136}{5} = 27{,}2$$

El lector se estará quizá preguntando si el valor obtenido (27,2) indica un grado de variabilidad pequeño, mediano o grande. En realidad, no tiene mucho sentido hablar de niveles altos o bajos de dispersión en términos absolutos, sino, en todo caso, en términos comparativos. Dado que valores de varianzas que pueden ser normales en ciertas variables y poblaciones podrían parecer exagerados en otros casos, no tiene sentido comparar varianzas halladas sobre variables distintas. La varianza sirve sobre todo para comparar el grado de dispersión de dos o más muestras de valores en una misma variable, llegando a conclusiones como la siguiente: «La población de hombres presenta una mayor variabilidad en su estatura que la población de mujeres, que son más homogéneas en esa característica». Efectivamente, la clave de este índice es que es sensible a los distintos grados de concentración en torno al valor medio. Esto podemos constatarlo al calcular la varianza de los valores del grupo *B* del ejemplo anterior, y que en la representación gráfica nos parecían más homogéneos que los del grupo *A*. Su varianza será:

$$S^2_{X_B} = \frac{(-2)^2 + (-1)^2 + (0)^2 + 1^2 + 2^2}{5} = \frac{10}{5} = 2$$

A partir de sus varianzas, concluimos que el grupo *A* tiene una mayor variabilidad, una mayor dispersión o que es más heterogéneo que el grupo *B*, dado que su varianza es mayor (27,2 frente a 2). Por su parte, el lector puede comprobar que la muestra *C* tiene una varianza igual a la del grupo *A* (27,2), a pesar de tener una media muy diferente.

Al estudiar por primera vez la varianza es frecuente que el estudiante sienta cierta desconfianza al observar los valores obtenidos, encontrando natural que la media de los valores del grupo *A* sea 12, puesto que es un valor intermedio que podría representar bien la magnitud general de los datos; por el contrario, el valor 27,2 no parece un número claramente relacionado con lo que se pretendía medir. Las mayores distancias que presentan esos valores con respecto a la media son de ocho puntos, y parece que una representación numérica de la magnitud general de esas distancias estaría bastante alejada de 27,2. La razón de esta discrepancia es que las distancias no se han tratado como tales, sino que para evitar el problema de que las diferenciales sumen cero éstas se han elevado al cuadrado. Por ello es frecuente que, con objeto de retomar las unidades originales de esas distancias, se calcule la raíz cuadrada de la cantidad obtenida. El índice así hallado se llama *desviación típica* y se representa por S_X; se define sencillamente como la raíz cuadrada positiva de la varianza:

$$S_X = \sqrt{S^2_X} \qquad [3.8]$$

En los tres grupos de valores del ejemplo, las desviaciones típicas serían $\sqrt{27,2} =$ $= 5,215$ para los grupos *A* y *C*, y $\sqrt{2} = 1,414$ para el *B*. Estos valores sí parecen guardar relación con lo esperable en términos de «separación promedio».

La desviación típica es un mejor descriptor de la variabilidad, aunque la varianza tenga algunas notables propiedades matemáticas que la hacen idónea para

basar en ella análisis estadísticos más complejos. De hecho, aunque anteriormente ya se utilizaba el concepto de desviación típica (Pearson, 1894), sólo Fisher desarrolló técnicas estadísticas potentes basadas en esta idea. Se propuso un nombre especial para el manejo del concepto de varianza, tal y como se refleja en la siguiente cita, en la que por primera vez se utiliza ese término: «Por tanto, al analizar las causas de la variabilidad es deseable tratar con el cuadrado de la desviación típica como medida de la variabilidad. Llamaremos varianza a esta cantidad» (Fisher, 1918, p. 399).

Cuando un estudiante aprende por primera vez a calcular varianzas puede cometer errores en los cálculos. Conviene que nos detengamos en un aspecto de la varianza que ayudará a detectar algunos de esos errores, pero que también ayudará a comprender mejor el sentido del índice. Un conjunto de valores puede mostrar un mayor o menor grado de homogeneidad, pero el grado más pequeño posible de homogeneidad se produce cuando todos los valores son idénticos. En ese caso las desviaciones de los valores con respecto a su media (diferenciales) son todas iguales a cero y, en consecuencia, también es igual a cero la media de sus cuadrados. Por tanto, ése es el mínimo valor que puede adoptar la varianza, y además la razón por la que como desviación típica se toma la raíz positiva de la varianza. En resumen, *la varianza y la desviación típica, como medidas de variación, son valores esencialmente positivos:*

$$S_X^2 \geq 0, \qquad S_X \geq 0$$

Al estudiante de psicología, o de ciencias sociales en general, le puede parecer caprichoso el estudio de una propiedad que parece tener sentido sólo desde el punto de vista matemático o estadístico, pero que no parece guardar relación con su objeto de estudio. Nada más lejos de la realidad. Las variaciones entre los datos están reflejando variaciones en las características que se están estudiando y que en psicología suelen ser indicadores de variables psicológicas o mediciones del comportamiento. Cuando estudiamos la variabilidad de las puntuaciones obtenidas por una muestra de amas de casa en una escala de extraversión, lo que estamos estudiando es hasta qué punto esos niveles de extraversión son o no homogéneos. La variabilidad de los datos está reflejando el hecho incuestionable de las diferencias individuales, las cuales conforman uno de los objetos de estudio de la psicología. De hecho, uno de los objetivos de la psicología es precisamente la explicación sistemática de esas diferencias, en tanto en cuanto presentan regularidades asociadas a segundas o terceras variables.

En muchos libros se define la varianza de una forma ligeramente distinta a como lo hemos hecho nosotros, dividiendo por $N - 1$ en lugar de dividir por N. Todavía no nos hemos adentrado lo suficiente en la estadística como para poder explicar las razones de esta sorprendente alteración de lo que intuitivamente parecería más natural para calcular una «separación promedio»: dividir por el número completo de observaciones. Por ahora nos limitaremos a señalar la existencia de esta alternativa, a la que a partir de aquí denominaremos *cuasivarianza* (también denominada *varianza insesgada*) y que representaremos por S_{N-1}^2. Representando por S_N^2 a la varianza en la que se divide por N, vamos a ver la relación

entre ambas. Como ambas fórmulas comparten el mismo numerador, la relación entre ellas es inmediata:

$$N \cdot S_N^2 = (N - 1) \cdot S_{N-1}^2 \quad [3.9]$$

Como veremos en los capítulos dedicados a la estadística inferencial, S_{N-1}^2 tiene mejores propiedades, dado que es un mejor estimador de su parámetro que S_N^2. De hecho, la mayoría de los programas informáticos comerciales para realizar los análisis estadísticos ofrecen este estadístico en lugar de la varianza (véase Ximénez y Revuelta, 2011, p. 26). Además, forma parte de algunas de las fórmulas que emplearemos en el capítulo 14.

3.3. DOS PROPIEDADES DE LA MEDIA Y LA VARIANZA

En este apartado vamos a exponer dos propiedades de la media y la varianza que nos resultarán útiles en el futuro, aunque con ellas no se agotan las propiedades de interés; en el capítulo siguiente estudiaremos los efectos que las transformaciones lineales tienen sobre estos dos estadísticos.

3.3.1. Media y varianza total a partir de las de varios grupos

Es bastante frecuente que contemos con la media y la varianza en la misma variable de varios grupos o muestras y nos interese conocer la media y la varianza de todas las observaciones juntas. Por ejemplo, conociendo las medias de las calificaciones de los grupos escolares *A*, *B* y *C*, nos interesa ahora conocer la media de todos los alumnos de esos tres grupos juntos. Naturalmente, podemos tomar todos los valores, sumarlos y dividir por el número total; la respuesta sería correcta, pero vamos a exponer otra forma de alcanzar la misma respuesta. El primer impulso que está dando sus primeros pasos por los caminos de la estadística consiste con frecuencia en proponer para ello el cálculo de la media de las medias $(\bar{X}_A + \bar{X}_B + \bar{X}_C)/3$. Esto no es, en general, correcto; sólo lo es en el caso particular de que los tamaños de los grupos sean iguales (véase el ejercicio 22 de este capítulo).

El método que vamos a exponer permite calcular la media del grupo total, conociendo las medias de cada uno de los grupos y sus tamaños. Veamos cuál es la fórmula y cómo se llega a ella. Supongamos que tenemos k grupos, cada uno constituido por $N_1, N_2, \ldots, N_k$ observaciones (obsérvese que los grupos no tienen que ser del mismo tamaño), y hallamos en cada uno la media en la variable X, representando a esas medias por $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$. Deseamos calcular la media en la variable X de todos los individuos juntos, es decir, de los $N_1 + N_2 + \ldots + N_k$ individuos, y que representaremos por N_T. Para cada uno de los k grupos podemos expresar la relación de la fórmula [3.2]:

$$N_1 \cdot \bar{X}_1 = \sum X_{i1}$$

$$N_2 \cdot \bar{X}_2 = \sum X_{i2}$$

$$\vdots$$

$$N_k \cdot \bar{X}_k = \sum X_{ik}$$

Sumando miembro a miembro las igualdades:

$$N_1 \cdot \bar{X}_1 + N_2 \cdot \bar{X}_2 + ... + N_k \cdot \bar{X}_k = \sum X_{i1} + \sum X_{i2} + ... + \sum X_{ik}$$

Dividiendo cada miembro de esta igualdad por N_T (que no es más que $N_1 + + N_2 + ... + N_k$):

$$\frac{N_1 \cdot \bar{X}_1 + N_2 \cdot \bar{X}_2 + ... + N_k \cdot \bar{X}_k}{N_T} = \frac{\sum X_{i1} + \sum X_{i2} + ... + \sum X_{ik}}{N_T}$$

Como el segundo miembro de esta igualdad es la suma de las puntuaciones de los k grupos, dividida por el número total de individuos (N_T), se trata de la media total que estamos buscando, $\bar{X}_T$. Por tanto, conociendo los datos que aparecen en la expresión del primer miembro se puede obtener la media total. En el numerador del primer miembro aparecen las medias de los grupos parciales y sus tamaños, mientras que en el denominador aparece el tamaño total, que es igual a la suma de las N_j:

$$\bar{X}_T = \frac{N_1 \cdot \bar{X}_1 + N_2 \cdot \bar{X}_2 + ... + N_k \cdot \bar{X}_k}{N_1 + N_2 + ... + N_k} \qquad [3.10]$$

Es fácil comprender que esta fórmula consiste, sencillamente, en una ponderación de las medias de los grupos con base en sus tamaños, puesto que el peso que cada media parcial tendrá en el grupo total dependerá de su aportación, en número de datos, a la media total. Por eso esta fórmula suele denominarse media ponderada. Dicho en palabras, *la media total de un grupo de puntuaciones, cuando se conocen los tamaños y medias de varios subgrupos hechos a partir del grupo total, mutuamente exclusivos y exhaustivos, se puede obtener ponderando las medias parciales a partir de los tamaños de los subgrupos en que han sido calculadas*.

En el caso de la varianza, la forma de obtenerla para el grupo total es bastante parecida. Sin embargo, la fórmula es más compleja y vamos a obviar su demostración. En concreto, la varianza total, S_T^2, se obtiene a partir de la siguiente fórmula, en la que intervienen las medias y varianzas de los de los grupos, así como sus tamaños:

$$S_T^2 = \frac{\sum N_j \cdot S_j^2}{\sum N_j} + \frac{\sum N_j \cdot (\bar{X}_j - \bar{X}_T)^2}{\sum N_j} \qquad [3.11]$$

En esta expresión, las S_j^2 son las varianzas de cada uno de los grupos. Dicho en palabras, *la varianza total de un grupo de puntuaciones, cuando se conocen las medias, las varianzas y los tamaños de varios subgrupos hechos a partir del grupo total, mutuamente exclusivos y exhaustivos, se puede obtener sumando la media (ponderada) de las varianzas y la varianza (ponderada) de las medias*.

Veamos un ejemplo de aplicación de las fórmulas [3.10] y [3.11]. Supongamos que conocemos las medias y varianzas de tres grupos de puntaciones (grupos 1, 2 y 3). En concreto, $\bar{X}_1 = 5$, $\bar{X}_2 = 7$, $\bar{X}_3 = 8$ y $S_1^2 = 12$, $S_2^2 = 10$, $S_3^2 = 15$; además, los tamaños de los grupos son $N_1 = 20$, $N_2 = 30$, $N_3 = 50$. Sustituimos en la fórmula [3.10] para obtener la media total:

$$\bar{X}_T = \frac{20 \cdot 5 + 30 \cdot 7 + 50 \cdot 8}{20 + 30 + 50} = \frac{710}{100} = 7{,}1$$

Con este valor ya tenemos todo lo necesario para obtener la varianza total mediante la fórmula [3.11]:

$$S_T^2 = \frac{20 \cdot 12 + 30 \cdot 10 + 50 \cdot 15}{20 + 30 + 50} +$$

$$+ \frac{20 \cdot (5 - 7{,}1)^2 + 30 \cdot (7 - 7{,}1)^2 + 50 \cdot (8 - 7{,}1)^2}{20 + 30 + 50} = 14{,}19$$

En resumen, el grupo total formado por los cien casos tiene una media de 7,1 y una varianza igual a 14,19.

3.3.2. Media y varianza de una combinación lineal de variables

Otra situación relativamente frecuente es aquella en la que se forma una variable a partir de una combinación lineal de dos o más variables, e interesa conocer la media y la varianza de la variable resultante. Una combinación lineal de las variables *V, X,.., Y*, a la que representaremos por *T*, es aquella que adopta la forma:

$$T_i = a \cdot V_i + b \cdot X_i + ... + c \cdot Y_i + k \qquad [3.12]$$

Es decir, se trata de una suma (a veces resta) de variables, multiplicadas por constantes (*a*, *b*, *c*...) y sumada otra constante *k*. Lo más frecuente es que se trate de la suma simple de las variables, algo que ocurre cuando las constantes multiplicadoras son todas iguales a 1 y la constante sumada es igual a 0. Esto es lo que ocurre, por ejemplo, cuando la puntuación total en un test (*T*) se define como la suma de las puntuaciones obtenidas en *k* ítems (I_j): $T = I_1 + I_2 + ... + I_k$. Sin embargo, a veces la combinación que interesa no es la simple suma, porque se considera que en la puntuación total de un test debemos dar más peso a la puntuación en algunos subtests, frente a otros subtests. Por ejemplo, supongamos que (por razones que no vienen al caso) decidimos que el índice que mejor refleja la

capacidad intelectual de un individuo es el que se obtiene dando doble peso a la escala verbal que a la escala manipulativa de un determinado test de inteligencia. Es decir, representando por I a la inteligencia, V a la capacidad verbal y M a la manipulativa, definimos ese índice como:

$$I = \frac{(2 \cdot V + M)}{3} = \frac{2}{3} \cdot V + \frac{1}{3} \cdot M$$

Lo que nos planteamos es si podemos hallar la media y la varianza de la nueva variable, I, conociendo los estadísticos de las variables que participan en su definición.

Se podría hallar la media calculando la puntuación combinada de cada observación del grupo, sumándolas y dividiendo por su número. Sin embargo, esta propiedad nos evitará ese directo pero laborioso camino, mediante la aplicación de una sencilla fórmula que sólo exigirá conocer las medias de las variables que intervienen en la ponderación. Veamos su demostración. Supongamos que definimos la variable T de la forma que aparece en la expresión [3.12]. La media de las N puntuaciones combinadas T_i será:

$$\bar{T} = (1/N) \cdot \sum T_i = (1/N) \cdot \sum (a \cdot V_i + b \cdot X_i + ... + c \cdot Y_i + k)$$

Vamos a desarrollar el sumatorio del paréntesis de la siguiente forma:

$$\bar{T} = (1/N) \cdot (a \cdot \sum V_i + b \cdot \sum X_i + ... + c \cdot \sum Y_i + \sum k)$$

$$\bar{T} = a \cdot (1/N) \cdot \sum V_i + b \cdot (1/N) \cdot \sum X_i + ... + c \cdot (1/N) \cdot \sum Y_i + (1/N) \cdot N \cdot k$$

De donde llegamos a la expresión final de esta propiedad:

$$\bar{T} = a \cdot \bar{V} + b \cdot \bar{X} + ... + \mathrm{c} \cdot \bar{Y} + k \qquad [3.13]$$

Podemos expresarla diciendo que *una variable definida como la combinación lineal de otras variables tiene como media la misma combinación lineal de las medias de las variables que intervienen en su definición.*

Veamos un ejemplo de su aplicación. Supongamos que disponemos de las puntuaciones de 4 personas en 3 variables (V, X, Y) y que definimos la variable T como la siguiente combinación lineal de esas tres variables:

$$T_i = 2 \cdot V_i - X_i + \frac{1}{3} \cdot Y_i + 25$$

En la siguiente tabla aparecen los valores y las medias de las variables que intervienen en la definición de la variable combinada, así como sus medias. También se incluyen las puntuaciones T_i y la media de las mismas.

Sujeto	*V*	*X*	*Y*	*T*
1	5	8	33	$2 \cdot 5 - 8 + (1/3) \cdot 33 + 25 = 38{,}00$
2	4	4	29	$2 \cdot 4 - 4 + (1/3) \cdot 29 + 25 = 38{,}67$
3	6	12	35	$2 \cdot 6 - 12 + (1/3) \cdot 35 + 25 = 36{,}67$
4	1	4	23	$2 \cdot 1 - 4 + (1/3) \cdot 23 + 25 = 30{,}67$
Media	4	7	30	36,00

Aunque por supuesto que es correcto hallar el valor T de cada sujeto y promediarlos (última columna de la tabla), esta propiedad nos permite obtenerla a partir de las medias de las variables que participan en su definición:

$$\bar{T} = 2 \cdot \bar{V} - \bar{X} + \frac{1}{3} \cdot \bar{Y} + 25 = 2 \cdot 4 - 7 + \frac{1}{3} \cdot 30 + 25 = 36$$

Con respecto a la varianza de T, se puede obtener también a partir de los estadísticos de las variables intervinientes. Sin embargo, nos hace falta un estadístico que aún no hemos expuesto. En el capítulo 5, una vez expuesto el estadístico que necesitamos (llamado *covarianza*), volveremos a retomar esta propiedad.

3.4. ASIMETRÍA

La asimetría es otra propiedad con la que se puede caracterizar a una distribución de valores y con la que comparar dos distribuciones. Esta propiedad hace referencia al grado en que los datos se reparten equilibradamente por encima y por debajo de la tendencia central. Podemos imaginarnos un examen muy fácil, en el que abundan las notas altas y escasean los suspensos. Dado que las calificaciones máxima y mínima son 10 y 0, respectivamente, la representación gráfica de la distribución de frecuencias quedaría inclinada hacia la derecha (asimetría *negativa*). Una figura parecida, pero en el sentido inverso, aparecería en un examen difícil, con muchos suspensos y pocas notas altas (asimetría *positiva*). Por el contrario, una distribución equilibrada sería aquella en la que las frecuencias se repartiesen imparcialmente en torno a la media. En la psicología encontramos fácilmente ejemplos de los tres tipos de distribución. Así, los tests de inteligencia suelen presentar distribuciones bastante simétricas cuando se administran a muestras relativamente grandes; una variable que se utiliza mucho en el estudio de los procesos superiores es el tiempo de respuesta, cuya distribución suele tener asimetría positiva; en tareas perceptivas de dificultad baja en las que se cuentan el número de «blancos» detectados se suele dar lo que se conoce como «efecto techo», puesto que hay muchos sujetos que detectan todos los «blancos» y, por tanto, la distribución suele mostrar asimetría negativa. Ejemplos de estos tres tipos de distribución aparecen en la figura 3.5.

	A	*B*	*C*
	Asimetría negativa n_i	**Simetría** n_i	**Asimetría positiva** n_i
0	1	2	5
1	2	5	10
2	3	8	17
3	6	10	20
4	8	15	16
5	12	20	12
6	16	15	8
7	20	10	6
8	17	8	3
9	10	5	2
10	5	2	1
11	0	0	0
	100	100	100

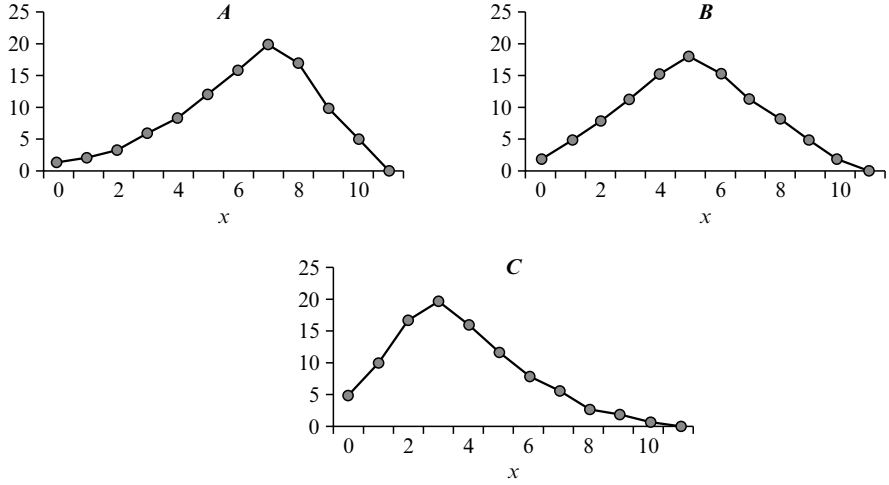


Figura 3.5.—Tres distribuciones de frecuencias con distintos tipos de asimetría y sus representaciones gráficas. El primero corresponde a un examen fácil, el segundo a un examen de dificultad media y el tercero a un examen difícil.

Aunque se han propuesto diferentes estadísticos con los que cuantificar esta propiedad, aquí vamos a exponer sólo uno, que es el que se emplea con mayor frecuencia. Se trata del índice de Pearson:

$$As = \frac{\Sigma\left(\frac{x_i}{S_X}\right)^3}{N} \qquad [3.14]$$

En esta fórmula se obtiene para cada caso el cociente entre su puntuación diferencial (x_i) y la desviación típica de la muestra (S_X); estos cocientes se elevan al cubo y se promedian (en el próximo capítulo veremos que esos cocientes se llaman «puntuaciones típicas» y les dedicaremos bastante atención, dada su importancia).

La valoración e interpretación de los valores que proporciona este índice es la siguiente. Los valores negativos indican asimetría negativa y son propios de distribuciones como las de la figura 3.4*A*. Los valores positivos indican asimetría positiva y son los esperables en distribuciones como la de la figura 3.4*C*. Por el contrario, las distribuciones simétricas proporcionarán valores próximos a cero. Veamos un ejemplo de uso de la fórmula 3.14. Los valores son 6, 7, 2, 5 y 5; estos valores tienen como media 5 y como desviación típica 1,673. Por tanto, el numerador de la fórmula será:

$$\Sigma\left(\frac{x_i}{S_X}\right)^3 = \left(\frac{6-5}{1{,}673}\right)^3 + \left(\frac{7-5}{1{,}673}\right)^3 + \left(\frac{2-5}{1{,}673}\right)^3 + \left(\frac{5-5}{1{,}673}\right)^3 + \left(\frac{5-5}{1{,}673}\right)^3 = -3{,}842$$

Por tanto, el grado de asimetría es $As = 3{,}842/5 = -0{,}768$, un valor que revela una asimetría negativa para esta muestra de datos.

3.5. CURTOSIS

La curtosis se refiere a una propiedad de uso muy técnico e infrecuente en la psicología aplicada. Aun así, vamos a exponer el índice de uso más frecuente. Está relacionado con el de la asimetría. Si en aquella se elevaban los cocientes entre las diferenciales y la desviación típica al cubo, en ésta se elevan a la cuarta potencia, aunque al final se le resta un 3:

$$Cr = \frac{\Sigma\left(\frac{x_i}{S_X}\right)^4}{N} - 3 \qquad [3.15]$$

Naturalmente, lo que más sorprenderá al lector es el hecho de que al promedio de esos cocientes elevados a la cuarta potencia se le reste 3. La razón es que existe un modelo de distribución, llamado «distribución normal» y del que hablaremos en temas posteriores, en el que ese promedio es exactamente igual a 3. Al restar un tres al promedio, lo que se consigue es utilizar ese modelo como patrón de comparación. Una distribución en la que el estadístico [3.15] sea igual a cero tiene un grado de curtosis similar al de la distribución normal y, siguiendo la terminología propuesta por Pearson (1906), se dice que es *mesocúrtica*. Si es positivo, su grado de apuntamiento es mayor que el de la distribución normal y se dice que es una distribución *leptocúrtica*, mientras que si es negativo su apuntamiento es menor que el de la distribución normal y se dice que es *platicúrtica*. Veamos un ejemplo de uso de la fórmula 3.15 con los mismos valores que empleamos para la asimetría. El numerador de la fórmula será:

$$\Sigma\left(\frac{x_i}{S_X}\right)^4 = \left(\frac{6-5}{1{,}673}\right)^4 + \left(\frac{7-5}{1{,}673}\right)^4 + \left(\frac{2-5}{1{,}673}\right)^4 + \left(\frac{5-5}{1{,}673}\right)^4 + \left(\frac{5-5}{1{,}673}\right)^4 = 12{,}499$$

Por tanto, el grado de curtosis es $Cr = \left(\frac{12{,}499}{5}\right) - 3 = -0{,}500$, un valor que revela una distribución ligeramente platicúrtica, pues es algo menor que el de la distribución normal.

PROBLEMAS Y EJERCICIOS

1. Calcule la media aritmética de los siguientes conjuntos de valores:

a) 4; 6; 7; 34; 1; 3; 12; 9; 2; 5.
b) −5; −8; −7; 0; −1; −2; −1; −4; −10; −7.
c) 33; 21; 24; 24; 36; 32; 26; 13; 10; 25; 14; 21.
d) 1,2; 1,6; 2,4; 1,4; 1,8; 1,2; 2,7; 1,5.
e) 37; 20; 21; 3; 35; 33; 33.

2. Se ha pedido a un sujeto que decida si está presente la letra *U* dentro de un conjunto de letras *V*, midiéndose el tiempo de respuesta (*TR*) que tarda el sujeto en realizar la tarea. Obtenga el *TR* medio a partir de los *TR* obtenidos en los diferentes ensayos, que se presentan a continuación:

288 297 253 249 200 249 200 259 273 261 261

3. Si el grupo de estudiantes de bachillerato seleccionados en el ejercicio 8 del capítulo 2 constituyen una muestra representativa de estudiantes de la Comunidad de Madrid, ¿qué podemos decir sobre los hábitos lectores de la población de estudiantes de Bachillerato?

4. En una investigación sobre el rendimiento académico en estudiantes de secundaria, se obtuvieron las calificaciones medias en la materia Ciencias Naturales en tres grupos de estudiantes de diferentes comunidades autónomas, los resultados se resumen en la siguiente tabla:

Comunidad autónoma	*N*	Calif. media CC. naturales
Navarra	20	7
Rioja	14	5,5
Extremadura	18	6

Calcule la media de la calificación en Ciencias Naturales de todos los estudiantes juntos.

5. Sabemos que al incorporar a la investigación del ejercicio anterior la calificación media de una muestra de 13 estudiantes de la Comunidad Autónoma de Galicia, la media total pasa a ser 6. ¿Cuál es la media del grupo de estudiantes de Galicia?

6. Considerando los datos del ejercicio 9 del capítulo 2, calcule la media en la puntuación de todos los participantes aplicando la fórmula [3.10].

7. Demuestre que cuando se tienen J grupos con el mismo tamaño muestral, la media total es igual a la media de las medias de cada grupo.

8. Obtenga la mediana de los siguientes conjuntos de datos:

a) 6, 11, 7, 8, 9, 12, 5.
b) 17, 19, 33, 37, 36, 31, 32, 18.
c) 9, 14, 10, 24, 26, 27, 11, 13, 25.
d) 25, 3, 8, 21, 29, 30, 14, 16, 15, 21, 10, 25.

9. Obtenga la mediana y la moda para los niños con TGD y niños DN del ejercicio 15 del capítulo 2. Interprete los resultados.

10. Tomando de nuevo los datos del ejercicio 15 del capítulo 2, forme un solo grupo y calcule la mediana y la moda.

11. En un centro de día se aplicó un programa conductual para mejorar la *autonomía* de ancianos con demencia moderada. Una de las conductas con las que se trabajó fue la de vestirse. Para ello, se seleccionó una muestra de diez personas mayores con demencia moderada. Se tomaron medidas del tiempo en vestirse de manera autónoma antes y después del programa conductual. Para evaluar la eficacia del programa se calculó la diferencia de tiempo antes y después de la aplicación. Sabiendo que la media de tiempo invertido en vestirse antes de la aplicación fue igual a 30 minutos y la diferencia media fue igual a 5 minutos, calcule la media de tiempo invertido después de la aplicación.

12. Se realizó la misma investigación del ejercicio anterior con una muestra de 20 ancianos de una residencia. Los resultados mostraron que la media antes de la aplicación fue igual a 35 y la media después de la aplicación fue igual a 25. Calcule:

a) La reducción media.
b) La media antes de la aplicación, la media después y la reducción media de una muestra formada por los participantes del ejercicio anterior y los de la residencia.

13. Proporcione un conjunto de diez datos que tengan de media 31, mediana 32 y moda 30.

14. Proponga razonadamente y calcule un índice de tendencia central para los datos del ejercicio 1 del capítulo 2.

15. Se ha aplicado un cuestionario de satisfacción laboral a cuatro grupos de funcionarios de diferentes ministerios. Calcular la media de todos los participantes si las medias en satisfacción laboral y los tamaños de cada grupo son:

	Grupo			
	A	***B***	***C***	***D***
Media	4	3,5	2	4,5
Tamaño	20	18	40	15

16. Tras aplicar el cuestionario del ejercicio anterior a un nuevo grupo, cuya media es 3,5, la media total es ahora 3,160. ¿Cuál es el tamaño del nuevo grupo?

17. Calcule la media, la varianza y la desviación típica de los datos del ejercicio 2 del capítulo 2 (*número de miembros de una unidad familiar*).

18. Teniendo en cuenta que la variable X toma en una muestra los valores: 2, 5, 4, 3, 2 y 1, obtenga la media y la varianza de X.

19. Teniendo en cuenta que la variable estatura (X, medida en cm) toma los siguientes valores para varones y mujeres:

Varones: 181, 185, 175, 172, 189.

Mujeres: 165, 167, 161, 168, 170.

Obtenga la media y la desviación típica de X en cada uno de los grupos. A continuación, comente los resultados.

20. Medida una variable X en una muestra de cinco personas, se ha obtenido una cuasivarianza (o varianza insesgada) de 5. A partir de estos únicos datos, calcule el valor de la varianza de X.

21. Calcule la varianza con los datos del ejercicio 9 del capítulo 2 e indique cuál de las dos muestras es más homogénea en conducta antinormativa (*ACA*), la de sujetos extravertidos o la de introvertidos.

22. Averigüe la media y la varianza total de las dos series de puntuaciones que se presentan a continuación, a partir de sus medias y varianzas parciales:

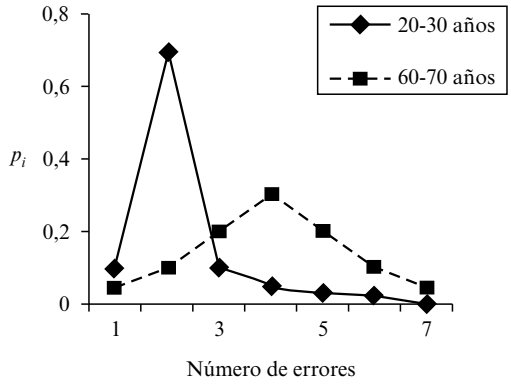
A: 8 6 5 4 2

B: 3 1 1 8 7

23. Una muestra de cien personas alcanza en la variable *rendimiento laboral* una varianza de 10 y en la variable *abandono* una varianza de 15. ¿Podemos afirmar que la variabilidad de este grupo es mayor en la variable *abandono*? Razone la respuesta.

24. Una muestra de cien varones obtuvo una varianza de 10 en la *escala de felicidad subjetiva* de Lyubomirsky (1999), mientras que una muestra de cien mujeres alcanzaron en la misma escala una varianza de 2. ¿Podemos afirmar que los varones varían mucho más que las mujeres en dicha escala? Razone su respuesta.

25. Medido el *número de errores* en la prueba de coordinación visual del test psicotécnico del carnet de conducir en una muestra aleatoria de diez conductores con edades comprendidas entre los 20 y 30 años y en otra de diez conductores con edades entre 60 y 70 años, se encuentra la siguiente distribución de frecuencias:



A la vista de los datos, y pese a no disponer de los estadísticos descriptivos univariados para cada grupo:

a) ¿Cuál de los grupos comete mayor número de errores?
b) ¿Cuál de las distribuciones es más homogénea?
c) Comente el grado de asimetría y curtosis en cada una de las distribuciones de la gráfica.

26. Calcule la varianza de los siguientes conjuntos de datos utilizando la fórmula [3.7]:

a) 8, 4, 5, 7, 5, 3, 5, 6, 2, 1.
b) 2, 8, 10, 5, 6, 1, 10, 9, 8, 20.
c) 31, 29, 30, 30, 30, 30, 28, 40, 35.

27. Calcule el dato que falta en el siguiente conjunto de datos: 6, 5, 8, X, 9. sabiendo que la media de la variable es 7 y el coeficiente de variación es 20,20. Resuélvalo tanto mediante la fórmula de la media [3.1] como mediante la fórmula alternativa de la varianza [3.7].

28. Obtenga el índice de asimetría de Pearson en los dos grupos siguientes de puntuaciones y compárelos en cuanto a esa propiedad:

A:	5	7	9	11	13
B:	5	7	9	10	9

29. Teniendo en cuenta que la media de una variable X es 12,5, la mediana 11 y la moda 7, indique, sin hacer ningún cálculo, qué tipo de asimetría esperaría observar en la distribución resultante.

30. Cuantifique la curtosis en los dos siguientes grupos de valores y compárelos en cuanto a esta propiedad:

A:	6	8	10	12	14
B:	5	9	10	11	5

31. Calcule el índice de asimetría para los datos del ejercicio 9 del capítulo 2 (conducta antinormativa en sujetos extravertidos e introvertidos).

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*) $\bar{X} = 8{,}3$.
b) $\bar{X} = -4{,}5$.
c) $\bar{X} = 23{,}25$.
d) $\bar{X} = 1{,}725$.
e) $\bar{X} = 26$.

2. $\overline{TR} = 248{,}308$.

3. Calculamos la media como índice que resume los hábitos de lectura: $\bar{X} = 1{,}533$.

Aunque para poder generalizar los resultados a la población de estudiantes se tendría que aplicar un procedimiento de inferencia estadística, el resultado permite hipotetizar que el nivel de lectura de la población de estudiantes de la Comunidad de Madrid es bajo, ya que en promedio no llegan a leer dos libros al año.

4. $\bar{X}_T = 6{,}25$

5. $\bar{X}_{Galicia} = 5$

6. $\bar{X}_T = 28{,}05$

7. Si se tienen J grupos, todos ellos con el mismo tamaño, se cumple que $N_1 = = N_2 = ... = N_J$. Denominemos al tamaño del grupo como N. El tamaño de todos los grupos juntos será: $J \cdot N = N_1 + N_2 + ... + N_J$. Aplicando la fórmula de la $\bar{X}_T$:

$$\bar{X}_T = \frac{N \cdot \bar{X}_1 + N \cdot \bar{X}_2 + ... + N \cdot \bar{X}_J}{N_1 + N_2 + ... + N_J} = \frac{N \cdot (\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_J)}{J \cdot N} =$$

$$= \frac{\bar{X}_1 + \bar{X}_2 + ... + \bar{X}_J}{J}$$

La última expresión es el promedio de las J medias.

8. *a*) $Mdn = 8$.
b) $Mdn = 31{,}5$.
c) $Mdn = 14$.
d) $Mdn = 18{,}5$.

9. Atendiendo a la distribución de frecuencias en el grupo de niños con *TGD* la moda es 8, ya que es el valor de la variable con mayor frecuencia absoluta; por tanto: $Mo_{TGD} = 8$. En el caso del grupo de niños con *DN*, el valor de la variable con mayor frecuencia absoluta es 5; por tanto: $Mo_{DN} = 5$. Recordando que la mediana es el C_{50}, se obtiene que:

Grupo	Mediana o C_{50}
TGD	7
DN	5

El valor de ambos estadísticos es superior en el grupo de niños con *TGD*, indicando que éstos muestran una mayor capacidad de cálculo. Además, en el caso de los niños con *DN*, la mediana y la moda coinciden.

10.

X_i	n_i	n_a	p_a
1	2	2	5
2	2	4	10
3	3	7	17,5
4	2	9	22,5
5	7	16	40
6	4	20	50
7	7	27	67,5
8	7	34	85
9	4	38	95
10	2	40	100

Recordando que la mediana es el C_{50}, $Mdn = 6$.

Recordando que la moda es el valor de la variable con mayor frecuencia absoluta, y atendiendo a la distribución, se observa que ésta tiene dos modas: $Mo_1 = 5$; $Mo_2 = 7{,}5$.

11. Si definimos a la diferencia entre antes (A) y después (D) como $E = = T_A - T_D$, entonces, $\bar{E} = \bar{T}_A - \bar{T}_D$; por tanto, $\bar{T}_D = 25$.

12. *a*) $\bar{E} = 35 - 25 = 10$.

b) Resumamos los datos en la siguiente tabla:

		Medias		
	Tamaño	**Antes**	**Después**	**Reducción**
Centro día	10	30	25	5
Residencia	20	35	25	10

Media total antes: $\bar{X}_T = 33{,}333$. Media total después: $\bar{X}_T = 25$. Media total de la reducción: $\bar{X}_T = 8{,}333$.

13. Hay infinitas soluciones. Se proponen dos:

a) 10 15 25 30 30 34 35 36 40 55.

b) 9 11 30 30 31 33 35 36 47 48.

14. Al ser una variable nominal, el índice de tendencia central adecuado es la moda. Atendiendo a la distribución de frecuencias, la moda es la asignatura Introducción a la psicología I (IP).

15. $\bar{X}_T = 3{,}124$.

16. $N_5 = 10$.

17. $\bar{X} = 3{,}925$; $S_X^2 = 1{,}419$; $S_X = 1{,}191$.

18. $\bar{X} = 2{,}833$; $S_X^2 = 1{,}806$.

19. Varones: $\bar{X}_V = 180{,}4$; $S_{X_V} = 6{,}25$. Mujeres: $\bar{X}_M = 166{,}2$; $S_{X_M} = 3{,}06$. Como cabía esperar, los varones son más altos que las mujeres. Sin embargo, la muestra de las mujeres es más homogénea.

20. $S_X^2 = 4$.

21. Extravertidos: $S^2_{X_{Extr}} = 8{,}40$. Introvertidos: $S^2_{X_{Intr}} = 8{,}79$.

Es más homogénea la muestra de sujetos extravertidos, aunque la diferencia es muy pequeña.

22. Medias parciales: $\bar{X}_A = 5$ y $\bar{X}_B = 4$. Varianzas parciales: $S^2_{X_A} = 4$ y $S^2_{X_B} = 8{,}80$. Media total: $\bar{X}_T = 4{,}5$

Nótese que este es el caso particular en que los tamaños muestrales son iguales ($N_A = N_B = 5$). Por tanto, también podríamos haber calculado la media total como el promedio de las medias. Es decir: $\bar{X}_T = \dfrac{5+4}{2} = 4{,}5$. Varianza total: $S^2_T = 6{,}65$.

23. No son, en principio, comparables, puesto que las dos variables vendrán dadas en unidades de medida distintas (por ejemplo, *rendimiento* en número de ventas y *abandono* en número de faltas en el trabajo).

24. Podríamos afirmarlo, a nivel descriptivo, si las características de los dos grupos son similares, en particular si las medias no difieren exageradamente.

25. *a*) Como se aprecia, cometen un mayor número de errores los conductores con edades comprendidas entre 60 y 70 años.

b) Es más homogénea la distribución de los conductores de 20-30 años. Obsérvese que la mayoría de estos sujetos (el 70 por 100) solamente cometen un error en la prueba, mientras que en la distribución de conductores de 60-70 años por término medio se cometen cuatro errores, pero hay varios sujetos (el 35 por 100) que cometen más de cuatro errores.

c) La distribución de conductores de 20-30 años presenta asimetría positiva y la de conductores con 60-70 años es simétrica. Adicionalmente, la distribución de conductores de 20-30 años presenta mayor apuntamiento o curtosis que la de conductores de 60-70 años, que es mesocúrtica.

26. La fórmula [3.7] es la fórmula alternativa de la varianza: $S^2_X = \dfrac{\sum X_i^2}{N} - \bar{X}^2$. Por tanto:

a) $S^2_X = 4{,}24$.

b) $S^2_X = 25{,}09$.

c) $S^2_X = 12{,}75$.

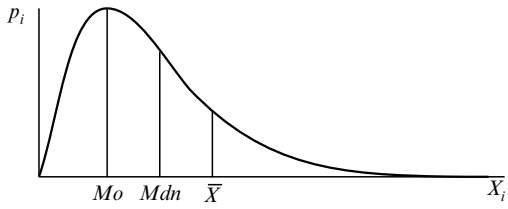
27. Como $CV = \dfrac{S_X}{\bar{X}} \cdot 100 = 20{,}20$ y $\bar{X} = 7$, se deduce que $S_X = 1{,}41$ y $S^2_X = 2$. Por tanto: $X = 7$ por ambas fórmulas.

28. *A*: $\bar{X}_A = 9$. $S_{X_A} = 2{,}83$. Puntuaciones diferenciales: $-4, -2, 0, 2, 4$. $As_{X_A} = 0$.

B: $\bar{X}_B = 8$. $S_{X_B} = 1{,}79$. Puntuaciones diferenciales: $-3, -1, 1, 2, 1$. $As_{X_B} = -0{,}63$.

Mientras la distribución del grupo *A* es simétrica, la del grupo *B* presenta asimetría negativa.

29. Como la *mediana* es menor que la *media,* y la *moda* es menor que la *mediana* y que la *media,* la distribución tendrá una forma similar a la siguiente. Es decir, probablemente presentará asimetría positiva.



30. Ambas distribuciones son platicúrticas, pues en las dos el índice de curtosis nos da negativo:

A: $\overline{X}_A = 10$. $S_{X_A} = 2{,}83$. Puntuaciones diferenciales: −4, −2, 0, 2, 4. $Cr_{X_A} = -1{,}30$.

B: $\overline{X}_B = 8$. $S_B = 2{,}53$. Puntuaciones diferenciales: −3, 1, 2, 3, −3. $Cr_{X_B} = -1{,}73$.

31. $As_{extravertidos} = 0{,}09$ y $As_{introvertidos} = 0{,}44$.

APÉNDICE

Fórmula alternativa de la varianza

Hemos definido la varianza (fórmula 3.5) como el promedio de las diferencias respecto a la media, elevadas al cuadrado. Desarrollamos el cuadrado del numerador de esa fórmula y separamos en tres quebrados:

$$S_X^2 = \frac{\sum(X_i - \bar{X})^2}{N} = \frac{\sum(X_i^2 + \bar{X}^2 - 2 \cdot X_i \cdot \bar{X})}{N} = \frac{\sum X_i^2}{N} + \frac{\sum \bar{X}^2}{N} - \frac{\sum 2 \cdot X_i \cdot \bar{X}}{N}$$

Como 2 y la media son constantes, aplicando las reglas del sumatorio llegamos a la fórmula [3.7].

$$S_X^2 = \frac{\sum X_i^2}{N} + \frac{N \cdot \bar{X}^2}{N} - \frac{2 \cdot \bar{X} \cdot \sum X_i}{N} = \frac{\sum X_i^2}{N} + \bar{X}^2 - 2 \cdot \bar{X}^2 = \frac{\sum X_i^2}{N} - \bar{X}^2$$

Índices alternativos de variación

No siempre se puede calcular la varianza, ni tampoco es siempre lo más apropiado. Aunque hay varias alternativas, con frecuencia específicas de campos muy concretos de la psicología (Botella, León, San Martín y Barriopedro, 2001), aquí vamos a señalar dos: la amplitud total y el coeficiente de variación.

Una forma muy sencilla de indicar el grado de variación de un conjunto de valores consiste en calcular la distancia entre el mayor y el menor de los valores observados. Este estadístico se llama *amplitud total* o *rango;* se obtiene sencillamente hallando la diferencia entre los valores extremos, es decir:

$$A_T = X_{\text{máx}} - X_{\text{mín}}$$

donde $X_{\text{máx}}$ y $X_{\text{mín}}$ representan los valores máximo y mínimo observados, respectivamente. A modo de ejemplo, la amplitud total de los valores de las tres variables del ejemplo del apartado 3.3.2 serían las siguientes: V: 6 − 1 = 5; X: 12 − 4 = 8; Y: 35 − 23 = 12.

A veces se desea comparar la variabilidad de dos o más grupos de valores. Comparar directamente sus varianzas puede ser inapropiado si esos grupos tienen medias muy diferentes. Para estos casos, Pearson (1896) propuso relativizar la desviación típica con respecto a la media. El estadístico así construido, expresado como un porcentaje, se denomina coeficiente de variación, se representa por sus iniciales (CV) y su fórmula es:

$$CV = \frac{S_X}{\bar{X}} \cdot 100$$

Además, este estadístico permite comparar la variabilidad de variables diferentes. También se puede considerar como un índice de la representatividad de la media. Cuanto mayor es el coeficiente de variación, menos representativa es la

media. Un ejemplo de este tipo de situaciones sería aquella en la que queremos comparar el grado de variabilidad en los tiempos empleados por dos grupos de varones y mujeres en correr los cien metros lisos. Los hombres son, en promedio, más rápidos que las mujeres, pero muchas veces una mayor media va acompañada de una mayor varianza. En este caso, se podrían comparar los coeficientes de variación en lugar de las varianzas. En el ejemplo numérico siguiente el grupo *A* tiene mayor varianza, pero como también tiene una media considerablemente mayor calculamos los coeficientes de variación. Al comparar éstos, concluimos que la variabilidad es mayor en el grupo *B*:

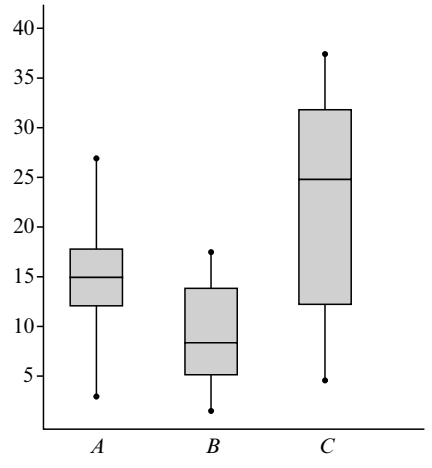
Grupo *A*: $\bar{X}_A = 20$ $S^2_{X_A} = 81$ $S_{X_A} = 9$ $CV_A = (9/20) \cdot 100 = 45{,}00$

Grupo *B*: $\bar{X}_B = 12$ $S^2_{X_B} = 49$ $S_{X_B} = 7$ $CV_B = (7/12) \cdot 100 = 58{,}33$

Representación gráfica de la variabilidad

Muchas veces interesa transmitir una idea directa y simple de la variabilidad observada en un conjunto de valores, reflejándola en una representación gráfica. Hay diversas formas de hacerlo, de las que aquí vamos a describir dos. En primer lugar, la técnica desarrollada por Tukey (1977) y denominada *box and whiskers,* que significa literalmente caja y bigotes; nosotros nos referiremos a ella simplemente como *diagrama de cajas*. Para su construcción se marcan señales, de tal forma que las distancias entre ellas sean proporcionales a las distancias entre la puntuación máxima, la mínima y los tres cuartiles. Con los tres cuartiles se forma una especie de ficha de dominó, mientras que la puntuación máxima y mínima se unen mediante líneas rectas a los bordes de esta forma geométrica. Se puede comparar la variabilidad de dos distribuciones haciendo representaciones paralelas de caja y bigotes, tal y como aparece en la siguiente figura, en la que se incluyen las representaciones de tres grupos de valores en una misma variable. Los estadísticos de los tres grupos son los que aparecen en la tabla de la izquierda:

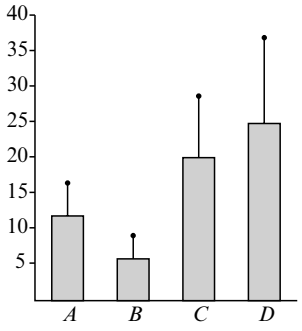
Estadístico:	*A*	*B*	*C*
Valor máximo	27	18	38
Valor mínimo	3	2	5
Q_1	12	5	12
Q_2	15	8	25
Q_3	18	14	32

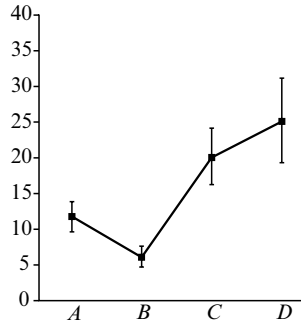


El dibujo conjunto de varios diagramas de cajas permite comparar gráficamente las distribuciones, tanto en su tendencia central como en sus niveles de variabilidad y sus grados de asimetría. Atendiendo a sus medianas, la muestra *C* presenta una mayor tendencia central, luego la *A* y por último la *B*. La muestra *C* parece bastante más heterogénea que las muestras *A* y *B*. En la muestra *A* se aprecia bastante simetría, mientras que la *B* parece asimétrica positiva y la *C* asimétrica negativa.

En otros casos se quiere representar la tendencia central de varios grupos de valores, pero se considera conveniente hacer alguna indicación sobre su grado de dispersión. En esos casos se pueden representar las medias mediante diagramas de barras o polígonos, pero añadiendo una línea (barras) o dos (polígono de frecuencias) verticales cuyo tamaño sea igual a la desviación típica. En el ejemplo siguiente tenemos las medias y desviaciones típicas de varios grupos (*A*, *B*, *C* y *D*), con el diagrama de barras y el polígono, con las líneas correspondientes a las desviaciones típicas.

Grupo	$\bar{X}$	S_X
A	12	4
B	6	3
C	20	8
D	25	12





Transformación de puntuaciones. Puntuaciones típicas y escalas derivadas

4

4.1. INTRODUCCIÓN

Con frecuencia resultará más útil trabajar con ciertas transformaciones de las puntuaciones que con las puntuaciones directas. Aunque ya hemos tratado la conversión de las puntuaciones directas en puntuaciones diferenciales, hay otras transformaciones aún más interesantes. En este capítulo vamos primero a exponer los efectos que la forma más habitual de transformación, las transformaciones lineales, tienen sobre los estadísticos más importantes: la media y la varianza. Conocer esos efectos nos permitirá justificar la definición del tipo de transformación más útil para nosotros, la transformación en *puntuaciones típicas,* así como sus principales propiedades. Por último, expondremos una transformación adicional de las puntuaciones típicas, que las convierten en las llamadas *escalas derivadas.*

4.2. MEDIA Y VARIANZA DE TRANSFORMACIONES LINEALES

Como ya hemos adelantado, con frecuencia realizaremos transformaciones lineales de las puntuaciones directas, por lo que nos interesa conocer los efectos de estas transformaciones sobre los estadísticos básicos. En concreto, supongamos una transformación lineal de las puntuaciones de la variable X en otras, que representaremos por Y, que se obtienen multiplicando las X_i por una constante (k) y sumando otra constante (c):

$$Y_i = k \cdot X_i + c \qquad [4.1]$$

Vamos a ver cómo la media y la varianza de las nuevas puntuaciones así transformadas, Y, se pueden obtener directamente conociendo la media y la varianza de las puntuaciones originales y los valores de las constantes empleadas en la transformación (k y c). No hace falta decir que si transformamos una a una las puntuaciones y luego aplicamos a esos valores las fórmulas de la media y la varianza que expusimos en el capítulo anterior, [3.1] y [3.5], el resultado será correcto. Sin embargo, con frecuencia aprovecharemos lo que vamos a exponer a

continuación para hacer demostraciones o para hacer cálculos abreviados de esos estadísticos.

Veamos primero el caso en que la transformación consiste sólo en sumar una constante (es decir, el caso en que $k = 1$ y c es cualquier valor distinto de 0). La media de las nuevas puntuaciones es, por definición, su suma dividida por su número. Sustituyendo el valor de cada nueva puntuación (Y_i) por el valor que lo define ($X_i + c$), llegamos a lo siguiente:

$$\bar{Y} = \frac{\sum Y}{N} = \frac{\sum(X + c)}{N} = \frac{\sum X}{N} + \frac{N \cdot c}{N} = \bar{X} + c \qquad [4.2]$$

Veamos lo que ocurre con la varianza:

$$S_Y^2 = \frac{\sum(Y - \bar{Y})^2}{N} = \frac{\sum[(X + c) - (\bar{X} + c)]^2}{N} =$$
$$= \frac{\sum[(X - \bar{X} + c - c)]^2}{N} = \frac{\sum[(X - \bar{X})]^2}{N} = S_X^2 \qquad [4.3]$$

Esto significa que *si sumamos la constante c a los N valores, su media se ve incrementada en esa misma constante [4.2], mientras que su varianza no se altera [4.3].*

Veamos ahora el caso en que la transformación consiste sólo en multiplicar los valores por una constante (es decir, el caso en que $c = 0$ y k es cualquier valor distinto 0 ó 1). Aplicando la misma lógica, obtenemos lo siguiente para la media:

$$\bar{Y} = \frac{\sum Y}{N} = \frac{\sum(k \cdot X)}{N} = \frac{k \cdot \sum X}{N} = k \cdot \bar{X} \qquad [4.4]$$

Con respecto a la varianza, por comodidad sustituiremos en la fórmula [3.7]:

$$S_Y^2 = \frac{\sum Y^2}{N} - \bar{Y}^2 = \frac{\sum(k \cdot X)^2}{N} - (k \cdot \bar{X})^2 = \frac{k^2 \cdot \sum X^2}{N} - k^2 \cdot \bar{X}^2 =$$
$$= k^2 \cdot \left(\frac{\sum X^2}{N} - \bar{X}^2\right) = k^2 \cdot S_X^2 \qquad [4.5]$$

Esto significa que *si multiplicamos por la constante k a los N valores, su media se ve multiplicada por esa constante [4.4] y su varianza por el cuadrado de esa constante [4.5].*

Aunque por razones didácticas hemos expuesto por separado los efectos de la suma y producto de constantes, en la práctica lo más frecuente es que las transformaciones impliquen simultáneamente valores de k distintos de 1 y valores de c distintos de 0. Podemos resumir estas propiedades de la siguiente forma:

Si se transforman linealmente unas puntuaciones, X, llamando Y a las puntuaciones transformadas (según [4.1]), la media y la varianza de las nuevas puntuaciones transformadas serán:

$$\bar{Y} = k \cdot \bar{X} + c \qquad [4.6]$$

$$S_Y^2 = k^2 \cdot S_X^2 \qquad [4.7]$$

Veamos un ejemplo numérico que ilustre lo expuesto hasta aquí con una transformación lineal completa. Supongamos una muestra de cuatro valores de la variable X, para los que hemos hallado su media y su varianza:

$$X_1 = 5, \quad X_2 = 4, \quad X_3 = 6, \quad X_4 = 1$$

$$\bar{X} = 4, \quad S_X^2 = 3{,}5$$

Supongamos que transformamos estos valores en otras puntuaciones, Y, mediante la transformación lineal:

$$Y_i = 2 \cdot X_i + 5$$

Las nuevas puntuaciones, una vez transformadas, serán: $Y_1 = 15$; $Y_2 = 13$; $Y_3 = 17$; $Y_4 = 7$. Si no conociéramos las fórmulas [4.2] a [4.5] o su expresión conjunta de [4.6] y [4.7], aplicaríamos las fórmulas de la media y la varianza a estos cuatro valores y obtendríamos el resultado correcto. Sin embargo, conociendo esas fórmulas podemos obtener más fácilmente los estadísticos de la variable Y:

$$\bar{Y} = 2 \cdot \bar{X} + 5 = 2 \cdot 4 + 5 = 13$$

$$S_Y^2 = 2^2 \cdot S_X^2 = 4 \cdot 3{,}5 = 14$$

4.3. PUNTUACIONES TÍPICAS

Dado que el valor observado en un individuo, sujeto o unidad de investigación representa la magnitud que manifiesta en la variable, una práctica común consistirá en comparar valores asociados a ellas. Sin embargo, ya vimos en capítulos anteriores que la comparación entre puntuaciones directas puede llevarnos a conclusiones engañosas. Ya hemos expuesto una solución, no del todo satisfactoria, que consiste en obtener las distancias a la media, o puntuaciones diferenciales. En este apartado y en el siguiente vamos a retomar este problema y a exponer mejores soluciones. Éstas se basan en aplicar transformaciones lineales de las puntuaciones observadas, produciendo otras que, sin perder o distorsionar la información contenida en las puntuaciones originales, permiten una comparación

más eficiente de las mismas. En concreto, en este apartado vamos a exponer las puntuaciones típicas y en el próximo las llamadas escalas derivadas.

Si se nos informa sólo de que un individuo obtiene una puntuación igual a 50 al evaluarle en una variable, y queremos hacer una valoración de este dato, podemos encontrarnos con la dificultad de carecer de referencias apropiadas para hacerlo. En el capítulo anterior hemos visto una forma de abordar el problema, calculando para este individuo lo que llamábamos como puntuación diferencial, y que no es más que la distancia, o diferencia, entre esa puntuación y la media del grupo de puntuaciones. Las puntuaciones diferenciales son sin duda más informativas e interesantes que las directas, pues al menos nos indican si la puntuación es superior o inferior a la media o si coincide con ella (según el signo de la puntuación diferencial). Sin embargo, esta información es insuficiente para comparar puntuaciones de personas pertenecientes a distintos grupos o que miden variables diferentes. Esto es fácil de ver en casos en los que dos sujetos de diferentes grupos o poblaciones obtienen la misma puntuación diferencial, pero a pesar de ello sus valores representan cosas bien diferentes. Supongamos que Pedro, perteneciente al grupo *A*, y Pablo, perteneciente al grupo *B*, obtienen la puntuación 50. Sabemos, además, que en ambos grupos la media es 40. Las puntuaciones relativas de ambos individuos con respecto a sus grupos son aparentemente las mismas, pues ambos tienen la misma puntuación diferencial:

$$x = 50 - 40 = 10$$

Sin embargo, esta igualdad en las puntuaciones diferenciales puede estar ocultando realidades bien distintas. Supongamos, por ejemplo, que los histogramas correspondientes a estos grupos son los que aparecen en la figura 4.1. Como se puede observar, las puntuaciones del grupo *A* son bastante más homogéneas que las del grupo *B*. Esto hace que, mientras que la puntuación 50 en el grupo *A* representa uno de los valores más altos (sólo los pocos valores 55 le superan), esa misma puntuación en el grupo *B* no es tan extrema; aun estando por encima de la media, parece ser una puntuación más bien cercana al Q_3. La diferencia entre los grupos *A* y *B* de la figura 4.1 estriba en su grado de variabilidad, mayor en el grupo *B* que en el grupo *A*. De hecho, la desviación típica del grupo *A* es 5, mientras que la del grupo *B* es 10.

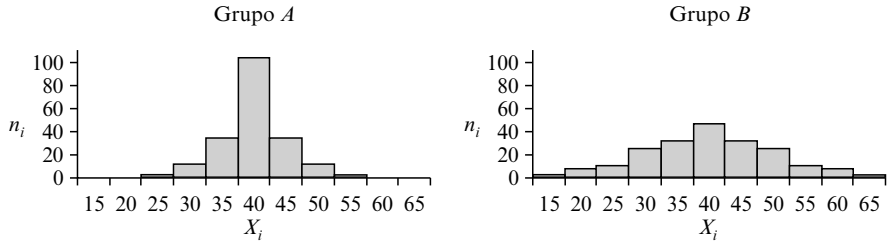


Figura 4.1.—Histogramas de dos distribuciones de frecuencias con distinta variabilidad.

Una solución a este problema de interpretación consiste en que en lugar de medir las distancias a la media en términos absolutos se haga con relación a la variabilidad del grupo de referencia. Se trataría de indicar cómo de grande es una distancia en términos de las distancias observadas en general en esas puntuaciones. Esa distancia general la habíamos medido en el capítulo anterior mediante la desviación típica; podemos utilizar ésta como unidad de medida. Las puntuaciones así conseguidas se denominan *puntuaciones típicas*, se representan por letras z minúsculas y su fórmula es:

$$z_i = \frac{X_i - \bar{X}}{S_X} \qquad [4.8]$$

También se pueden expresar en términos de las puntuaciones diferenciales:

$$z_i = \frac{x_i}{S_X} \qquad [4.9]$$

No siempre se ha utilizado la desviación típica como módulo divisor; en el siglo XIX se solía dividir por $S_X \cdot \sqrt{2}$, mientras que Galton prefería utilizar lo que se llamaba el «error probable» (Walker, 1975). Ya entrado el siglo XX se comenzó a utilizar S_X como denominador (Sheppard, 1902); sólo en 1914 se empezó a utilizar la expresión medida típica (Kelley, 1914). Actualmente es ya universal el uso de S_X como denominador.

Al proceso de obtención de las puntuaciones típicas se le llama *tipificación*. Así, Pedro, que pertenecía al grupo *A* en el ejemplo anterior, tenía una puntuación directa de 50, mientras que la media y la desviación típica del grupo eran 40 y 5, respectivamente. Por tanto, su puntuación diferencial era 10 y su puntuación típica es:

$$z_i = \frac{50 - 40}{5} = 2$$

Por el contrario, la tipificación de la puntuación de Pablo nos da:

$$z_i = \frac{50 - 40}{10} = 1$$

Como se puede apreciar, la diferencia entre ambos individuos en relación al grupo al que pertenecen queda perfectamente reflejada en estos valores, mientras que en las puntuaciones diferenciales no quedaba de manifiesto: Pedro se separa de la media de su grupo en dos desviaciones típicas por encima de ésta, mientras que Pablo sólo se separa una desviación típica de la media del suyo. La definición de las puntuaciones típicas, de hecho, se basa en esta idea y se puede expresar diciendo lo siguiente:

> La puntuación típica de una observación indica el número de desviaciones típicas que esa observación se separa de la media de su grupo.

Las puntuaciones típicas permiten, por tanto, hacer comparaciones entre unidades de distintos grupos, entre variables medidas de distintas formas o incluso entre variables diferentes. En todos los casos las puntuaciones típicas siempre nos indicarán el número de desviaciones típicas (de las de ese grupo y del variable) que cada valor se separa de la media (de ese grupo y del variable), y si se trata de una desviación por encima o por debajo de la media (según el signo de la puntuación típica). Esta transformación es de suma utilidad, pues se traduce en que las puntuaciones típicas tienen unas características de tendencia central y variabilidad constantes, tal y como vamos a ver a continuación al deducir su media y su varianza.

La media y la varianza de las puntuaciones típicas se deducen fácilmente a partir de lo expuesto en el apartado anterior si la fórmula de éstas se expresa en el formato de las transformaciones lineales de [4.1]. En concreto, las puntuaciones típicas no son más que una transformación lineal; consiste en multiplicar las directas por una constante (el inverso de la desviación típica: $k = 1/S_X$) y luego sumar a esos productos otra constante (el cociente entre la media y la desviación típica, con signo negativo: $c = -\bar{X}/S_X$):

$$z_i = \frac{X_i - \bar{X}}{S_X} = \frac{1}{S_X} \cdot X_i - \frac{\bar{X}}{S_X}$$

Aplicando lo que hemos aprendido en el apartado anterior sobre la media y la varianza de las transformaciones lineales, podemos deducir la media y la varianza de las puntuaciones típicas (fórmulas [4.6] y [4.7]):

$$\bar{z} = \frac{1}{S_X} \cdot \bar{X} - \frac{\bar{X}}{S_X} = 0 \qquad [4.10]$$

$$S_z^2 = \left(\frac{1}{S_X}\right)^2 \cdot S_X^2 = 1 \qquad [4.11]$$

Estas características de las puntuaciones típicas son universales; no dependen del tipo de puntuaciones directas, ni de su dispersión, ni de su número. Su universalidad es precisamente su razón de ser; por ello lo destacamos especialmente:

> Las puntuaciones típicas de un grupo de puntuaciones siempre tienen como media 0 y como desviación típica 1.

Las puntuaciones típicas reflejan, en cierto sentido, las relaciones esenciales entre las puntuaciones, con independencia de la unidad de medida que se haya utilizado en la medición. Por eso, cuando en dos conjuntos de puntuaciones, emparejadas con algún criterio, a los elementos de cada par les corresponde la misma puntuación típica dentro de su conjunto, se puede decir que mantienen la misma estructura interna; se dice entonces que son *puntuaciones equivalentes*.

Expresado más formalmente, diríamos que dos conjuntos de N puntuaciones, X_i e Y_i, son equivalentes si sus puntuaciones típicas son iguales:

$$\frac{X_i - \bar{X}}{S_X} = \frac{Y_i - \bar{Y}}{S_Y} \qquad (i = 1, 2, 3, ..., N) \qquad [4.12]$$

Debido a sus atractivas propiedades, las puntuaciones típicas se utilizan en la definición de muchos estadísticos y en la aplicación de muchas técnicas y procedimientos. Recordemos, por ejemplo, las fórmulas que hemos propuesto en el capítulo 3 para valorar la asimetría y la curtosis. En ellas aparecían expresiones en las que ahora podemos reconocer a las puntuaciones típicas. De hecho, las fórmulas de la asimetría y la curtosis, [3.14] y [3.15], se basan en el promedio de las puntuaciones típicas elevadas al cubo y la cuarta potencia.

4.4. ESCALAS DERIVADAS

A pesar de que las puntuaciones típicas tienen las indudables ventajas que hemos mencionado, también tienen algún inconveniente. Así, dado que la media de las típicas es cero y su desviación típica uno, aproximadamente la mitad de las puntuaciones son negativas y casi todas decimales. Esto hace que resulte incómodo su tratamiento y que muchas veces se busquen procedimientos que permitan superar esta dificultad. Un procedimiento consiste en transformar a su vez las puntuaciones típicas en otras que retengan todas las relaciones que manifiestan las puntuaciones originales y que, por tanto, sean puntuaciones equivalentes, pero evitando las dificultades de cálculo y tratamiento. Estas nuevas puntuaciones constituyen lo que se denominan *escalas derivadas*.

Estas transformaciones se fundamentan en la siguiente propiedad de las puntuaciones típicas. Supongamos que disponemos de un conjunto de N puntuaciones directas, X_i, que las transformamos en puntuaciones típicas, z_i, y que después volvemos a transformar las típicas en otras, que llamaremos T_i, mediante la transformación lineal siguiente, en la que a y b son dos constantes:

$$T_i = a \cdot z_i + b \qquad [4.13]$$

Apliquemos de nuevo lo que hemos aprendido en el apartado 4.1 sobre las transformaciones lineales, sabiendo, además, que la media y la varianza de las puntuaciones z son siempre 0 y 1, respectivamente. La media y la varianza de estas nuevas puntuaciones T serán:

$$\bar{T} = a \cdot \bar{z} + b = a \cdot 0 + b = b$$

$$S_z^2 = a^2 \cdot S_z^2 = a^2 \cdot 1 = a^2 \qquad [4.14]$$

Naturalmente, la desviación típica será igual a la constante a tomada en valor absoluto ($|a|$). Como se puede apreciar, dado que podemos elegir las constantes a y b que nos parezcan más apropiadas para cada caso, podemos conseguir trans-

formaciones en escalas derivadas que tengan la media y la varianza que nos parezcan más cómodas para trabajar. Por tanto, *si transformamos linealmente las puntuaciones típicas, multiplicándolas por una constante a y sumando una constante b, entonces las puntaciones transformadas tendrán como media la constante sumada, b, como desviación típica el valor absoluto de la constante multiplicada, $|a|$, y como varianza el cuadrado de esta constante, a^2.*

En resumen, la construcción de una escala derivada parte de unas puntuaciones directas, éstas se tipifican y después se transforman linealmente en otras puntuaciones, según aparece en el esquema de la figura 4.2.

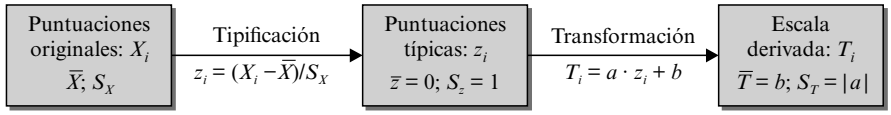


Figura 4.2.—Esquema del proceso de transformación de unas puntuaciones en una escala derivada, pasando por la tipificación como paso intermedio.

Este proceso se podría realizar en un solo paso si la transformación se realiza según la expresión (siendo *a* una constante positiva):

$$T_i = \frac{a}{S_X} \cdot X_i + \left(b - \frac{a \cdot \overline{X}}{S_X} \right)$$

En este caso, la media y la varianza de estas puntuaciones serán:

$$\overline{T} = \frac{a}{S_X} \cdot \overline{X} + \left(b - \frac{a \cdot \overline{X}}{S_X} \right) = b + \frac{a}{S_X} \cdot \overline{X} - \frac{a}{S_X} \cdot \overline{X} = b$$

$$S_T^2 = \left(\frac{a}{S_X} \right)^2 \cdot S_X^2 = a^2$$

Veamos un ejemplo numérico de esta doble transformación. En la tabla siguiente aparecen en la primera columna los cinco valores originales, en la variable *X*, con su media y su desviación típica. En la columna central aparecen sus puntuaciones típicas, *z*, que naturalmente tienen media 0 y desviación típica 1. Por último, éstas se transforman en otras puntuaciones, T_i, multiplicando las típicas por 3 y sumando 10; naturalmente, estas últimas tienen media 10 y desviación típica 3.

X_i	z_i	T_i
2	$z_1 = (2 - 5)/2 = -1{,}5$	$T_1 = 3 \cdot z_1 + 10 = 5{,}5$
5	$z_2 = (5 - 5)/2 = 0$	$T_2 = 3 \cdot z_2 + 10 = 10{,}0$
6	$z_3 = (6 - 5)/2 = 0{,}5$	$T_3 = 3 \cdot z_3 + 10 = 11{,}5$
4	$z_4 = (4 - 5)/2 = -0{,}5$	$T_4 = 3 \cdot z_4 + 10 = 8{,}5$
8	$z_5 = (8 - 5)/2 = 1{,}5$	$T_5 = 3 \cdot z_5 + 10 = 14{,}5$
$\overline{X} = 5$ $S_X = 2$	$\overline{z} = 0$ $S_z = 1$	$\overline{T} = 10$ $S_T = 3$

Recordemos que la cuestión fundamental de las escalas derivadas consiste en transformar las puntuaciones originales, X_i, en otras puntuaciones transformadas, T_i, tales que sean más cómodas de tratar e interpretar, pero que a la vez retengan las relaciones esenciales entre los valores; es decir, que sean puntuaciones equivalentes. Sólo nos queda, por tanto, demostrar que las puntuaciones X_i y T_i son equivalentes.

Dado que cualquier puntuación T_i se obtiene a partir de una puntuación z_i mediante la expresión [4.13] y que las puntuaciones T_i tienen media b y desviación típica $|a|$, si en la expresión [4.13] despejamos z_i constatamos que la tipificación de cada puntuación T_i da igual al valor z_i correspondiente a la X_i de la que procede. Es decir:

$$\frac{T_i - b}{|a|} = z_i = \frac{X_i - \bar{X}}{S_X}$$

y, por tanto, queda demostrado que son puntuaciones equivalentes. De hecho, cualquier transformación lineal en la que la constante multiplicadora sea positiva da lugar a unas puntuaciones equivalentes. Por ejemplo, si transformamos las puntuaciones X_i según la expresión:

$$Y_i = k \cdot X_i + c$$

entonces las puntuaciones Y_i tendrán como media $(k \cdot \bar{X} + c)$ y como desviación típica $(|k| \cdot S_X)$. La puntuación típica de cualquier puntuación X_i será:

$$\frac{X_i - \bar{X}}{S_X} = z_i$$

mientras que la típica de su transformada será:

$$\frac{Y_i - \bar{Y}}{S_Y} = \frac{(k \cdot X_i + c) - (k \cdot \bar{X} + c)}{k \cdot S_X} = \frac{k \cdot (X_i - \bar{X}) + c - c}{k \cdot S_X} = \frac{X_i - \bar{X}}{S_X}$$

y, por tanto, las X_i e Y_i son puntuaciones equivalentes.

Aunque los valores que se pueden utilizar como constantes a y b para construir escalas derivadas pueden ser cualesquiera, ha sido frecuente en la historia de la psicología utilizar ciertas constantes concretas, quizá con la intención de generalizar su uso y llegar incluso a hacerlas universales. Si así hubiera sido, nos encontraríamos ahora con que todas las escalas tendrían las mismas características, facilitando su interpretación. Por ejemplo, si las puntuaciones en cualquier instrumento de medición se transformasen en una escala con media 500 y desviación típica 100, bastaría con informar de la propia puntuación para poder valorarla en términos relativos. Así, se han propuesto entre otras las siguientes escalas con nombre propio:

$$T_i = 10 \cdot z_i + 50$$

$$S_i = 2 \cdot z_i + 5$$

Lamentablemente, estas iniciativas no triunfaron, y ahora las diferentes escalas tienen sus características particulares (véase el apéndice del capítulo presente). De todas estas transformaciones, la más conocida es la que convierte las puntuaciones directas en inteligencia en una escala derivada de *CI* (cociente intelectual), que son puntuaciones en las que la media es igual a 100 y la desviación típica es igual a 15:

$$CI_i = 15 \cdot z_i + 100$$

PROBLEMAS Y EJERCICIOS

1. Se ha aplicado a una muestra de 45 participantes un cuestionario que mide la dependencia de campo (variable X). Cada uno de los valores de X ha sido multiplicado por una constante igual a 4 y se le ha sumado una constante igual a 20, obteniéndose la variable Y. Sabiendo que $\overline{X} = 5$ y $S_X^2 = 4$, obtenga la media y varianza de la variable Y.

2. Si a cada uno de los valores de la variable X:

a) Se le suma una constante igual a −4, ¿cuánto valdrá la nueva media sabiendo que la media de X es 15?

b) Si multiplicáramos los valores originales de X por 0,5, ¿qué valor tomaría la nueva media de X?

3. Para evitar los números negativos de unos datos se añade la constante 80, tras haberlos multiplicado por la constante 100 para evitar a su vez los números decimales. La media aritmética resultante tras la transformación ha sido 40. ¿Cuál es el valor de la media aritmética de los datos originales?

4. Sabiendo que $S_X^2 = 36$, obtenga la varianza y desviación típica para cada una de las transformaciones propuestas en el ejercicio 2.

5. A una muestra de 130 participantes se les aplicó una prueba de estimación de la distancia percibida a un blanco estático. Sabiendo que la distancia real a la que se encontraba el blanco era igual a 3 metros y que la media de la distancia estimada por los participantes era igual a 2,85 metros, obtenga la media de los errores de estimación cometida por los participantes. ¿Cuál sería le media de los errores si la estimación se hubiera hecho en centímetros?

6. Siguiendo con el ejercicio anterior, sabiendo que la varianza de la estimación fue igual a 2, obtenga la varianza de los errores de estimación cuando se hizo en metros. ¿Qué valor tendría la varianza de los errores si estos se hubieran tomado en centímetros?

7. Se ha medido en una muestra de 75 participantes el tiempo de respuesta, TR, en segundos que se tarda en responder a la tarea de Wisconsin. Los resultados se han resumido en los siguientes estadísticos: $\overline{TR} = 3{,}2$ y $S_{TR}^2 = 0{,}810$. ¿Cuanto valdrían la media y la varianza de los TR si se hubieran medido en décimas de segundo?

8. A todos los valores de una variable X obtenidos en una muestra se les ha sumado una constante c desconocida. Sabiendo que el valor de la media de X antes de sumar la constante era igual a 13 y después de la suma es 7,6, obtenga el valor de la constante c.

9. A los valores de una variable X se les aplicó una transformación lineal, formando la nueva variable Y, donde las constantes fueron mayores que cero. Se sabe que: la media de X es igual a 8 y su varianza es 4; la media de Y es 31 y su desviación típica es 6. Obtenga el valor de las constantes.

10. Tras aplicar una prueba que mide competencia lingüística en inglés a un grupo de estudiantes, se obtuvo que la media del grupo era igual a 6,3 y la varianza 3,24. Tras revisar las puntuaciones se observó que la puntuación dada a cada estudiante es un 10 por 100 inferior a la que realmente tendrían que haber obtenido. Obtenga la media y varianza de las puntuaciones corregidas.

11. Si las puntuaciones típicas se multiplican por una constante igual a 4 y se les suma una constante igual a 7, formándose la nueva variable Y, ¿cuáles serían la media, varianza y desviación típica de la variable Y?

12. Si transformamos las puntuaciones directas de un test X en las puntuaciones Y del siguiente modo: $Y = 5 \cdot X + 20$, ¿serán iguales las puntuaciones típicas de X y las de Y?

13. A dos niños, AP y MR, se les evalúa con una prueba de capacidad psicomotriz. Si la puntuación directa de AP es mayor que la de MR, ¿también lo será su puntuación típica?

14. Dos estudiantes, Luis y Álvaro, hacen dos exámenes parciales, X e Y, de matemáticas. A continuación conteste a las siguientes preguntas:

a) Si Luis obtiene mayor puntuación directa en el primer parcial (X) que en el segundo (Y), ¿también obtendrá mayor puntuación típica en el primer parcial (X)?

b) Si Luis obtiene en el primer parcial (X) una puntuación típica mayor que Álvaro, ¿también tendrá un centil mayor?

c) Si Álvaro obtiene una puntuación típica en el primer parcial (X) mayor que en el segundo (Y), ¿también tendrá un centil mayor en X que en Y?

15. Obtenidas las puntuaciones en un test de aptitud numérica (X), se transforman las puntuaciones originales en una escala derivada con media 50 y varianza 100. Dos participantes, Jaime y Pedro, hacen la prueba: en la escala derivada Jaime recibe un valor de 75 y Pedro de 45. A partir de estos datos, calcule sus respectivas puntuaciones típicas.

16. Se sabe que el cociente intelectual de un sujeto en una prueba espacial es 105. ¿Cuál es la puntuación de ese sujeto en una nueva escala derivada de la misma prueba espacial que tiene media 10 y desviación típica 2?

17. Supongamos que 30 estudiantes de primer curso de psicología pasan dos pruebas de atención (X e Y). Las medias y las desviaciones típicas de ambas prue-

bas y las puntuaciones directas obtenidas por los estudiantes *A* y *B* aparecen en la tabla siguiente:

	Media	Desviación típica	Puntuaciones de *A*	Puntuaciones de *B*
X	50	12,25	45	60
Y	23	3,54	20	30

a) Calcule la puntuación típica de ambos estudiantes en cada una de las pruebas.

b) Calcule las puntuaciones de ambos estudiantes para la prueba *X* en una escala derivada con media 10 y desviación típica 5.

c) Globalmente, ¿cuál de los dos estudiantes tiene mejor rendimiento en las pruebas de atención?

18. Tras pasar a un grupo de 35 ejecutivos un test de conservadurismo (*C*) y otro de autoritarismo (*A*), las puntuaciones obtenidas tienen los estadísticos descriptivos que se muestran en la tabla siguiente:

	Media	Desviación típica
C	40	15
A	70	20

a) Si Cristina obtuvo en *C* una puntuación igual a 64 y en *A* una puntuación de 102, ¿en cuál de los dos tests obtuvo mejor puntuación?

b) Si transformamos las puntuaciones de la variable *C* en otras de una escala derivada con media 30 y desviación típica 1,5, ¿cuál será la nueva puntuación de Cristina?

19. Tras tipificar ocho puntuaciones, hemos perdido una de ellas. Calcule cuánto vale la puntuación típica perdida y explique por qué, sabiendo que las otras son: −1,30; −1,15; 0,25; 1,40; 2,40; −0,90; −0,55.

20. Transforme los grupos de puntuaciones del ejercicio 26 del capítulo 3 en puntuaciones típicas.

21. Medida la variable *X*, sabemos que su media es 10 y su desviación típica 6. A continuación, conteste a las siguientes preguntas:

a) ¿Qué puntuación típica tiene un sujeto que tuvo una directa igual a 12?

b) ¿Qué puntuación directa tiene un sujeto con una puntuación típica igual a −1,5?

c) Si transformamos los datos en una escala derivada con media 30 y desviación típica 9, un sujeto que obtuvo una directa de 7, ¿qué valor tiene en la escala derivada?

d) Si tras la transformación del apartado anterior un sujeto queda con la puntuación 22,5, ¿cuál era su puntuación original en X?

22. Sabiendo que la variable X tiene como media 40 y como desviación típica 10, complete la tabla inferior con las puntuaciones directas, diferenciales y típicas de los cuatro sujetos que en ella aparecen, así como las puntuaciones en una escala derivada con media 100 y desviación típica 20.

Sujeto	Puntuación directa, X_i	Puntuación diferencial, x_i	Puntuación típica, z_i	Escala derivada, T_i
1	60			
2		−5		
3			−1,50	
4				145

23. En una prueba de percepción sensorial, cuyas puntuaciones tienen una distribución simétrica con media 5 y desviación típica 3, se sabe que Raquel obtiene una puntuación típica de 0,5, Victoria ocupa el centil 40 y Joaquín obtiene una puntuación directa de 6 puntos. A partir de estos datos, conteste quién de los tres obtuvo la mayor puntuación.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. $\bar{Y} = 40$; $S_Y^2 = 64$.

2. *a*) $\bar{Y} = 11$.
b) $\bar{Y} = 7{,}5$.

3. $\bar{X} = -0{,}40$.

4. *a*) $S_Y^2 = 36$ y $S_Y = 6$. Sumar una constante a los valores de una variable no modifica el valor de la varianza ni el de la desviación típica de la variable original.
b) $S_Y^2 = 9$; $S_Y = 3$.

5. Si definimos $E_i = X_i - 3$, entonces:
a) $\bar{E} = -0{,}15$ metros.
b) $\bar{E} = -15$ centímetros.

6. *a*) $S_E^2 = 2$.
b) $S_E^2 = 20.000$.

7. $\bar{Y} = 32$; $S_Y^2 = 81$.

8. $c = -5{,}4$.

9. $k = 3$; $c = 7$.

10. $\bar{X} = 7$; $S_X^2 = 4$.

11. $\bar{Y} = 7$; $S_Y^2 = 16$; $S_Y = 4$.

12. Sí, pues las puntuaciones transformadas (Y) serán equivalentes a las originales (X).

13. Sí, porque la puntuación típica es una mera transformación lineal creciente de la puntuación directa, y en este caso la transformación sería idéntica para ambas puntuaciones directas. Les restaríamos la misma media y dividiríamos por la misma desviación típica.

14. *a*) No necesariamente; dependerá de las medias y de las desviaciones típicas de cada examen.
b) Sí, pues en una misma distribución, a mayor puntuación típica mayor porcentaje de sujetos dejará por debajo de sí.
c) No necesariamente; dependerá de los estadísticos y de la forma de la distribución de cada variable.

15. Jaime: $z_i = 2{,}5$; Pedro: $z_i = -0{,}50$.

16. $T = 10{,}67$.

17. *a*) Prueba X: $z_A = -0{,}41$; $z_B = 0{,}82$. Prueba Y: $z_A = -0{,}85$; $z_B = 1{,}98$.
b) Prueba X: $T_A = 7{,}95$; $T_B = 14{,}1$.
c) El estudiante B rinde mejor, pues sus típicas son mayores que las del estudiante A; este último puntúa por debajo de la media de su grupo tanto en la prueba X como en la prueba Y.

18. *a*) Obtuvo la misma puntuación en ambos tests (una típica de 1,6).
b) $T = 32{,}4$.

19. La puntuación típica que nos falta es -0,15.

20. *a*) 1,65; −0,29; 0,19; 1,17; 0,19; −0,78; 0,19; 0,68; −1,26; −1,75.
b) −1,18; 0,02; 0,42; −0,58; −0,38; −1,38; 0,42; 0,22; 0,02; 2,42.
c) −0,12; −0,68; −0,40; −0,40; −0,40; −0,40; −0,96; 2,40; 1,00.

21. *a*) $z_i = 0{,}33$
b) $X_i = 1$
c) $T_i = 25{,}5$
d) $X_i = 5$.

22.

Sujeto	Puntuación directa, X_i	Puntuación diferencial, x_i	Puntuación típica, z_i	Escala derivada, T_i
1	60	20	2	140
2	35	−5	−0,50	90
3	25	−15	−1,50	70
4	62,5	22,5	2,25	145

23. Raquel.

APÉNDICE

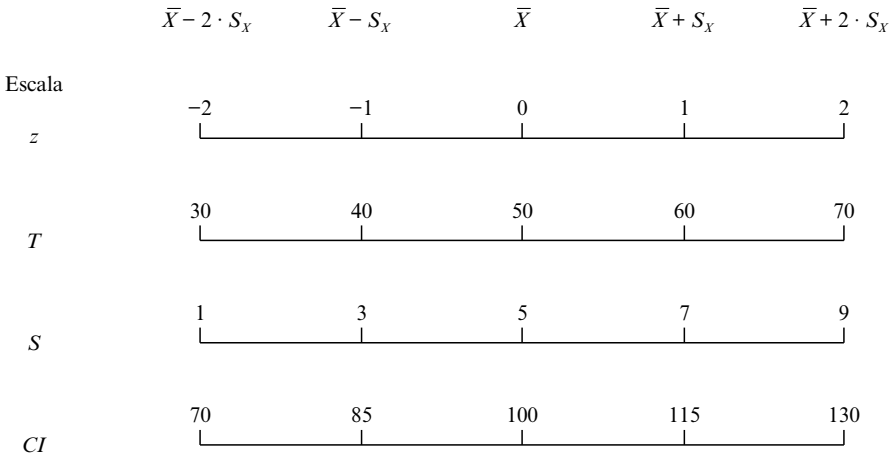
Sobre las escalas derivadas

Las puntuaciones *T* tienen media 50 y desviación típica 10; es la transformación general más conocida. También son bastante conocidas las puntuaciones *S*, o de *estaninos,* que tienen media 5 y desviación típica 2, o las que se desarrollaron en el ejército norteamericano para sus pruebas de clasificación, que tienen media 100 y desviación típica 20. Sin embargo, la transformación más conocida es la del cociente intelectual o *CI*, que se refiere a la medición de la inteligencia; tiene media 100 y desviación típica 15. El término «cociente intelectual» no tiene mucho que ver con lo que se mide, pero se ha mantenido por tradición histórica. Concretamente, Stern (1912) propuso obtener la razón entre la edad mental y la edad cronológica, multiplicando luego por 100:

$$CI = \frac{\text{Edad mental}}{\text{Edad cronológica}} \cdot 100$$

Si el desarrollo mental de un niño coincide con su edad cronológica, su *CI* sería igual a 100; por el contrario, valores de CI superiores o inferiores a 100 indicarían adelantos o retrasos, respectivamente, con respecto al desarrollo mental medio de los niños de su edad.

En cualquier caso, existe una equivalencia entre las puntuaciones en las escalas derivadas, que podemos resumir en el siguiente esquema:



PARTE SEGUNDA

Estadística descriptiva con dos variables

5 Correlación lineal

5.1. INTRODUCCIÓN

Uno de los objetivos principales de la psicología, como en otras ciencias, es descubrir relaciones entre variables. Así, en el campo de la psicología podemos preguntarnos si el rendimiento laboral en un tipo de puesto de trabajo guarda relación con la personalidad del trabajador, si el fracaso escolar está asociado con determinadas circunstancias personales y familiares, si un determinado estilo de vida fomenta los estados depresivos, si hay tareas en las que la práctica masiva facilita más el aprendizaje que la práctica distribuida, si determinado tipo de publicidad genera un mayor consumo del producto o qué rasgos de personalidad están asociados a una mayor propensión al suicidio. La observación de relaciones claras y estables entre las variables ayuda a comprender los fenómenos y a encontrar explicaciones de los mismos, además de sugerir vías de intervención. Para la búsqueda de estas relaciones, la estadística proporciona valiosos instrumentos.

En términos matemáticos, las relaciones entre variables pueden ser de muchos tipos; son funciones teóricas, ideales, que podrían ilustrar relaciones deterministas. Pero en la psicología nunca se encuentran relaciones deterministas, sino más bien conjuntos de observaciones que manifiestan una configuración algo irregular. Nos preguntaremos si esa configuración (que refleja la relación entre las variables) se parece a algún modelo teórico; en caso afirmativo, concluiremos que ese modelo captura bien la relación. Aquí vamos a centrarnos exclusivamente en el estudio de las relaciones lineales, que son las más sencillas. Lo que vamos a exponer son las formas más habituales de observar y cuantificar las relaciones lineales entre dos variables. Advertimos, por tanto, que aunque en este capítulo tratemos sobre relaciones o correlaciones entre variables, estrictamente hablando deberíamos utilizar la expresión «relación lineal»; si no lo hacemos será exclusivamente por razones de economía de espacio. Además, los estadísticos de asociación que vamos a describir son aplicables exclusivamente a las variables cuantitativas.

5.2. REPRESENTACIÓN GRÁFICA DE UNA RELACIÓN

Los procedimientos para determinar la existencia y grado de relación lineal entre dos variables deben ser capaces también de discriminar entre los tipos de relación lineal; las diferencias entre esos tipos quedarán más claras al hacer representaciones gráficas y exponer algunos ejemplos. Supongamos que registramos dos variables en un grupo de estudiantes: al comienzo del curso medimos su nivel de inteligencia mediante un test apropiado, y al final del curso evaluamos su rendimiento mediante la nota media obtenida. Es habitual que el resultado de la inspección de estos dos conjuntos de puntuaciones sea la constatación de que, en general, los estudiantes con inteligencia alta tienden a obtener mejores calificaciones que los estudiantes con inteligencia baja. Esta relación no es mecánica, puesto que siempre hay estudiantes con inteligencia alta que, por una enfermedad a lo largo del curso, por problemas familiares o motivacionales o por otros factores externos, no consiguen un nivel alto de rendimiento, mientras que otros con una inteligencia baja consiguen compensar esa menor capacidad con una dedicación mayor, consiguiendo al final buenas calificaciones. Sin embargo, estos casos especiales suelen ser minoría y, además, los factores externos pueden ejercer su influencia sobre estudiantes con cualquier nivel de inteligencia. En la mayoría de los casos sí se podrá apreciar esa tendencia general en la relación entre las variables: valores altos en inteligencia tienden a emparejarse con valores altos en rendimiento, y valores bajos en la primera tienden a emparejarse con valores bajos en la segunda. Cuando ocurra esto, diremos que entre las variables hay una *relación lineal directa.*

Supongamos ahora que a esa misma muestra de estudiantes les administramos una prueba de atención de lápiz y papel, consistente en tachar sólo las letras R en una hoja abigarrada de símbolos, en el menor tiempo posible. De cada participante anotamos dos variables: el tiempo invertido en completar la tarea y el número de errores cometidos al realizarla. Al inspeccionar los resultados probablemente observaremos un fenómeno bien conocido en psicología, llamado *balance entre velocidad y precisión* (Pew, 1969), que se manifiesta en que las personas que acaban antes suelen cometer más errores. Es decir, aquellas personas que adoptan una estrategia que prima la velocidad acaban pronto la tarea, pero a costa de cometer más errores, mientras que otros participantes actúan con más prudencia y prefieren tardar más con tal de equivocarse menos. En este caso también hay una cierta relación entre las variables, pero de un tipo muy diferente al observado entre la inteligencia y el rendimiento. Los valores bajos en la variable «tiempo invertido» tienden a estar emparejados con valores altos en la variable «número de errores» y viceversa. Cuando ocurra esto, diremos que entre las variables hay una *relación lineal inversa*.

Supongamos, por último, que en un grupo de estudiantes medimos también la estatura y escribimos los valores obtenidos por cada uno, emparejados con sus puntuaciones en inteligencia. Salvo coincidencias causales e inesperadas, la inspección de esos pares de valores probablemente nos indicará que no existe relación entre las variables. Con frecuencias parecidas, los valores altos de estatura estarán emparejados con valores de inteligencia altos, medios o bajos, y lo mismo ocurri-

rá con los valores de estatura medios y bajos. En este caso no podríamos decir que haya relación directa o inversa entre las variables. Cuando ocurra esto diremos que entre las variables hay una *relación lineal nula* (o independencia lineal).

Al hacer una representación gráfica conjunta de dos variables, se pueden apreciar visualmente estos tres tipos de relación. Para ello se identifican los pares de valores y se señalan los correspondientes puntos en unos ejes de coordenadas. Así, en el ejemplo primero se localizarían los puntos (9, 4) (12, 5) (6, 1) ... (13, 6). Esta nube de puntos recibe el nombre de *diagrama de dispersión*. En la figura 5.1 hemos construido los diagramas de dispersión de los tres ejemplos anteriores (figuras 5.1 *a*, *b* y *c*).

En estos ejemplos se puede apreciar algo que no surge por casualidad: en la representación de la relación entre inteligencia y rendimiento el diagrama tiende a tomar una forma alargada (tendencia a la linealidad) e inclinada hacia arriba; en la de la relación entre el tiempo de ejecución y el número de errores el diagrama muestra también una tendencia a la linealidad, pero descendente, y en la de la relación entre inteligencia y estatura no se aprecia tendencia alguna a la linealidad.

5.3. CUANTIFICACIÓN DE UNA RELACIÓN LINEAL

Al igual que con otros conceptos que hemos tratado en capítulos anteriores, vamos a desarrollar procedimientos precisos que sean capaces de distinguir entre los tres tipos de relación lineal que hemos descrito y de cuantificar su intensidad. La descripción de estos índices se basará en las representaciones de los tres ejemplos anteriores, según aparecen en la figura 5.2. En ella se representan los diagramas de dispersión, a los que hemos añadido unas líneas punteadas verticales y horizontales que indican los valores correspondientes a las medias de las variables de cada ejemplo (se han diseñado de forma tal que las medias en X e Y son iguales en los tres ejemplos, $\overline{X} = 8$; $\overline{Y} = 3$).

Un primer procedimiento para cuantificar la relación consiste en hallar el promedio de los productos cruzados de las puntuaciones diferenciales. Llamaremos *producto cruzado* al resultado de multiplicar, para cada individuo o caso, sus valores en las dos variables; es decir, $X_i \cdot Y_i$ (adviértase que en estos índices N ya no es, como hasta aquí, el número de valores observados, sino el número de pares de valores). Estos productos se pueden obtener con puntuaciones directas, diferenciales o típicas. Por tanto, el procedimiento de hallar el promedio de los productos cruzados de las puntuaciones diferenciales consiste en calcular:

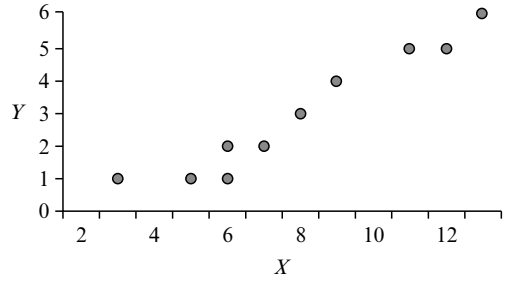
$$\frac{\Sigma(x_i \cdot y_i)}{N}$$

Veamos el comportamiento de este índice en los distintos casos de la figura 5.2. Es fácil darse cuenta de que cada figura está segmentada en cuatro cuadrantes y de que los puntos estarán en uno u otro dependiendo de que la observación supere o no la media de X y/o la media de Y. En concreto, si supera ambas medias, como el par (11, 5) del primer ejemplo, el punto aparecerá en el cuadrante superior derecho

a)

Inteligencia (X)	Rendimiento (Y)
9	4
12	5
6	1
5	1
8	3
7	2
3	1
6	2
11	5
13	6

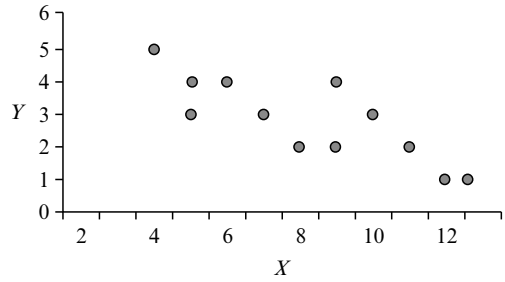
Relación directa



b)

Tiempo (X)	Errores (Y)
7	3
12	1
5	4
5	3
6	4
9	4
13	1
9	2
4	5
10	3

Relación inversa



c)

Estatura (X)	Inteligencia (Y)
7	3
8	4
5	3
10	3
9	5
11	4
7	1
6	2
8	3
9	2

Relación nula

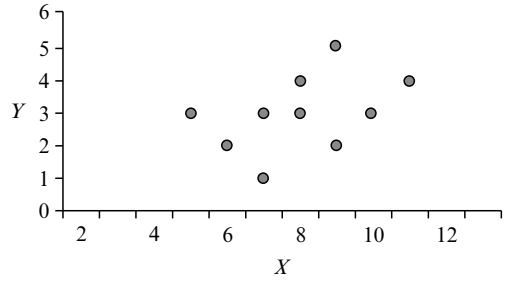


Figura 5.1.—Diagramas de dispersión de tres ejemplos de relaciones lineales entre variables: *a*) relación lineal directa; *b*) relación lineal inversa, y *c*) relación lineal nula.

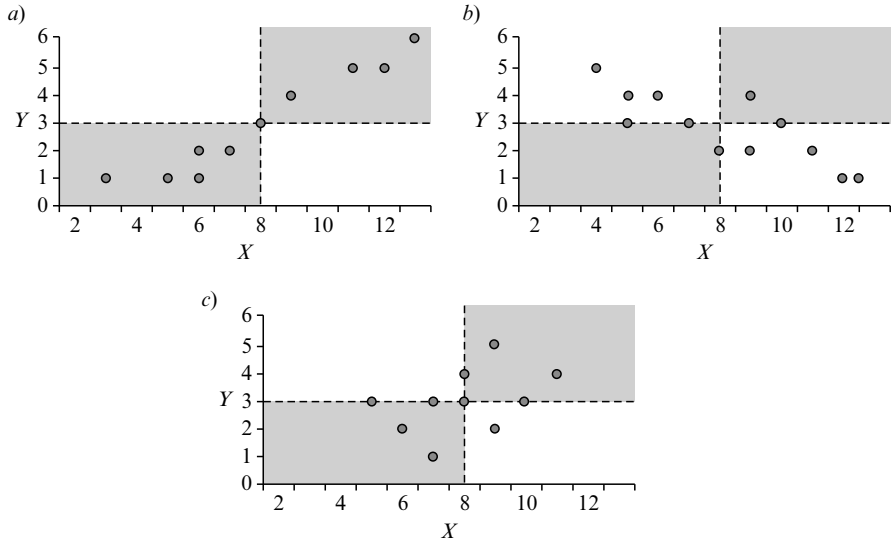


Figura 5.2.—Diagramas de dispersión de los tres ejemplos.

(noreste, NE); si supera la media de X pero no la de Y, como el par (9, 2) del segundo ejemplo, aparecerá en el cuadrante inferior derecho (sureste, SE); si supera la de Y pero no la de X, como el par (4, 5) del segundo ejemplo, aparecerá en el cuadrante superior izquierdo (noroeste, NO); por último, si no supera ninguna de las medias, como el par (5, 2) de todos los ejemplos, aparecerá en el cuadrante inferior izquierdo (suroeste, SO). Al tratar con puntuaciones diferenciales, éstas serán positivas si superan la media y negativas en caso contrario. Por tanto, aquellas observaciones que aparezcan en los cuadrantes NE o SO tendrán productos cruzados positivos (sus puntuaciones diferenciales serán ambas positivas o ambas negativas), mientras que las que aparezcan en los cuadrantes NO o SE tendrán productos cruzados negativos (una puntuación diferencial positiva y la otra negativa). Es fácil constatar que la diferencia entre los diagramas de dispersión que reflejan relaciones directas y los que reflejan relaciones inversas es que en los primeros hay muchos puntos en los cuadrantes NE y SO (cuyos productos cruzados de diferenciales son positivos) y pocos en los otros dos cuadrantes (cuyos productos cruzados son negativos), mientras que en los segundos ocurre lo contrario. En resumen, si existe una relación lineal directa entre las variables, en el numerador de $\sum(x_i \cdot y_i)/N$ habrá muchos sumandos positivos y pocos negativos, por lo que esa expresión tenderá a ser positiva. En las relaciones inversas serán mayoría los productos negativos y esa expresión tenderá a ser negativa. Por el contrario, en casos como el de la gráfica inferior de la figura 5.2 habrá una cantidad parecida de productos positivos y negativos que, al tender a compensarse entre sí, harán que esa expresión tienda a dar valores próximos a cero.

Por otra parte, si la nube de puntos está perfectamente en línea recta, todos estarán en los cuadrantes NE y SO o en los cuadrantes NO y SE. Cuanto menos acusada sea la tendencia a la linealidad, más equilibrada será la distribución de

los puntos entre los cuadrantes y más se compensarán los productos positivos y los negativos. Dicho de otra forma: el promedio de productos cruzados de diferenciales tenderá a dar positivo si la relación es directa, negativo si es inversa y en torno a cero si es nula; además, su valor absoluto será mayor cuanto más acusada sea la tendencia a la linealidad en el diagrama de dispersión.

El índice de la relación lineal del que estamos hablando se llama *covarianza*, se representa por S_{xy} y su fórmula es:

$$S_{xy} = \frac{\Sigma(x_i \cdot y_i)}{N} \qquad [5.1]$$

Sustituyendo las diferenciales por la expresión en directas y aplicando las reglas del sumatorio del apéndice del primer capítulo, se llega a una fórmula equivalente, generalmente más operativa, que emplea las puntuaciones directas:

$$S_{xy} = \frac{\Sigma(x_i \cdot y_i)}{N} = \frac{\Sigma[(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{N} = \frac{\Sigma[(X_i Y_i - X\bar{Y} - \bar{X}Y_i + \bar{X} \cdot \bar{Y})]}{N} =$$

$$= \frac{\Sigma X_i Y_i}{N} - \frac{\bar{Y}\Sigma X_i}{N} - \frac{\bar{X}\Sigma Y_i}{N} + \frac{N \cdot \bar{X} \cdot \bar{Y}}{N} = \frac{\Sigma X_i Y_i}{N} - \bar{Y} \cdot \bar{X} - \bar{X} \cdot \bar{Y} + \bar{X} \cdot \bar{Y} =$$

$$= \frac{\Sigma X_i Y_i}{N} - \bar{X} \cdot \bar{Y}$$

Por tanto, la covarianza se puede expresar también como el promedio de productos cruzados de las puntuaciones directas, menos el producto de las medias:

$$S_{xy} = \frac{\Sigma X_i Y_i}{N} - \bar{X} \cdot \bar{Y} \qquad [5.2]$$

Aplicamos esta última fórmula a los datos del ejemplo *a*) de la figura 5.1 (con inteligencia y rendimiento), dejando al lector la tarea de comprobar los de los otros dos ejemplos.

	X	*Y*	*X* · *Y*
	9	4	36
	12	5	60
	6	1	6
	5	1	5
	8	3	24
	7	2	14
	3	1	3
	6	2	12
	11	5	55
	13	6	78
Suma	80	30	293
Media	8	3	

Inteligencia y rendimiento: $S_{xy} = \frac{293}{10} - 8 \cdot 3 = 5{,}3$

Tiempo y número de errores: $S_{xy} = \frac{209}{10} - 8 \cdot 3 = -3{,}1$

Inteligencia y estatura: $S_{xy} = \frac{248}{10} - 8 \cdot 3 = -0{,}8$

Tal y como esperábamos, este índice es capaz de discriminar entre los tres casos expuestos, pues proporciona un resultado positivo en el primero, negativo en el segundo y cercano a cero en el tercero.

Sin embargo, la covarianza tiene una dificultad como índice de la asociación lineal. ¿Son 5,3 y −3,1 valores grandes o pequeños? ¿Una covarianza de −0,8 indica una asociación lineal relevante, o es un valor tan pequeño que podemos concluir que refleja prácticamente independencia lineal? Efectivamente, la covarianza carece de unos valores máximo y mínimo fijos; no tiene límites comunes a todos los casos, que permitan su interpretación inmediata. Esta dificultad se apreciará todavía más claramente si ponemos un ejemplo de cómo se alteraría la covarianza con un simple cambio en las unidades de medida. Así, si medimos la estatura en metros y el peso en kilos de un grupo de cinco individuos, la covarianza entre esas variables será distinta de si a esos mismos individuos les miden, como se hace en Estados Unidos, la estatura y el peso en pies y libras, respectivamente. Sin embargo, como son las mismas dimensiones e individuos, la asociación lineal entre sus estaturas y pesos debería ser constante. En el ejemplo siguiente hemos incluido unos datos de cinco individuos en las primeras unidades (X e Y) y su conversión en las segundas (V y W), así como las covarianzas entre las dos primeras y las dos segundas. Como se puede apreciar, el cambio de unidades tiene un importante efecto sobre S_{xy} (vale 0,283 al calcularla en metros y kilos y 2,053 al hacerlo en pies y libras).

X	Y	$X \cdot Y$
1,71	78	133,38
1,60	65	104,00
1,57	63	98,91
1,66	74	122,84
1,67	73	121,91
8,21	353	581,04

V	W	$V \cdot W$
5,61	171,96	964,70
5,25	143,30	752,33
5,15	138,89	715,28
5,45	163,14	889,11
5,48	160,94	881,95
26,94	778,23	4.203,37

$$\bar{X} = 1{,}642 \qquad \bar{Y} = 70{,}6 \qquad \bar{V} = 5{,}388 \qquad \bar{W} = 155{,}646$$

$$S_{xy} = \frac{581{,}04}{5} - 1{,}642 \cdot 70{,}6 = 0{,}283$$

$$S_{vw} = \frac{4.203{,}37}{5} - 5{,}388 \cdot 155{,}646 = 2{,}053$$

¿Es 0,283 un valor que indica una asociación lineal fuerte? ¿Indica 2,053 una asociación más fuerte que 0,283? Este problema de la valoración de la covarianza se solucionaría si el índice tuviera unos valores máximo y mínimo que sirvieran de referencia. La solución a este problema consistirá en utilizar unas puntuaciones cuyas varianzas no cambien al modificar las unidades de medida de las puntuaciones originales, sino que sean siempre iguales. Naturalmente, estas puntuaciones van a ser las puntuaciones típicas, cuya varianza es siempre igual a 1. Efectivamente, un segundo índice de la asociación lineal consistirá en hallar también un promedio de productos cruzados, pero no de los productos cruzados de las puntuaciones diferenciales, sino de las puntuaciones típicas. Este índice fue desarrollado y propuesto por Francis Galton y Karl Pearson, aunque fue este último el que le dio su fórmula definitiva (Pearson, 1896), por lo que se conoce como *coeficiente de correlación* de Pearson; se representa por la letra *r*. La correlación de Pearson entre las variables *X* e *Y* será:

$$r_{xy} = \frac{\sum z_{x_i} z_{y_i}}{N} \qquad [5.3]$$

La correlación no es más que una covarianza hallada sobre puntuaciones típicas; por eso a veces se dice que la correlación es una covarianza estandarizada o que es una covarianza adimensional.

La fórmula [5.3] no resulta muy práctica a la hora de hacer cálculos. Exige la tipificación de cada puntuación, por lo que para aplicarla hay que hallar previamente las medias y desviaciones típicas de cada variable. Para facilitar los cálculos manuales se han derivado otras fórmulas alternativas equivalentes, que en la mayoría de los casos resultarán más prácticas. La primera, [5.4], se obtiene sustituyendo en [5.3] las puntuaciones típicas por su expresión en diferenciales ($z_x = x/S_x$ y $z_y = y/S_y$). La segunda, [5.5], se obtiene advirtiendo en [5.4] que el numerador dividido por *N* no es otra cosa que la covarianza. Para la tercera se sustituye en [5.4] las puntuaciones diferenciales por su expresión en directas ($x_i = X_i - \bar{X}$; $y_i = Y_i - \bar{Y}$) y las desviaciones típicas según su fórmula; tras varias transformaciones algebraicas se llega a [5.6].

$$r_{xy} = \frac{\sum x_i y_i}{N \cdot S_x \cdot S_y} \qquad [5.4]$$

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} \qquad [5.5]$$

$$r_{xy} = \frac{N \cdot \sum X_i Y_i - \sum X_i \cdot \sum Y_i}{\sqrt{N \cdot \sum X_i^2 - (\sum X_i)^2} \cdot \sqrt{N \cdot \sum Y_i^2 - (\sum Y_i)^2}} \qquad [5.6]$$

Cuando, por la razón que sea, dispongamos ya tanto de los productos cruzados entre diferenciales como de las desviaciones típicas, será mejor utilizar la fórmula [5.4]. En cambio, cuando dispongamos de los productos cruzados en directas y de las medias y las desviaciones típicas, entonces será más cómodo

calcular primero la covarianza y luego utilizar la fórmula [5.5]. Por último, si no tenemos hecho ningún cálculo previo, resulta más práctica la fórmula [5.6], a pesar de su mayor aparatosidad. Veamos un ejemplo de aplicación de la tercera fórmula con nuestro ejemplo de la inteligencia y el rendimiento. En realidad, basta con tomar los cálculos hechos para hallar la covarianza y añadir los cuadrados de los valores de ambas variables.

	X	Y	$X \cdot Y$	X^2	Y^2
	9	4	36	81	16
	12	5	60	144	25
	6	1	6	36	1
	5	1	5	25	1
	8	3	24	64	9
	7	2	14	49	4
	3	1	3	9	1
	6	2	12	36	4
	11	5	55	121	25
	13	6	78	169	36
Σ	80	30	293	734	122

$$r_{xy} = \frac{N \cdot \sum X_i Y_i - \sum X_i \cdot \sum Y_i}{\sqrt{N \cdot \sum X_i^2 - (\sum X_i)^2} \cdot \sqrt{N \cdot \sum Y_i^2 - (\sum Y_i)^2}} = \frac{10 \cdot 293 - 80 \cdot 30}{\sqrt{10 \cdot 734 - 80^2} \cdot \sqrt{10 \cdot 122 - 30^2}} =$$

$$= \frac{530}{\sqrt{940} \cdot \sqrt{320}} = 0{,}966$$

Queremos resaltar dos últimas ideas acerca de la definición de r_{xy} que se desprenden de la fórmula [5.5]. La primera es que la correlación entre dos variables siempre tendrá el mismo signo que la covarianza, puesto que lo que aparece en el denominador de esa fórmula son desviaciones típicas, que siempre son positivas. La segunda es que si la desviación típica de alguna de las variables es cero, la correlación es indeterminada. En términos gráficos, esto ocurre cuando los puntos están en perfecta línea recta, pero esta línea es paralela a alguno de los ejes. Por ejemplo, las dos muestras de pares de valores siguientes, que tienen varianza nula en alguna de las variables y cuyos diagramas de dispersión aparecen en la figura 5.3, presentan esta característica, que recibe el nombre de *colinealidad*.

X	Y
12	3
6	3
8	3
5	3
10	3
2	3

X	Y
6	3
6	1
6	5
6	4
6	6
6	2

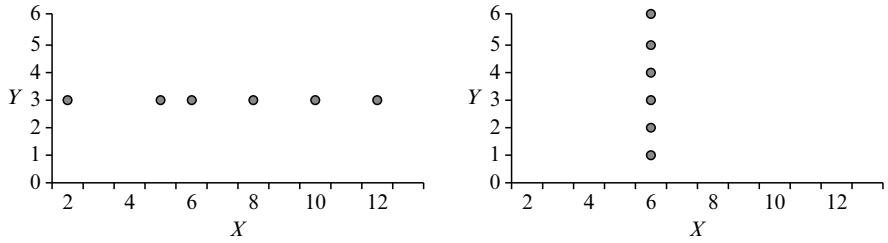


Figura 5.3.—Ejemplos de pares de valores que presentan colinealidad.

5.3.1. Propiedades del coeficiente de correlación de Pearson

La razón principal por la que la covarianza no llegaba a satisfacer completamente la necesidad de un índice de la asociación lineal era la dificultad de su valoración, dado que carecía de un máximo y un mínimo prefijados. Puesto que hemos destacado su alternativa principal, el coeficiente de correlación de Pearson, precisamente porque no tiene esa dificultad, subrayaremos este hecho como primera propiedad, enunciándola primero y demostrándola después: *el coeficiente de correlación de Pearson no puede valer más de +1 ni menos de −1 (*$-1 \leq r_{xy} \leq 1$*)*.

Amón (1993) nos proporciona una elegante demostración de esta propiedad, que reproducimos a continuación. Supongamos que definimos las siguientes variables, siendo z_x y z_y dos variables tipificadas:

$$U = z_x + z_y$$

$$V = z_x - z_y$$

Obtengamos a continuación la varianza de estas dos variables. Dado que $\bar{z}_x = \bar{z}_y = 0$ (véase el capítulo 4), la media de U será igual a 0; por tanto, la varianza de U será igual a:

$$S_U^2 = \frac{\sum(U_i - \bar{U})^2}{N} = \frac{\sum(z_{x_i} + z_{y_i} - 0)^2}{N} = \frac{\sum(z_{x_i} + z_{y_i})^2}{N} = \frac{\sum(z_{x_i}^2 + z_{y_i}^2 + 2 \cdot z_{x_i} \cdot z_{y_i})}{N} =$$

$$= \frac{\sum z_{x_i}^2}{N} + \frac{\sum z_{y_i}^2}{N} + \frac{2 \cdot \sum z_{x_i} \cdot z_{y_i}}{N}$$

Dado que las medias de las puntuaciones típicas son iguales a cero, las expresiones de los dos primeros quebrados son, en realidad, las varianzas de las puntuaciones típicas de X e Y, respectivamente, que son iguales a 1, mientras que el tercero no es más que la correlación de Pearson multiplicada por 2. Haciendo un desarrollo similar para la variable V, llegamos a las dos expresiones siguientes:

$$S_U^2 = 2 + 2 \cdot r_{xy} \qquad S_V^2 = 2 - 2 \cdot r_{xy}$$

Observando cuidadosamente estas dos expresiones, advertimos en la primera que un valor de r_{xy} menor de −1 conduciría a una varianza de U negativa, mientras que en la segunda ocurriría lo mismo con valores de r_{xy} mayores de +1; por tanto, r_{xy} no puede adoptar esos valores. El lector puede estar seguro de que en caso de que sus cálculos le lleven a un coeficiente de correlación de Pearson fuera del rango [−1; 1] ha cometido algún error y debe repasar las operaciones hasta encontrarlo.

La máxima correlación entre dos variables se obtiene cuando las puntuaciones en ambas variables son equivalentes, es decir, cuando para cada caso se verifica que $z_x = z_y$, pues entonces:

$$r_{xy} = \frac{\sum z_{x_i} \cdot z_{y_i}}{N} = \frac{\sum z_{x_i} \cdot z_{x_i}}{N} = \frac{\sum z_{x_i}^2}{N} = 1$$

Puesto que al ser nula la media de las puntuaciones típicas, la expresión del último quebrado es igual a la varianza de las puntuaciones típicas, que ya sabemos que es necesariamente igual a 1. Por el contrario, el mínimo valor de r_{xy} se obtiene cuando la puntuación típica de un par en una variable está asociada con una puntuación típica igual en la otra variable, pero con el signo opuesto; es decir, cuando para cada caso se verifica que $z_x = -z_y$, pues en este caso la correlación es exactamente igual a −1.

Un hecho interesante que merece ser resaltado sobre este índice es que la correlación de Pearson de una variable consigo misma es necesariamente igual a 1; esto es así porque, en ese caso, $r_{xx} = \sum z_x \cdot z_x / N = \sum z_x^2 / N$. Ya hemos visto más arriba que esta expresión no es otra cosa que la varianza de las puntuaciones típicas, que necesariamente es igual a 1. Como r es igual a 1 cuando las puntuaciones típicas de las dos variables son estrictamente equivalentes, es fácil entender la afirmación de Runyon y Haber (1971, p. 96) de que «r representa el grado en que los mismos individuos o eventos ocupan la misma posición relativa en dos variables».

Para exponer la segunda propiedad de la correlación de Pearson vamos a detenernos en el estudio de los efectos de las transformaciones lineales sobre r_{xy}. Supongamos que tenemos dos variables, X e Y, cuya correlación es r_{xy}, y hacemos las siguientes transformaciones lineales, siendo a y c dos constantes positivas:

$$U_i = a \cdot X_i + b$$

$$V_i = c \cdot Y_i + d$$

Sustituyendo en la fórmula [5.4]:

$$r_{uv} = \frac{\sum u_i v_i}{N \cdot S_U \cdot S_V} = \frac{\sum [a \cdot (X_i - \bar{X}) \cdot c \cdot (Y_i - \bar{Y})]}{N \cdot a \cdot S_x \cdot c \cdot S_y} = \frac{a \cdot c}{a \cdot c} \cdot \frac{\sum x_i y_i}{N \cdot S_x \cdot S_y} = r_{xy}$$

Podemos resumir esta relación como nuestra segunda propiedad: *si hacemos transformaciones lineales de una o de las dos variables, en las que las constantes multiplicadoras son positivas, la correlación de Pearson no se altera.* Es decir, si $U = a \cdot X + b$ y $V = c \cdot Y + d$, siendo (a, $c > 0$), entonces $r_{uv} = r_{xy}$.

Aunque nos hemos centrado en las transformaciones con constantes multiplicadoras positivas, que son las más frecuentemente utilizadas, podemos encontrarnos con casos en los que una o las dos constantes multiplicadoras sean negativas. Aun sin demostración, enunciaremos los efectos de tales transformaciones sobre *r*. Si las dos constantes son negativas, el valor de *r* sigue quedando inalterado. Pero si una es positiva y la otra negativa, entonces la correlación es igual en valor absoluto, pero con el signo cambiado.

En realidad, las fórmulas [5.4], [5.5] y [5.6] son casos particulares de esta propiedad. Es indiferente obtener r_{xy} con puntuaciones directas, diferenciales o típicas, puesto que las diferenciales y las típicas no son más que transformaciones lineales de las directas. Como ejemplo de esta propiedad, animamos al lector a comprobar que las correlaciones entre estaturas y pesos medidos en diferentes unidades en un ejemplo anterior de este mismo apartado dan exactamente lo mismo, a pesar de que sus covarianzas son diferentes.

5.3.2. Valoración e interpretación de una correlación

Hay algunas cuestiones importantes que se deben tener en cuenta para valorar e interpretar un coeficiente de correlación. En este apartado hemos reunido algunas orientaciones al respecto.

En la interpretación de una correlación de Pearson hay que separar dos aspectos diferentes: su cuantía y su sentido. La cuantía se refiere al grado en que la relación entre dos variables queda bien descrita con un índice de asociación lineal como *r*, mientras que el sentido se refiere al tipo de relación. Una correlación en torno a cero indica una relación lineal baja o nula; una correlación positiva indica una relación lineal directa, mientras que una correlación negativa indica una relación lineal inversa. Podemos imaginarnos a los coeficientes de correlación situados en un eje imaginario (figura 5.4). Cuanto más cercano quede un coeficiente del valor cero, menos apto es el modelo lineal como descripción de la relación entre las variables. Por el contrario, cuanto más se acerque a los extremos, mejor describe esa relación.

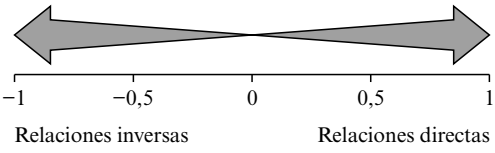


Figura 5.4.—Eje para la representación de coeficientes de correlación.

Sin embargo, la base para valorar *r* no debe ser su valor simple. De hacerlo así, se cae en la tentación de establecer conclusiones del tipo de que una correlación de 0,60 indica que hay un 60 por 100 de asociación lineal, o que una correlación de 0,80 indica el doble de asociación lineal que una correlación de 0,40. Aunque por ahora el lector sólo puede confiar en lo que decimos, puesto que la exposición y demostración de ello se incluirá en un capítulo posterior, la base para valorar un

coeficiente de correlación debe ser el cuadrado de su valor, r^2. Como consecuencia, cuando $r_{xy} = 0{,}80$ y $r_{uv} = 0{,}40$ no diremos que en el primer caso el grado de asociación lineal es el doble que en el segundo ($r_{xy}/r_{uv} = 0{,}80/0{,}40 = 2$), sino que diremos que es cuatro veces mayor ($r^2_{xy}/r^2_{uv} = 0{,}64/0{,}16 = 4$).

Además, hay que tener otras precauciones al interpretar coeficientes de correlación. La obtención de una correlación igual (o cercana) a cero puede llevar a pensar erróneamente que no hay relación entre las variables. La correlación de Pearson mide el grado de adecuación de unos datos a un modelo lineal, pero entre las variables puede existir otro tipo de relación. Un ejemplo típico de esta circunstancia es la relación entre activación y rendimiento. Desde hace bastantes años es sabido que los niveles bajos de activación están asociados con rendimiento bajo, mientras que con niveles medios de activación el rendimiento es alto. Si entre las variables hubiera una relación directa sistemática, el rendimiento seguiría incrementándose indefinidamente al incrementar la activación, pero no sucede así. Cualquier estudiante sabe que con estados altos de ansiedad se reduce el rendimiento en los exámenes. Es decir, el rendimiento máximo se obtiene con niveles medios de activación, mientras que con niveles demasiado bajos o demasiado altos el rendimiento disminuye. Esta relación es conocida como ley de Yerkes-Dodson (1908) y establece que la relación entre activación y rendimiento adopta la forma de una U invertida. En la figura 5.5 hemos representado el conjunto de pares de valores que aparecen a su lado, cuya correlación es prácticamente cero y que sirve para ilustrar precisamente este caso, pues entre las variables sí hay una relación, aunque no sea de tipo lineal.

X	*Y*
4	3
11	3
3	2
13	1
5	3
3	1
4	2
12	3

X	*Y*
4	4
10	5
7	5
12	2
8	5
10	4
14	2

$r_{xy} = -0{,}013$

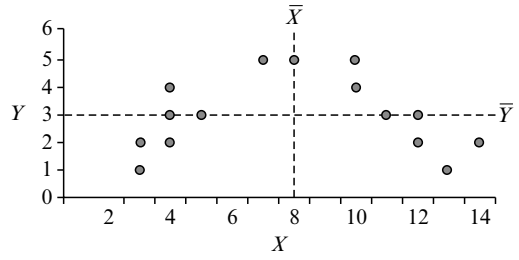


Figura 5.5.—Representación típica de unos puntos que reflejan una relación en forma de U invertida, como la relación entre ansiedad y rendimiento que trató de explicar la ley de Yerkes-Dodson.

Hay, además, otros factores que afectan a las expectativas sobre el valor de *r*, como por ejemplo la variabilidad o la mediación de terceras variables. También hay una cierta especificidad en los campos de estudio concretos. Por ejemplo, para estudiar la estabilidad de las puntuaciones que ofrece un test se suele aplicar el test dos veces, dejando un intervalo de tiempo, y se calcula la correlación entre las puntuaciones obtenidas en las dos administraciones de la prueba. Así se obtiene lo que se llama la fiabilidad del test (test-retest). Pues bien, se debe desconfiar de cualquier test de inteligencia con una correlación test-retest menor de 0,90. Por el contrario, en el estudio de la personalidad una correlación superior a 0,60 es un resultado que se puede considerar importante. En cada área de estudio se va desarrollando un conocimiento que permite valorar los coeficientes de correlación en términos relativos. En el cuadro 5.1 presentamos algunos resultados de la literatura psicológica que ayudarán al lector a hacerse una idea de lo diferentes que pueden ser las correlaciones prototípicas en diferentes áreas y variables. Ésta es la razón por la que preferimos no proponer categorías generales de valoración, como a veces se hace, calificando como correlaciones altas a las que en valor absoluto son superiores a cierta cantidad. Los coeficientes de correlación se deben valorar comparando unos con otros, o comparándolos con los valores que típicamente se suelen encontrar en el campo de estudio específico del que se trate. No obstante, no hay que olvidar que estamos exponiendo estadística descriptiva. En la estadística inferencial se desarrollan procedimientos plenamente justificados para valorar e interpretar los coeficientes de correlación (véase capítulo 15).

CUADRO 5.1

Ejemplos de coeficientes de correlación de Pearson obtenidos en la literatura psicológica

- Carrobles, Remor y Rodríguez-Alzamora (2003) obtuvieron, en un estudio con adultos con infección por VIH, una correlación de -0,31 entre el grado en que asumen una estrategia de afrontamiento activa-positiva y los niveles de depresión que muestran.
- Entrenando en habilidades sociales a niños entre 6 y 12 años, O'Connor, Frankel, Paley, Schonfeld, Carpenter, Laugeson y Marquardt (2006) obtuvieron una correlación de 0,43 entre la inteligencia de los niños (medida como CI) y el resultado del entrenamiento en habilidades sociales; los niños más inteligentes se benefician más de la intervención.
- Schweizera y Moosbrugger (2004) informan de una correlación de 0,51 entre inteligencia general y atención sostenida.
- Redzuan, Juhari, Yousefi, Mansor y Talib (2010) obtuvieron una correlación de -0,22 entre depresión y rendimiento académico en estudiantes de edades comprendidas entre 15 a 19 años.
- Fan-peng y Dong (2010), investigando sobre la adquisición del inglés en población de estudiantes chinos, informan de una correlación de 0,545 entre el nivel de concentración para pronunciar bien el inglés y el nivel de ansiedad cuando hablan este idioma.
- Ríos-Rísquez, Sánchez-Meca y Godoy-Fernández (2010) encontraron una correlación de 0,218 entre la edad y la manifestación de síntomas de ansiedad en una muestra de profesionales de enfermería de Urgencias y Cuidados Intensivos.

Lo que se desprende de todo esto es que no conviene analizar la relación entre dos variables exclusivamente mediante el mero cálculo del coeficiente de correlación, sino que conviene representar gráficamente el diagrama de dispersión para observar esa relación. Una representación gráfica puede ser más informativa que un simple valor de *r*. Otros ejemplos de situaciones engañosas son aquellas en las que se trata con variables que tienen un rango restringido o aquellas en las que se mezclan grupos no homogéneos. Los primeros se refieren a aquellos casos en los que los valores de *X* observados no son representativos de los valores posibles. Supongamos que la relación entre las variables *X* e *Y* fuera la representada en la figura 5.6.*a*), pero debido al procedimiento de muestreo nosotros recogemos sólo los pares de valores oscuros. A partir de ellos concluiríamos que no hay relación lineal entre las variables, a pesar de que en la población sí la hay. Un ejemplo de grupos no homogéneos sería el de la figura 5.6.*b*), en la que se han representado los pares de valores observados en una muestra. Al analizarlos globalmente se aprecia una cierta tendencia a la linealidad, pero en realidad es una situación engañosa. Al separar los casos correspondientes a hombres (•) y mujeres (◦), se observa que dentro de cada subpoblación la situación es muy diferente de la que aparece en el total.

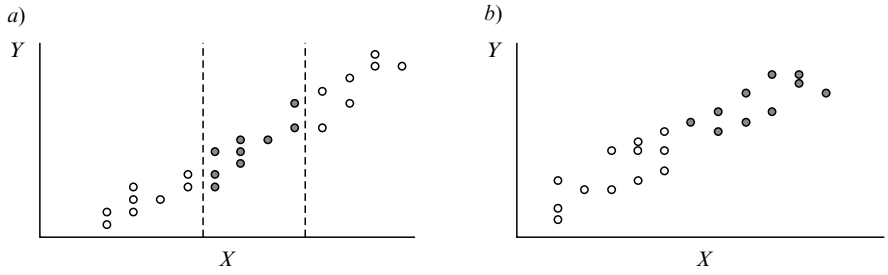


Figura 5.6.—Representación de dos casos especiales: *a*) ausencia aparente, pero engañosa, de correlación debida a una restricción en el rango de valores de *X* observados; *b*) presencia aparente, pero engañosa, de correlación debida a la mezcla de dos grupos heterogéneos con correlación baja dentro de cada uno.

Otro peligro que corren los recién llegados a la estadística (y a veces algunos que ya llevan tiempo trabajando con ella) es el de interpretar los coeficientes de correlación en términos de relaciones causales entre las variables. Aunque no es éste el lugar apropiado para abordar la cuestión, queremos al menos dedicarle unas líneas. Una correlación no sólo puede aparecer entre una variable-causa y una variable-efecto, sino también entre variables entre las que no hay relaciones de causalidad; por tanto, no se deben inferir relaciones de causalidad sólo a partir del hecho de haber obtenido un valor de *r* sustancialmente diferente de cero (León y Montero, 2003). De lo contrario, podríamos llegar a paradojas como la siguiente. Si medimos en los países de la ONU el número de coches por cada mil habitantes y el nivel cultural medio de sus habitantes, probablemente encontraremos una relación directa y, por tanto, una correlación positiva. Pero, ¿quiere decir esto que si regalamos coches a los habitantes de un país se incrementará su nivel cul-

tural? Es claro que en este argumento hay una falacia, que reside en el hecho de que entre estas variables no hay una relación causal, sino que probablemente ambas tienen una relación causal con una tercera variable, el nivel económico del país. Un mayor nivel económico tiene efectos, simultáneamente, sobre el nivel cultural y sobre los bienes de consumo. Por tanto, mientras no tengamos otras informaciones adicionales, cuando encontremos un valor de r diferente de cero nos limitaremos a decir que hemos encontrado un patrón de covariación entre las variables.

5.3.3. Las matrices de correlaciones y de varianzas y covarianzas

En muchos estudios se miden conjuntos de variables y se cuantifican sus relaciones lineales, dos a dos, mediante sendos coeficientes de correlación de Pearson. La forma más habitual de incluir estos valores en un informe de investigación es confeccionar una *matriz de correlaciones* (a veces denominada *Matriz* **R**). Ésta consiste en una tabla con el mismo número de filas y columnas que de variables, en la que en cada casilla aparece la correlación entre las variables correspondientes a la fila y la columna. Es decir, si trabajamos con las variables U, V, W, ..., X, entonces la matriz tendría la siguiente forma:

	U	V	W	...	X
U	r_{uu}	r_{uv}	r_{uw}	...	r_{ux}
V	r_{vu}	r_{vv}	r_{vw}	...	r_{vx}
W	r_{wu}	r_{wv}	r_{ww}	...	r_{wx}
...	...	...	...	...	...
X	r_{xu}	r_{xv}	r_{xw}	...	r_{xx}

Parte de la información contenida en esta matriz es redundante. Así, en la diagonal principal aparecen las correlaciones de cada variable consigo misma; ya hemos visto anteriormente que estas correlaciones son necesariamente iguales a 1. Igualmente, la matriz es simétrica con respecto a la diagonal principal; todos los valores que aparecen por encima de ella se repiten de nuevo por debajo. La razón es que, necesariamente, $r_{xy} = r_{yx}$. Por todo ello, es frecuente que las matrices de correlaciones que aparecen en la literatura sean del tipo conocido como matriz triangular superior (si se presentan los valores en las posiciones superiores a la diagonal principal) o como matriz triangular inferior (si se presentan en las posiciones inferiores). La siguiente es un ejemplo de matriz de correlaciones triangular superior:

	U	V	W	X
U		−0,17	−0,11	−0,30
V			0,46	0,17
W				0,10
X				

Ximénez y Revuelta (2010) nos ofrecen un ejemplo de una matriz de correlaciones triangular inferior. En su estudio sobre la invarianza factorial de una medida del ajuste de la persona a su organización de trabajo, se preguntó a una muestra de 490 trabajadores (todos ellos antiguos alumnos de la UAM) sobre sus preferencias en cuanto a diferentes aspectos del trabajo y el grado en que éstos se ven satisfechos por su organización. La matriz de correlaciones y los nombres de algunas de las variables incluidas en el estudio aparecen a continuación:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1										
X_2	0,05									
X_3	0,05	0,33								
X_4	0,23	0,17	0,19							
X_5	0,09	0,19	0,18	0,24						
X_6	0,08	0,02	0,08	−0,02	0,07					
X_7	−0,04	0,15	0,09	−0,08	0,10	0,29				
X_8	0,02	0,12	0,12	0,02	0,08	0,13	0,36			
X_9	−0,08	0,05	0,08	0,03	0,14	0,44	0,44	0,33		
X_{10}	−0,05	0,07	0,08	−0,06	0,22	0,37	0,49	0,32	0,60	

X_1: Seguridad deseada.
X_2: Autonomía deseada.
X_3: Consecución de objetivos deseada.
X_4: Oportunidades de promoción deseadas.
X_5: Estatus deseado.
X_6: Seguridad obtenida.
X_7: Autonomía obtenida.
X_8: Consecución de objetivos obtenida.
X_9: Oportunidades de promoción obtenidas.
X_{10}: Estatus obtenido.

Otro instrumento muy útil para exponer los estadísticos muestrales de una forma compacta y clara y para realizar ciertos cálculos es la matriz de varianzas y covarianzas (también denominada *Matriz* **S**). Como la de correlaciones, la matriz de varianzas y covarianzas es una matriz cuadrada en la que cada fila y cada columna corresponden a una variable. En cada casilla aparece el promedio de los productos cruzados de puntuaciones diferenciales entre las variables correspondientes a su fila y columna. Así, por ejemplo, en la casilla correspondiente a la fila de la variable X y a la columna de la variable Y aparecerá la expresión:

$$\frac{\Sigma(x_i \cdot y_i)}{N}$$

Esta expresión no es más que S_{xy}, la covarianza entre las variables X e Y. Sin embargo, en la diagonal principal coincidirán las variables de la fila y la columna; por tanto, aparecería la covarianza de una variable consigo misma. Es fácil advertir que la covarianza de una variable, X, consigo misma no es más que S_x^2, la varianza de esa variable:

$$\frac{\sum(x_i \cdot x_i)}{N} = \frac{\sum x_i^2}{N} = S_x^2$$

En consecuencia, la matriz de varianzas y covarianzas de las variables U, V, X,... W será:

	U	V	X	...	W
U	S_u^2	S_{uv}	S_{ux}	...	S_{uw}
V	S_{vu}	S_v^2	S_{vx}	...	S_{vw}
X	S_{xu}	S_{xv}	S_x^2	...	S_{xw}
...	...	...	...	...	...
W	S_{wu}	S_{wv}	S_{wx}	...	S_w^2

Esta matriz comparte con la matriz de correlaciones la propiedad de la simetría. Es decir, como la covarianza entre X e Y es la misma que entre Y y X, los coeficientes que aparecen sobre la diagonal principal aparecen repetidos bajo la misma.

Como ya se ha mencionado, la matriz de varianzas y covarianzas es un instrumento muy útil para: *a*) realizar los cálculos dirigidos a calcular la varianza de una combinación lineal de varianzas (como expondremos en el próximo capítulo), y *b*) ofrecer los estadísticos más importantes de un conjunto de variables en un formato compacto y sencillo de entender, sobre todo si a la matriz se le añade un vector con las medias de las variables.

Con frecuencia nos interesará pasar de la matriz de correlaciones a la de varianzas y covarianzas, o viceversa. Para ello se emplea la fórmula [5.5], que pone la covarianza en función de la correlación y viceversa. Veamos un ejemplo. Supongamos que queremos transformar la siguiente matriz de varianzas y covarianzas en la correspondiente matriz de correlaciones (una operación que llamaremos el *volcado* de la matriz):

	U	V	W	X
U	16	10	−15	8
V	10	9	−12	6
W	−15	−12	25	−8
X	8	6	−8	16

Para obtener cualquiera de las correlaciones aplicamos la ecuación [5.5]. Por ejemplo, para obtener la correlación entre las variables U y V localizamos en la matriz de varianzas y covarianzas los valores que debemos sustituir en ella (las varianzas de U y V son 16 y 9, mientras que la covarianza entre ellas es 10):

$$r_{uv} = \frac{S_{uv}}{S_u \cdot S_v} = \frac{10}{\sqrt{16} \cdot \sqrt{9}} = 0{,}833$$

Tras aplicar esta fórmula para cada par de variables obtenemos la siguiente matriz de correlaciones final.

	U	V	W	X
U		0,833	−0,750	0,500
V			−0,800	0,500
W				−0,400
X				

A veces, los resultados de un estudio se ofrecen también como una matriz de correlaciones a la que se añaden las medias y varianzas de cada variable. En estas ocasiones se puede hacer el volcado inverso, para confeccionar la matriz de varianzas y covarianzas. Supongamos la siguiente matriz de correlaciones, junto con las varianzas de cada variable:

	U	V	X
U		0,825	0,750
V			0,400
X			
Varianza	20	35	60

Despejando la covarianza en la fórmula [5.5] obtenemos la siguiente expresión:

$$S_{xy} = r_{xy} \cdot S_x \cdot S_y \qquad [5.7]$$

Así, la covarianza entre las variables U y V se obtiene sustituyendo en esta fórmula los valores de las varianzas de esas variables (20 y 35) y la correlación entre ellas (0,825):

$$S_{uv} = r_{uv} \cdot S_u \cdot S_v = 0{,}825 \cdot \sqrt{20} \cdot \sqrt{35} = 21{,}83$$

Tras aplicar esta fórmula para cada casilla y colocar las varianzas en la diagonal principal, se obtiene la siguiente matriz de varianzas y covarianzas:

	U	*V*	*X*
U	20	21,83	25,98
V	21,83	35	18,33
X	25,98	18,33	60

La mayoría de los programas informáticos comerciales, como el SPSS (Norusis, 2011), ofrecen las matrices de correlaciones y de varianzas-covarianzas en este mismo formato (véase Ximénez y Revuelta, 2011, pp. 33-35).

PROBLEMAS Y EJERCICIOS

1. Disponemos de los siguientes datos en las variables X e Y:

Sujeto	1	2	3	4	5	6
X	2	4	7	9	3	5
Y	1	5	9	8	3	2

a) Estudie, en primer lugar, de forma gráfica, la posible relación lineal entre las variables X e Y.

b) Calcule el índice de correlación de Pearson e interprete los resultados.

2. Calcule la covarianza y el coeficiente de correlación de Pearson entre las variables X e Y en la siguiente muestra de valores, e interprete los resultados obtenidos:

Sujeto	1	2	3	4	5	6
X	7	10	11	9	8	3
Y	5	3	2	6	5	11

3. Elabore el diagrama de dispersión y calcule el coeficiente de correlación de Pearson para las variables X e Y en la siguiente muestra de 20 sujetos:

Sujeto	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
X	2	5	7	3	4	5	5	6	1	4	3	9	8	6	7	5	2	2	1	8
Y	5	1	6	6	5	4	3	2	1	5	4	8	2	1	3	3	4	2	3	7

4. ¿Cuáles son los posibles valores de r_{xy} cuando la muestra está compuesta sólo de dos sujetos? (conteste a esta pregunta razonando la respuesta y sin hacer ningún tipo de cálculo).

5. Si $r_{xy} = 0$, ¿son las variables X e Y entonces independientes?

6. Llevado a cabo un estudio sobre la relación entre la motivación de logro con diferentes facetas de la satisfacción laboral en una muestra de 84 trabajadores, se obtuvo la siguiente matriz de correlaciones:

$$\mathbf{R} = \begin{array}{c|ccccc} & SS & SH & SR & MC & OP \\ \hline SS & & 0{,}82 & 0{,}61 & 0{,}42 & -0{,}49 \\ SH & & & 0{,}35 & 0{,}15 & -0{,}15 \\ SR & & & & 0{,}75 & 0{,}31 \\ MC & & & & & 0{,}45 \\ OP & & & & & \end{array}$$

Donde: *SS*: Satisfacción con el sueldo obtenido.
SH: Satisfacción con el horario realizado.
SR: Satisfacción con el reconocimiento obtenido.
MC: Motivación por la consecución de objetivos.
OP: Oportunidades de promoción.

A continuación, conteste a las siguientes cuestiones:

a) ¿Qué variable correlaciona más con *MC*?
b) ¿Qué variable correlaciona menos con *SH*?
c) ¿Cuál es la mayor correlación lineal encontrada?
d) ¿*SR* se relaciona más con *SS* o con *MC*?
e) ¿Cómo se interpreta la correlación lineal negativa entre *OP* y *SS*?

7. Medidas las variables *X*: *Rendimiento académico* e *Y*: *Tiempo dedicado al ocio* (horas/semana), se obtuvieron los siguientes resultados en dos muestras, una de estudiantes de Madrid y otra de estudiantes de Valencia:

	Madrid										**Valencia**									
Sujeto	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X	4	7	2	8	9	6	7	5	7	3	1	4	3	5	5	4	7	9	4	8
Y	30	22	27	20	18	21	23	25	17	15	35	31	29	28	31	27	22	23	24	19

A partir de estos datos, conteste a las siguientes cuestiones:

a) Indique cuál de las dos muestras dedica un mayor número de horas semanales a su tiempo de ocio.
b) Calcule el coeficiente de correlación de Pearson en cada muestra y diga en cuál de las dos muestras existe mayor correlación lineal entre el rendimiento académico y el tiempo dedicado al ocio.
c) Elabore una representación gráfica conjunta del diagrama de dispersión de ambas muestras.

8. Disponemos de la siguiente matriz de varianzas-covarianzas para las variables X, Y y Z:

	X	Y	Z
X	9	3,84	−5,29
Y		4	2,69
Z			6

A partir de los datos anteriores:

a) Elabore la matriz de correlaciones para las variables X, Y y Z.
b) Indique con qué variable tiene mayor grado de relación lineal la variable X, si con Y o con Z.

9. Diga cómo esperaría que fueran las correlaciones entre los siguientes pares de variables:

a) Cantidad ingerida de alcohol y tiempo de reacción de un conductor ante un peatón que trata de cruzar por un paso de cebra.
b) «Capacidad psicomotriz» de los niños y edad a la que empiezan a andar.
c) Calificaciones obtenidas en Análisis de Datos I y en Metodología de la Psicología por los estudiantes de grado en psicología.
d) Estatura e inteligencia.
e) Deterioro de las actividades cotidianas y calidad de la red social en ancianos.
f) Grado de depresión y grado de ansiedad en pacientes con estrés postraumático.
g) Edad y número de multas de tráfico recibidas en el último año.
h) Grado de hiperactividad y tiempo que el niño es capaz de estarse quieto y sentado.

10. Tras obtener una correlación de 0,80 entre el nivel de consumo de alcohol en jóvenes y sus infracciones de tráfico, concluimos que el uso del alcohol produce un incremento en las infracciones de tráfico; comente esta conclusión.

11. Si las varianzas de las variables X e Y son, respectivamente, 100 y 256, diga entre qué valores puede oscilar la covarianza entre esas variables y por qué.

12. Obtenga los datos que faltan en las siguientes matrices de varianzas-covarianzas y de correlaciones para las variables X, Y y Z:

S =

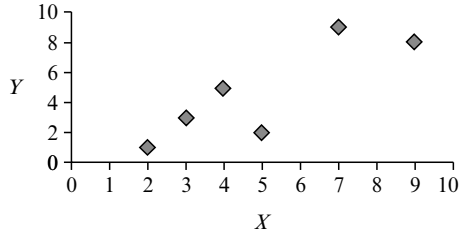
	X	Y	Z
X	5	( )	2,51
Y		7	−4,52
Z			4

R =

	X	Y	Z
X		0,63	( )
Y			( )
Z			

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

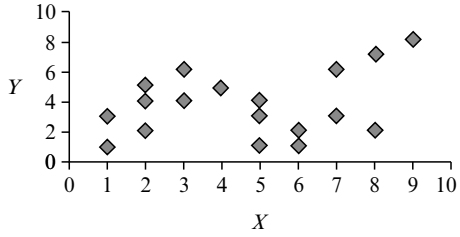
1. *a*)



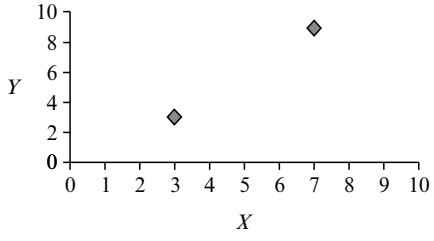
b) $r_{xy} = 0{,}845$. La correlación entre X e Y es positiva y sugiere un valor alto, por lo que existe relación lineal directa entre las variables X e Y.

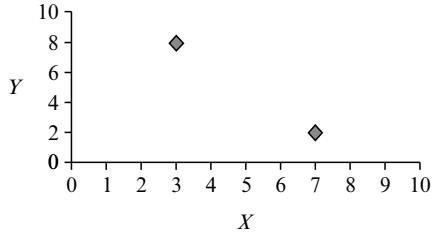
2. $S_{xy} = -7$; $r_{xy} = -0{,}945$. Como se observa, los datos sugieren una fuerte correlación lineal, en este caso inversa, entre las variables X e Y.

3. $r_{xy} = 0{,}294$. Como se observa, tanto los datos como el diagrama de dispersión sugieren que no hay un patrón claro de relación lineal entre las variables X e Y.



4. Sólo +1 o −1, puesto que dos puntos forman una línea recta perfecta. Como se ilustra en las figuras inferiores, si ésta tiene inclinación positiva la correlación será $r_{xy} = 1$, y si es negativa, $r_{xy} = -1$.

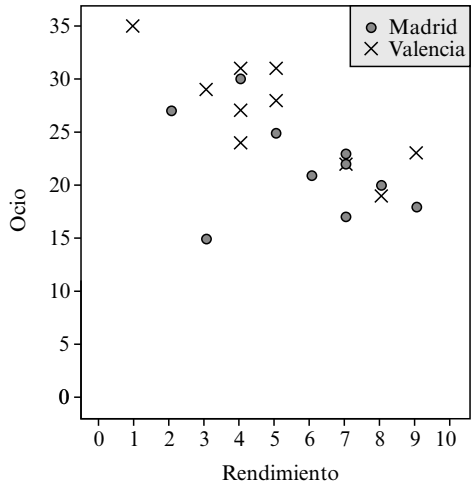




5. $r_{xy} = 0$ indica independencia lineal, pero puede haber relación de otros tipos entre las variables X e Y, como se ilustra en el ejemplo de la figura 5.5.

6. *a*) *SR.*
b) *MC* y *OP.*
c) Entre *SS* y *SH.*
d) Correlaciona más con *MC.*
e) Significa que, a mayor satisfacción con el sueldo obtenido, menor puntuación en oportunidades de promoción y viceversa. Es decir, cuanto más satisfechos están los trabajadores con el sueldo que obtienen, menor es su motivación por las oportunidades de promoción que les ofrece su empresa.

7. *a*) Dedican un mayor número de horas los de Valencia ($\bar{Y} = 26{,}9$) que los de Madrid ($\bar{Y} = 21{,}8$).
b) Mayor en los estudiantes de Valencia ($r_{xy} = -0{,}823$; $r^2_{xy} = 0{,}68$) que en los de Madrid ($r_{xy} = -0{,}409$ y $r^2_{xy} = 0{,}17$).
c)



8. *a*)

	X	*Y*	*Z*
X		0,64	−0,72
Y			0,55
Z			

b) *X* tiene mayor relación lineal con *Z* (pues $r^2_{xy} = 0{,}41$ y $r^2_{xz} = 0{,}52$).

9. *a*) Positiva.
b) Negativa.
c) Positiva.
d) Nula.
e) Negativa.
f) Positiva.
g) Negativa.
h) Negativa.

10. En muchos casos, se encuentran correlaciones entre variables que no tienen una relación causal entre sí, sino que únicamente siguen un patrón de covariación debido a una tercera variable que es causa de ambas simultáneamente. Por eso, se suele admitir que a partir sólo de una correlación no puede inferirse una relación causal. En este ejemplo concreto es probable que haya alguna variable o conjunto de variables que fomenten las infracciones de tráfico, entre ellas el consumo de alcohol, aunque también podrían influir, entre otras, la experiencia al volante, las distracciones y el cansancio.

11. Teniendo en cuenta que $r_{xy} = S_{xy}/S_x S_y$: el máximo es 160 (valor que se alcanzaría en caso de que la correlación entre las variables fuera +1) y el mínimo es −160 (valor que se alcanzaría con $r_{xy} = -1$).

12. Los datos que faltan son los de S_{xy}, r_{xz} y r_{yz}, los cuales pueden obtenerse aplicando las fórmulas [5.5] y [5.7], con lo que se llega a:

$$\mathbf{S} = \begin{array}{c|ccc} & X & Y & Z \\ X & 5 & 3{,}73 & 2{,}51 \\ Y & & 7 & -4{,}52 \\ Z & & & 4 \end{array} \qquad \mathbf{R} = \begin{array}{c|ccc} & X & Y & Z \\ X & & 0{,}63 & 0{,}56 \\ Y & & & -0{,}85 \\ Z & & & \end{array}$$

6 Combinación lineal de variables

6.1. INTRODUCCIÓN

Es frecuente encontrarse en situaciones en las que hay que trabajar con variables que proceden de la combinación de otras variables, especialmente de combinaciones lineales. Por ejemplo, muchas veces la puntuación total de un test psicológico se obtiene sumando las puntuaciones obtenidas en dos o más subescalas. Ya hemos abordado esta cuestión en capítulos anteriores. En éste vamos a exponer la obtención de los principales estadísticos (la media y la varianza) de una variable que se crea como una combinación lineal de dos o más variables. Supongamos, por ejemplo, que representamos por U, V y X a las puntuaciones en tres subescalas, mientras que T representa las puntuaciones totales que se obtienen sumando esas puntuaciones; la expresión formal sería:

$$T_i = U_i + V_i + X_i$$

En estas circunstancias es probable que queramos obtener la media y la varianza de esa nueva variable. Desde luego, un procedimiento correcto para hacerlo consistiría en calcular para cada sujeto su puntuación T_i y luego operar sobre esos valores mediante las fórmulas y procedimientos que ya hemos expuesto en los capítulos anteriores:

$$\bar{T} = \frac{\Sigma T_i}{N} \qquad \text{y} \qquad S_T^2 = \frac{\Sigma (T_i - \bar{T})^2}{N}$$

Sin embargo, este procedimiento puede resultar muy laborioso, sobre todo cuando el número de sujetos y variables es muy grande. En este capítulo vamos a exponer procedimientos que nos permitirán deducir las características de estas variables a partir de las características de las variables componentes y de sus relaciones. Primero estudiaremos el caso más sencillo, el de la suma o resta de dos variables; luego expondremos el caso más frecuente, el de la suma de un número cualquiera de variables, y por último describiremos el caso general, aplicable a cualquier combinación lineal de variables.

6.2. SUMA Y RESTA DE DOS VARIABLES: MEDIA Y VARIANZA

La combinación lineal de variables más sencilla es aquella en la que una variable se define a partir de la simple suma o resta de otras dos variables. Definiremos de esas dos formas las variables T y U:

$$T_i = X_i + Y_i \qquad [6.1]$$

$$U_i = X_i - Y_i \qquad [6.2]$$

Veamos primero cuánto valdrán las medias de las nuevas puntuaciones:

$$\bar{T} = \frac{\Sigma T_i}{N} = \frac{\Sigma(X_i + Y_i)}{N} = \frac{\Sigma X_i}{N} + \frac{\Sigma Y_i}{N} = \bar{X} + \bar{Y}$$

$$\bar{U} = \frac{\Sigma U_i}{N} = \frac{\Sigma(X_i - Y_i)}{N} = \frac{\Sigma X_i}{N} - \frac{\Sigma Y_i}{N} = \bar{X} - \bar{Y}$$

Por su parte, la varianza de T será:

$$S_T^2 = \frac{\Sigma(T_i - \bar{T})^2}{N} = \frac{\Sigma[(X_i + Y_i) - (\bar{X} + \bar{Y})]^2}{N}$$

Reorganizamos el numerador, lo desarrollamos como un binomio al cuadrado y aplicamos la regla de distribución del sumatorio (véase el apéndice del capítulo 1):

$$S_T^2 = \frac{\Sigma[(X_i - \bar{X}) + (Y_i - \bar{Y})]^2}{N} =$$

$$= \frac{\Sigma[(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2 + 2 \cdot (X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{N} =$$

$$= \frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2 + 2 \cdot \Sigma(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{N}$$

Por último, separamos la expresión en tres quebrados diferentes, y en cada uno de ellos reconocemos un elemento ya descrito en capítulos anteriores:

$$S_T^2 = \boxed{\frac{\Sigma(X_i - \bar{X})^2}{N}} + \boxed{\frac{\Sigma(Y_i - \bar{Y})^2}{N}} + 2 \cdot \boxed{\frac{\Sigma(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{N}}$$

$$\downarrow \qquad \downarrow \qquad \downarrow$$

$$S_x^2 \qquad S_y^2 \qquad S_{xy}$$

Dejamos en manos del lector interesado en poner a prueba su pericia algebraica el desarrollo de la demostración de que la varianza de U será similar a la de T, con la única diferencia de que el tercer quebrado irá precedido por un signo negativo, en lugar del signo positivo. Por tanto, podemos resumir lo expuesto hasta aquí de la siguiente forma:

a) *Si se define una variable como la suma de dos variables*, su media es igual a la suma de las medias de esas variables y su varianza es igual a la suma de las varianzas de esas variables más dos veces la covarianza entre ellas. Es decir:

$$\text{Si} \quad T_i = X_i + Y_i$$

$$\text{entonces} \quad \overline{T} = \overline{X} + \overline{Y} \quad \text{y} \quad S_T^2 = S_x^2 + S_y^2 + 2 \cdot S_{xy} \qquad [6.3]$$

b) *Si se define una variable como la diferencia entre dos variables,* su media es igual a la diferencia entre las medias de esas variables y su varianza es igual a la suma de las varianzas de esas variables menos dos veces la covarianza entre ellas. Es decir:

$$\text{Si} \quad T_i = X_i - Y_i$$

$$\text{entonces} \quad \overline{T} = \overline{X} - \overline{Y} \quad \text{y} \quad S_T^2 = S_x^2 + S_y^2 - 2 \cdot S_{xy} \qquad [6.4]$$

Veamos un ejemplo numérico de la aplicación de estas fórmulas. Supongamos que disponemos de las puntuaciones de cuatro personas en dos subtests, X e Y; definimos T como la suma de los subtests y U como su diferencia; es decir, $T_i = X_i + Y_i$ y $U_i = X_i - Y_i$. Si no conociésemos las fórmulas [6.3] y [6.4] obtendríamos las puntuaciones T y U de cada sujeto y hallaríamos las medias y varianzas de esas puntuaciones, tal y como mostramos en la siguiente tabla:

	X	Y	T	U	T^2	U^2	$X \cdot Y$
	6	5	11	1	121	1	30
	3	1	4	2	16	4	3
	5	4	9	1	81	1	20
	2	2	4	0	16	0	4
Suma	16	12	28	4	234	6	57

$$\overline{T} = 28/4 = 7 \qquad \overline{U} = 4/4 = 1$$

$$S_T^2 = \frac{234}{4} - 7^2 = 9{,}5 \qquad S_U^2 = \frac{6}{4} - 1^2 = 0{,}5$$

Las fórmulas [6.3] y [6.4] nos permiten obtener estas características de forma más directa, sobre todo cuando el número de observaciones es muy grande. Para ello, debemos disponer de los siguientes estadísticos de las variables originales, X e Y:

$$\bar{X} = 16/4 = 4 \qquad \bar{Y} = 12/4 = 3 \qquad S_{xy} = 2{,}25$$

$$S_x^2 = 2{,}5 \qquad S_y^2 = 2{,}5$$

Ahora sustituimos en [6.3] y [6.4]:

$$\bar{T} = \bar{X} + \bar{Y} = 4 + 3 = 7 \qquad \bar{U} = \bar{X} - \bar{Y} = 4 - 3 = 1$$

$$S_T^2 = S_x^2 + S_y^2 + 2 \cdot S_{xy} = 2{,}5 + 2{,}5 + 2 \cdot 2{,}25 = 9{,}5$$

$$S_U^2 = S_x^2 + S_y^2 - 2 \cdot S_{xy} = 2{,}5 + 2{,}5 - 2 \cdot 2{,}25 = 0{,}5$$

obteniendo los mismos estadísticos que hemos obtenido directamente con las puntuaciones de T y U.

6.3. SUMA DE *J* VARIABLES

Expondremos ahora un procedimiento más general, aplicable a todos aquellos casos en los que la nueva variable se define como la suma de un número cualquiera de variables. Por comodidad, representaremos a partir de aquí las variables con la misma letra, X, pero con diferentes subíndices para diferenciarlas. Con el primer subíndice nos referiremos a la observación y con el segundo a la variable; así, X_{ij} representará al valor i-ésimo en la variable j-ésima. Expuesto con esta nomenclatura, el procedimiento que vamos a describir es aplicable a toda variable definida de la siguiente forma:

$$T_i = X_{i1} + X_{i2} + \ldots + X_{iJ} \qquad [6.5]$$

Esta expresión no es más que la suma de J variables. Extendiendo lo expuesto al tratar la suma de dos variables, *la media de esta variable es igual a la suma de las medias de las variables sumadas*. Es decir:

$$\text{Si} \qquad T_i = X_{i1} + X_{i2} + \ldots + X_{iJ}$$

$$\text{entonces} \quad \bar{T} = \bar{X}_1 + \bar{X}_2 + \ldots + \bar{X}_J \qquad [6.6]$$

El procedimiento para obtener la varianza de una variable definida de esta forma se simplifica bastante si utilizamos la matriz de varianzas y covarianzas, que ya hemos expuesto en el capítulo anterior. En concreto, en [6.1] se definía una

variable como la suma de dos variables. Si construimos la matriz **S**, o de varianzas y covarianzas, de las variables sumadas, observamos que ésta contiene precisamente los elementos de la fórmula de la varianza, definida en [6.3]: las varianzas de las dos variables y dos veces la covarianza entre ellas. Esto no es casualidad. Ahorramos al lector la demostración de que esto ocurre con la suma de cualquier suma de variables. Es decir, que si una variable se define como la suma de J variables, *su varianza es igual a la suma de todos los elementos de la matriz* **S** *construida con esas J variables* (adviértase que se multiplica por 2 la suma de los elementos que quedan por encima o por debajo de la diagonal). Es decir:

$$\begin{aligned} &\text{Si} && T_i = X_{i1} + X_{i2} + \ldots + X_{iJ} \\ &\text{entonces} && S_T^2 = S_1^2 + S_2^2 + \ldots + S_J^2 \\ & && + 2 \cdot (S_{12} + S_{13} + \ldots + S_{1J} + \ldots S_{(J-1)J}) \end{aligned} \qquad [6.7]$$

Veamos también un ejemplo numérico. Supongamos que tras administrar un test compuesto por cuatro subtests (X_1, X_2, X_3, X_4) a un colectivo obtenemos la siguiente matriz **S**, de varianzas y covarianzas, así como las medias que también se indican para los subtests. Queremos hallar la media y la varianza de una variable, T, definida como la suma de los valores en los cuatro subtests ($T_i = X_{i1} + + X_{i2} + X_{i3} + X_{i4}$):

	X_1	X_2	X_3	X_4
X_1	32	15	7	18
X_2	15	40	22	12
X_3	7	22	20	14
X_4	18	12	14	26
$\bar{X}_J$	118	33	83	56

La media de T se obtiene como la suma de las medias, $\bar{T} = 118 + 33 + 83 + + 56 = 290$. La varianza se obtiene como la suma de los elementos de la matriz **S** de esas cuatro variables:

$$S_T^2 = 32 + 40 + 20 + 26 + 2 \cdot (15 + 7 + 18 + 22 + 12 + 14) = 294$$

6.4. COMBINACIÓN LINEAL DE *J* VARIABLES

Aunque los casos que hemos visto hasta aquí son los más sencillos y frecuentes, hay otros que no encajan en esas situaciones. Lo que vamos a abordar ahora es el caso general, aplicable a cualquier combinación lineal de variables. Los an-

teriores no son más que casos particulares del presente. Vamos a exponer cómo obtener la media y la varianza de una variable que se define de la siguiente forma:

$$T_i = k_1 \cdot X_{i1} + k_2 \cdot X_{i2} + ... + k_J \cdot X_{iJ} + c \qquad [6.8]$$

En esta expresión, k_j es el peso que tienen los valores de la variable X_j. Nótese que lo que hemos desarrollado en los apartados anteriores no son más que casos particulares de la expresión [6.8]. La suma y resta de dos variables serían aquellos casos en los que interviniesen dos variables, y las constantes multiplicadoras fueran ambas iguales a 1 (suma de dos variables) o una igual a 1 y la otra igual a −1 (diferencia de variables). La suma de J variables es el caso en el que todas las k_j son iguales a 1. Además, en todos esos casos la constante sumada, c, es igual a 0.

La media de una variable definida según [6.8] se obtiene aplicando la siguiente definición: *la media de una variable definida como una combinación lineal de variables es igual a la misma combinación lineal de las medias de las variables componentes*. Es decir:

$$\text{Si} \qquad T_i = k_1 \cdot X_{i1} + k_2 \cdot X_{i2} + ... + k_J \cdot X_{iJ} + c$$

$$\text{entonces} \quad \overline{T} = k_1 \cdot \overline{X}_1 + k_2 \cdot \overline{X}_2 + ... + k_J \cdot \overline{X}_J + c \qquad [6.9]$$

La varianza de una combinación lineal de variables se obtiene por el procedimiento que describimos a continuación; omitimos su laboriosa demostración. *La varianza de una variable formada como una combinación lineal de variables es igual a la suma de las varianzas de las variables componentes multiplicadas por los cuadrados de sus coeficientes, más dos veces las covarianzas por pares entre las variables, multiplicadas por los coeficientes correspondientes*. Es decir,

$$\text{Si} \qquad T_i = k_1 \cdot X_{i1} + k_2 \cdot X_{i2} + ... + k_J \cdot X_{iJ} + c$$

$$\text{entonces} \quad S_T^2 = k_1^2 \cdot S_1^2 + k_2^2 \cdot S_2^2 + ... + k_J^2 \cdot S_J^2 + 2 \cdot (k_1 \cdot k_2 \cdot S_{12} + \\ + k_1 \cdot k_3 \cdot S_{13} + ... + k_1 \cdot k_J \cdot S_{1J} + ... + k_{J-1} \cdot k_J \cdot S_{(J-1)J}) \qquad [6.10]$$

PROBLEMAS Y EJERCICIOS

1. Obtenga:

a) La matriz de correlaciones a partir de la siguiente matriz de varianzas-covarianzas.

b) Si se define la variable $V = X_3 - X_1$, calcule S_V^2.

$$\mathbf{S} = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline X_1 & 25 & 10 & -5{,}7 & 26{,}25 \\ X_2 & 10 & 16 & 0{,}6 & 2{,}8 \\ X_3 & -5{,}7 & 0{,}6 & 9 & 5{,}25 \\ X_4 & 26{,}25 & 2{,}8 & 5{,}25 & 49 \end{array}$$

2. Considerando las matrices incompletas de correlaciones y varianzas-covarianzas que se muestran a continuación, obtenga las varianzas de las siguientes variables:

a) $U = X_1 + X_3$

b) $V = X_4 - X_1$

$$\mathbf{R} = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline X_1 & & ( ) & 0{,}2 & 0{,}4 \\ X_2 & & & ( ) & -0{,}5 \\ X_3 & & & & ( ) \\ X_4 & & & & \end{array}$$

$$\mathbf{S} = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline X_1 & 4 & ( ) & ( ) & ( ) \\ X_2 & & 9 & ( ) & ( ) \\ X_3 & & & 36 & ( ) \\ X_4 & & & & 49 \end{array}$$

3. Complete la matriz de correlaciones y la de varianzas-covarianzas (recuerde que ambas matrices son simétricas) del ejercicio anterior, sabiendo que: entre las variables X_1 y X_2 existe una relación lineal nula, así como una relación lineal directa perfecta entre las variables X_2 y X_3; y $S_{x_3x_4} = -42$.

4. Una prueba de capacidad verbal, formada por tres ítems *linealmente independientes*, ha sido aplicada a un estudiante de primer curso del Grado en Psicología, obteniendo una puntuación directa en la prueba igual a 9. Sabiendo que la puntuación en la prueba es igual a la suma de las puntuaciones de cada ítem, obtenga la puntuación típica del estudiante en la prueba, teniendo en cuenta los datos que aparecen en la siguiente tabla:

	I_1	I_2	I_3
S_X	2	8	5
$\bar{X}$	1	3	4

5. Si a la matriz de varianzas-covarianzas del ejercicio 1 le añadimos el vector de medias que aparece más abajo, obtenga la media y la varianza de las siguientes variables: $U = 0{,}2 \cdot X_1 - 0{,}8 \cdot X_2$; $V = 5 \cdot X_2 + 2 \cdot X_4$.

	X_1	X_2	X_3	X_4
$\bar{X}$	3	7	1	8

6. Demuestre qué condición se ha de satisfacer para que la varianza de la suma de dos variables sea igual a la varianza de su diferencia.

7. Para mejorar el tiempo de respuestas, *TR*, ante situaciones de alarma, se aplicó un programa de mejora a un grupo de personal de seguridad. Para ello se midió el *TR* de cada participante ante una situación de peligro antes y después del programa. La variable eficacia del programa, *EP*, se evaluó mediante la diferencia del *TR* antes y después del programa. Sabiendo que la varianza del TR_{Antes} fue igual a 25, que la varianza del $TR_{Después}$ fue igual a 16, y que la de *EP* fue igual a 17, obtenga el valor de la correlación entre las variables TR_{Antes} y $TR_{Después}$.

8. Supongamos que $U_i = X_i - Y_i$ y que la varianzas de X e Y son iguales, aunque las desconocemos. Obtenga el valor de S_U^2 si:

a) Existe una relación lineal directa perfecta entre X e Y.
b) Existe una relación lineal inversa perfecta entre X e Y.
c) Hay independencia lineal entre X e Y.

9. Se han medido las variables U, V, X e Y en una muestra, calculándose la matriz de varianzas-covarianzas que aparece a continuación, así como las medias. Entonces:

a) Si se define la variable $W = U + V + Y$, calcule la media y la varianza de W.
b) Si se define la variable $T = \dfrac{V + X + Y}{3}$, calcule la media y la varianza de T.

S =		U	V	X	Y
	U	4	3	9,6	2,8
	V	3	9	10,8	−6,3
	X	9,6	10,8	36	4,2
	Y	2,8	−6,3	4,2	49
Medias		2	4	5	8

10. Considerando la matriz de varianzas-covarianzas y el vector de medias que se presentan más abajo, calcule la media y la varianza de:

a) La variable $T = X + Y + V + W$.
b) La variable $U = (X/4) - V$.

S =		X	Y	V	W
	X	25	1	12	6
	Y	1	4	−9,6	1,8
	V	12	−9,6	64	−2,4
	W	6	1,8	−2,4	9
Medias		6	3	10	5

11. La puntuación V, utilizada en un proceso de evaluación clínica, se obtiene mediante el promedio de las puntuaciones típicas de los valores obtenidos en dos cuestionarios (A y B) de personalidad. ¿Cuánto valdrá la varianza de V si la correlación entre las puntuaciones de esas dos pruebas vale: *a*) 1; *b*) 0 y *c*) −1?

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*)

R =		X_1	X_2	X_3	X_4
	X_1	1	0,5	−0,38	0,75
	X_2	0,5	1	0,05	0,1
	X_3	−0,38	0,05	1	0,25
	X_4	0,75	0,1	0,25	1

b) $S_V^2 = 45{,}4$.

2. $S_U^2 = 44{,}8$; $S_V^2 = 41{,}8$.

3.

$$\mathbf{R} = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ X_1 & & 0 & -0{,}2 & 0{,}4 \\ X_2 & & & 1 & -0{,}5 \\ X_3 & & & & -1 \\ X_4 & & & & \end{array}$$

$$\mathbf{S} = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ X_1 & 4 & 0 & -2{,}4 & 5{,}6 \\ X_2 & & 9 & 18 & -10{,}5 \\ X_3 & & & 36 & -42 \\ X_4 & & & & 49 \end{array}$$

4. $z = 0{,}104$.

5. U: $\bar{U} = -5$; $S_U^2 = 8{,}04$. V: $\bar{V} = 51$; $S_V^2 = 652$.

6. La condición es que $S_{x_1x_2}$ sea igual a cero.

7. $r_{Antes\,Después} = 0{,}6$.

8. Como $S_y^2 = S_x^2$, entonces $S_U^2 = 2 \cdot S_x^2 \cdot (1 - r_{xy})$. Por tanto:

a) Si $r_{xy} = 1$, entonces $S_U^2 = 0$.
b) Si $r_{xy} = -1$, entonces $S_U^2 = 4 \cdot S_x^2$.
c) Si $r_{xy} = 0$, entonces $S_U^2 = 2 \cdot S_x^2$.

9. *a*) $\bar{W} = 14$; $S_W^2 = 61$.
b) $\bar{T} = 5{,}667$; $S_T^2 = 12{,}378$.

10. *a*) $\bar{T} = 24$; $S_T^2 = 119{,}6$ (el mismo resultado si se suman todos los elementos de la matriz de varianzas-covarianzas).
b) $\bar{U} = -8{,}5$; $S_U^2 = 59{,}5625$.

11. Se demuestra que, en estas condiciones, $S_V^2 = \frac{1}{2} + \frac{1}{2} \cdot r_{AB}$, entonces,

a) Si $r_{AB} = 1$; $S_V^2 = 1$.
b) Si $r_{AB} = 0$; $S_V^2 = \frac{1}{2}$.
c) Si $r_{AB} = -1$; $S_V^2 = 0$.

7 Regresión lineal

7.1. INTRODUCCIÓN

Probablemente haya algunos lectores que, al terminar el capítulo 5, hayan tenido la sensación de que las ideas expuestas en él se quedaban cortas. Concluir estableciendo la presencia o ausencia de relación lineal, así como su grado, no parece suficiente. Efectivamente, se podrían explotar más las relaciones lineales y encontrarles otras utilidades. Por ejemplo, continuando con los ejemplos del capítulo sobre la correlación, podemos preguntarnos si sabiendo que existe una relación lineal directa entre la inteligencia y el rendimiento y conociendo el valor de inteligencia de una persona, ¿no podríamos utilizar la información sobre la relación entre esas variables para hacernos una idea, una predicción, de cuál podría ser su rendimiento al final del curso? Como la relación es directa, si esta persona tiene una inteligencia alta parece razonable esperar (o predecir) un rendimiento alto, mientras que una inteligencia baja nos haría esperar un rendimiento bajo. Lo mismo podríamos hacer en el segundo ejemplo con un individuo del que conociéramos sólo su tiempo de respuesta. Por el contrario, dada la ausencia de relación lineal entre la estatura y la inteligencia, conocer las puntuaciones de la primera no nos ayudaría a hacer conjeturas sobre las puntuaciones de la segunda.

En definitiva, la cuestión principal que vamos a abordar en este capítulo es la utilización de la información contenida en las relaciones lineales observadas entre variables para, conociendo el valor en una variable, hacer una conjetura creíble sobre su valor en otra. En este ámbito, el término *regresión* se utiliza como sinónimo de *predicción,* a pesar de que etimológicamente estos términos no tengan ninguna relación. La razón es de naturaleza histórica, dado que estas técnicas se desarrollaron en el contexto de los estudios realizados por Francis Galton sobre los fenómenos de la herencia (en el apartado 7.3.4 se explica el origen del término).

Las situaciones prototipo a las que se aplican las técnicas de regresión son aquellas en las que primero se miden dos variables, X e Y, en un conjunto de unidades; posteriormente se da la circunstancia de que de una unidad adicional se conoce su valor en una de ellas (por ejemplo X) y se quiere hacer una predicción o conjetura para esa unidad en Y. Quizá algún lector esté ya anticipando procedimientos para hacer tales pronósticos. Uno más bien burdo y simple con-

sistiría en calcular la media de las unidades en la variable que se quiere predecir y utilizar ese valor medio como predicción. Si los *N* individuos con valores conocidos tienen una media igual a 8, pronosticar al nuevo individuo un 8 es una opción razonable. Sin embargo, en esta predicción no interviene el valor de ese individuo en la otra variable ni la relación entre *X* e *Y*. De hecho, en este procedimiento a todo nuevo individuo se le haría la misma predicción. Otro procedimiento podría consistir en buscar alguno de los individuos con valores conocidos, que tuviera en la variable que utilizamos para pronosticar el mismo valor que el individuo del que queremos hacer predicciones, y pronosticarle en la otra el mismo valor que obtuvo aquél. Esta estrategia proporcionaría mejores predicciones que la asignación de la media, pero todavía no se explota completamente la relación entre las variables, dado que cada predicción se basaría en una sola observación. Por el contrario, los modelos de regresión son instrumentos de predicción basados en los *N* pares de valores; proporcionarán pronósticos en general más precisos, aunque esta precisión dependerá del grado en que los puntos se ajusten a una función lineal.

Por otra parte, las predicciones tenderán a ser más ajustadas cuanto mayor sea la información en la que se basen. Así, si los pronósticos en rendimiento no sólo se basan en la relación de esta variable con la inteligencia, sino también con otras variables relevantes, como la motivación, la memoria, etc., las predicciones serán aún más precisas. Cuando nos basemos en una sola variable predictora hablaremos de *regresión simple*, mientras que al utilizar más de una variable predictora hablaremos de *regresión múltiple*.

Aunque entre las variables pueden existir relaciones de muchos tipos, no sólo lineales, nosotros nos vamos a restringir a éstas por tres razones: *a*) son las más sencillas; *b*) en muchos casos son suficientes para describir las relaciones entre las variables, y *c*) nos bastan para exponer la lógica en la que se basa la búsqueda de modelos predictivos. Para aquellos lectores interesados en la ampliación a otros modelos, baste con indicar que mientras que la lógica en la que se basan es exactamente la misma, lo que cambia es el número de coeficientes que intervienen en el modelo y sus fórmulas. En el libro de Amón (1993) se pueden consultar las fórmulas apropiadas para las funciones cuadrática, potencial, exponencial y logarítmica.

En este capítulo, tras repasar algunas generalidades sobre los modelos lineales y establecer una terminología, describiremos la forma de trabajar con los modelos de regresión simple, separando las tareas de identificación, valoración y aplicación del modelo. En el apéndice de este capítulo trataremos la extensión de estos conceptos a la regresión múltiple, además de exponer las demostraciones de algunas de las fórmulas o relaciones de la regresión.

7.2. FUNCIONES LINEALES

La relación entre dos variables, *X* e *Y*, es lineal si es de la forma:

$$Y = A + B \cdot X \qquad [7.1]$$

donde A y B son dos constantes. Por ejemplo, la relación entre el diámetro (X) y el perímetro (Y) de un círculo es:

$$Y = 3{,}1416 \cdot X$$

Ésta es una función lineal en la que A es igual a 0 y B es igual a 3,1416 (o número π). El perímetro de un círculo con un diámetro de 20 centímetros será igual a $3{,}1416 \cdot 20 = 62{,}832$. Otro ejemplo es la relación entre el consumo de electricidad y la cantidad facturada. Supongamos que la cantidad que cobra la compañía eléctrica por mantener contratada una cierta potencia es de 6 euros al mes y que cobra 0,125 euros por cada kw/h. La relación entre el consumo (X) y el importe de la factura será lineal, siendo las constantes A y B iguales a 6 y 0,125, respectivamente:

$$Y = 6 + 0{,}125 \cdot X$$

El importe de un recibo mensual en el que se hayan consumido 300 kw/h será: $Y = 6 + 0{,}125 \cdot 300 = 43{,}5$ euros. Las relaciones entre variables que adoptan la forma [7.1] reciben el nombre de funciones lineales, porque al representarlas en unos ejes de coordenadas aparecen como líneas rectas. En la figura 7.1 aparecen las representaciones de los dos ejemplos anteriores.

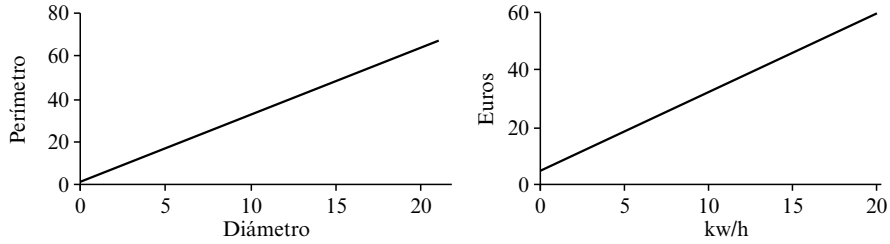


Figura 7.1.—Representación gráfica de dos funciones lineales: *a*) relación entre el diámetro y el perímetro de una circunferencia, y *b*) relación entre el consumo de electricidad y el importe del recibo (véanse los ejemplos en el texto).

Cualquier función lineal se puede identificar completamente mediante dos cantidades. La primera es la constante sumada, representada por A, que recibe el nombre de *origen*. Corresponde al valor que adopta la variable Y cuando la variable X adopta el valor 0. Se llama así porque, como se aprecia en la figura 7.1, indica el punto en el que la recta corta el eje de ordenadas (0 en la primera y 6 en la segunda). Cuando A es una cantidad negativa, la recta corta al eje de ordenadas por debajo del cruce de los ejes. La segunda cantidad es la constante multiplicada, representada por B, que recibe el nombre de *pendiente*. Se interpreta en términos de tasa de cambio, es decir, indica en cuánto cambia el valor de Y por cada incremento de una unidad en X. Así, en nuestro primer ejemplo el pe-

rímetro se incrementaría en 3,1416 por cada unidad que se incrementa el diámetro. En el segundo ejemplo el recibo se incrementa en 12,5 céntimos por cada unidad que incrementemos el número de kw/h consumidos. Cuando B es una cantidad positiva, los incrementos en X se acompañan de incrementos en Y; la recta está inclinada hacia arriba. Cuando es negativa, los incrementos en X se acompañan de decrementos en Y; la recta se inclina hacia abajo.

En los dos ejemplos anteriores las relaciones lineales entre variables son perfectas; el valor de Y se puede obtener con precisión aplicando la ecuación lineal. Sin embargo, en ciencias sociales nunca nos encontramos con situaciones como éstas. El trabajo en el contexto de la regresión es algo diferente al que acabamos de ver; incluso podríamos decir que, en cierto sentido, tiene una orientación inversa. Mientras que en el ejemplo del consumo eléctrico el importe del recibo (valores de Y) se genera precisamente mediante la ecuación lineal, en los estudios realizados en psicología dispondremos de observaciones hechas en las dos variables estudiadas y nos preguntaremos hasta qué punto la relación entre ellas se parece a (y, por tanto, podría describirse como) un modelo lineal.

En ciencias sociales, la tarea para la que se utiliza la regresión, en conexión con la correlación, es la búsqueda de variables X que expliquen las variaciones en Y. Así, en el capítulo 5 (véase la figura 5.1*b*) exponíamos el ejemplo de la relación entre la velocidad de ejecución y el número de errores. El diagrama de dispersión de los pares de valores observados no formaba una línea recta perfecta, aunque se apreciaba una tendencia a la linealidad inversa. En ese ejemplo probablemente haya terceras variables que también tengan un efecto sobre el número de errores. De hecho, es bien sabido que en este tipo de tareas una mayor agudeza perceptiva y una mayor introversión están asociadas a un mejor rendimiento. Estas variables generarán unas oscilaciones que no quedan explicadas si nuestro estudio se restringe exclusivamente a las relaciones entre la velocidad y el número de errores. Sin embargo, podemos decir que el grado en que el diagrama de dispersión entre estas dos variables se parezca a un modelo lineal será un índice del grado en que la variable X sea capaz de explicar por sí sola la variable Y.

En muchos casos contaremos con la constatación de unas oscilaciones para las que en principio no tenemos explicación. No obstante, eso no nos impedirá buscar predictoras. Veamos otro ejemplo. Uno de los tipos de tareas laborales más estudiadas por los psicólogos incluye aquellas que exigen mantener la atención muy concentrada durante largos períodos de tiempo para detectar la presencia de un cierto evento que ocurre con muy poca frecuencia. Son las llamadas tareas de *atención sostenida* o de *vigilancia.* El interés por estas tareas surgió en conexión con la optimización del uso de los radares durante la segunda guerra mundial, pero después su estudio se ha extendido a muchos otros campos, como por ejemplo el control de calidad en la industria. Una tarea típica en este contexto consistiría en observar el paso en una cinta transportadora de las piezas fabricadas y en detectar a simple vista aquellas que pudieran ser defectuosas. El rendimiento en este tipo de tareas se puede medir mediante el porcentaje de piezas defectuosas detectadas. Supongamos que un fabricante, cansado de observar una gran variabilidad en el rendimiento de los trabajadores en ese puesto y deseoso de encontrar un procedimiento que le permita anticipar qué trabajadores rendi-

rían mejor, encarga un estudio sobre la cuestión a un psicólogo. Dadas las características de la tarea, el psicólogo parte de la sospecha de que el rendimiento en este tipo de tareas puede estar relacionado con ciertas variables de personalidad, especialmente la introversión. Desde su punto de vista, dado que los introvertidos se distraen menos tenderán a rendir más. Para comprobarlo, prepara una simulación con unos conjuntos de piezas que incluyen algunas piezas defectuosas y mide el rendimiento de los miembros del equipo de control de calidad con este material. Posteriormente les pasa un test de introversión y empareja cada puntuación en introversión (X) con el rendimiento del operador (Y). Tras analizar los datos encuentra una tendencia a la linealidad en esa relación. Aunque para él esta tendencia sea un apoyo a su interpretación del fenómeno, en realidad lo único que puede decir es que la relación entre las variables «introversión» (X) y «porcentaje de piezas defectuosas detectadas» (Y) se puede describir mediante lo que en el capítulo 5 llamábamos una relación directa. Si esta relación es suficientemente estrecha, es decir, si r_{xy} se aleja de cero, habrá descubierto una buena regla predictiva. Mediante la evaluación de la introversión de los trabajadores podrá hacer una conjetura (o predicción) del rendimiento en ese puesto laboral, pero de ninguna manera esto se puede considerar como una demostración de que la introversión sea la causa de las variaciones en rendimiento.

En pocas palabras, lo que se estudia en el contexto de la regresión simple es la posibilidad de aplicar un modelo lineal, basado en una única variable, como instrumento para la predicción de una segunda variable.

7.3. REGRESIÓN SIMPLE

Definiremos primero algunos términos. A partir de ahora llamaremos *variable predictora* a la que utilizamos para hacer pronósticos y *variable criterio* a aquella en la que se hacen pronósticos; cuando la variable predictora es una variable que se manipula, o se cumplen las condiciones que permiten hacer inferencias causales entre ésta y la variable criterio, entonces reciben también los nombres de *variable independiente* y *variable dependiente*, respectivamente (León y Montero, 2003). Lo que vamos a determinar es la recta de regresión de Y sobre X, es decir, aquella que permite predecir Y a partir de los valores de X. En los subíndices de los coeficientes A y B aparecerán las letras YX (A_{yx} y B_{yx}), indicando que se refieren a la recta de regresión de Y sobre X. Por simplicidad, en muchos casos omitiremos esos índices, pero el lector debe estar sobre aviso de que en la otra recta de regresión, la de X sobre Y, aparecerían esos índices en orden inverso y las fórmulas cambiarían. Igualmente, omitiremos los subíndices del coeficiente de correlación de Pearson cuando éste se refiera a las variables X e Y. Por otro lado, distinguiremos entre el valor que cada individuo tiene en la variable Y del valor que se predice en esa variable a ese mismo individuo, valor que designaremos como Y'; mientras Y es el valor empírico, Y' es la predicción o conjetura que se hace mediante el modelo de regresión.

Las tareas que se desarrollan en el contexto de la regresión son fundamentalmente tres. En primer lugar, la identificación del modelo, es decir, la construcción

de una ecuación de la forma [7.1] que sirva como representación de la relación entre las variables. En segundo lugar, la valoración del modelo identificado, mediante la evaluación de su capacidad predictiva. En tercer lugar, la aplicación de los modelos a la predicción en situaciones reales. En los apartados siguientes abordaremos sucesivamente estas tres tareas.

7.3.1. Identificación del modelo: ecuaciones

Cuando los puntos no están exactamente en línea recta, la determinación de un modelo lineal representativo puede parecer una tarea algo arbitraria. Así, a los pares de valores en las variables X e Y que aparecen en la figura 7.2 podríamos representarlos en principio por cualquiera de las rectas que aparecen en la figura, pues lo único que las diferencia son los valores de sus orígenes y pendientes.

X	*Y*
4	1
10	5
8	2
2	3

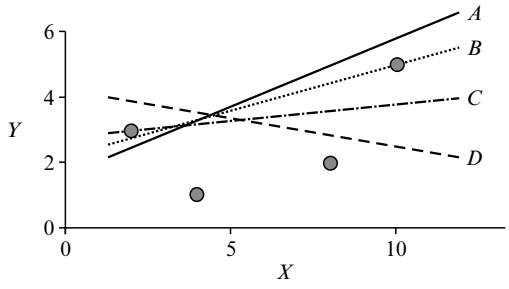


Figura 7.2.—Algunas rectas candidatas a representar la relación entre X e Y que sintetizan los cuatro pares de valores que aparecen en la tabla.

¿Cuál de esas rectas es mejor representación de la relación entre las dos variables? Parece claro que esa decisión no se puede dejar a la apreciación subjetiva ni a otros criterios personales. Como en los capítulos anteriores, vamos a exponer técnicas que permitan llegar inequívocamente a modelos completamente determinados mediante criterios objetivos. De todos los procedimientos posibles, en los modelos de regresión se utiliza el llamado *criterio de mínimos cuadrados*, que se basa en la siguiente pregunta: ¿qué modelo lineal es aquel con el que, en caso de haberlo utilizado como predictor de los valores de Y, hubiéramos cometido una cantidad de *error* lo más pequeña posible? Esto quiere decir que si nos dedicáramos a poner a prueba un número indeterminadamente grande de modelos

lineales y calculáramos algún índice de la capacidad predictiva de cada uno de ellos, nos quedaríamos con aquel modelo para el que la capacidad predictiva fuese mejor. Cuando se aplica este procedimiento con el índice del error que definiremos unas líneas más adelante, se está aplicando el criterio de mínimos cuadrados; lo que se busca es el modelo lineal para el que ese índice sea lo más pequeño posible.

Nuestro índice de la capacidad predictiva de los modelos lineales, para un conjunto dado de pares de valores, es el promedio de los errores elevados al cuadrado. Pongamos a prueba este procedimiento con los datos del ejemplo anterior. Vamos a medir con este índice la capacidad predictiva de un modelo cualquiera para esos datos, como por ejemplo el modelo:

$$Y' = 1 + 0{,}50 \cdot X$$

Para ello calculamos los valores de Y que asociaríamos a cada uno de los valores de X mediante la fórmula anterior (recordemos que se representan por Y'); después calcularemos la diferencia entre cada valor pronosticado y el valor real de Y que se asoció a X (es decir, $Y - Y'$), y elevaremos esas diferencias al cuadrado; por último, las sumaremos y dividiremos por N; veámoslo con la recta del ejemplo (representado en la figura 7.3 superior):

Ejemplo. Calculamos el promedio de los errores al cuadrado cometidos con la recta arbitrariamente elegida:

TABLA 7.1

X	Y	$Y' = 1 + 0{,}5 \cdot X$	$(Y - Y')$	$(Y-Y')^2$
4	1	3	−2	4
10	5	6	−1	1
8	2	5	−3	9
2	3	2	1	1
				15

$$\frac{\sum(Y_i - Y'_i)^2}{N} = \frac{15}{4} = 3{,}75$$

Pero esa recta es arbitraria. Vamos a ver cuál es la auténtica recta de regresión. El criterio de mínimos cuadrados es aquel que nos proporciona el modelo lineal para el que el promedio de los errores al cuadrado:

$$\frac{\sum(Y_i - Y'_i)^2}{N} \qquad [7.2]$$

es lo más pequeño posible. Quizá algún lector se esté preguntando por qué no se toman los errores simples, en lugar de elevarlos al cuadrado. La razón es que los

errores pueden ser positivos o negativos y la suma compensaría unos con otros, dando una impresión engañosa del tamaño de los errores. Por el contrario, al elevarlos al cuadrado todos son positivos, y su suma será tanto mayor cuanto mayores sean sus distancias en promedio hasta la recta, sin importar si esas diferencias lo son por exceso o por defecto.

La recta que hace mínima la expresión [7.2] se obtiene sustituyendo en ella Y' por su valor ($Y'_i = A + B \cdot X_i$), derivando con respecto a A y con respecto a B, igualando a cero y despejando. Las fórmulas que se obtienen son las siguientes:

$$B_{yx} = \frac{N \cdot \sum XY - \sum X \cdot \sum Y}{N \cdot \sum X^2 - (\sum X)^2} \qquad [7.3]$$

$$A_{yx} = \bar{Y} - B_{yx} \cdot \bar{X} \qquad [7.4]$$

Aplicamos estas fórmulas a los datos del ejemplo de la tabla 7.1 y hallamos nuestro índice de la capacidad predictiva del modelo, construido con las cantidades calculadas con estas fórmulas; primero obtenemos los coeficientes de la ecuación:

X	Y	X^2	$X \cdot Y$
4	1	16	4
10	5	100	50
8	2	64	16
2	3	4	6
24	11	184	76

$$B_{yx} = \frac{4 \cdot 76 - 24 \cdot 11}{4 \cdot 184 - 24^2} = 0{,}25$$

$$A_{yx} = \frac{11}{4} - 0{,}25 \cdot \frac{24}{4} = 1{,}25$$

Ahora calculamos los errores, hallamos sus cuadrados y promediamos:

X	Y	$Y' = 1{,}25 + 0{,}25 \cdot X$	$(Y - Y')$	$(Y-Y')^2$
4	1	2,25	−1,25	1,5625
10	5	3,75	1,25	1,5625
8	2	3,25	−1,25	1,5625
2	3	1,75	1,25	1,5625
				6,25

El promedio de los errores al cuadrado será:

$$6{,}25/4 = 1{,}5625$$

Como se puede apreciar, esta cantidad es menor que la obtenida con la ecuación arbitraria que utilizábamos anteriormente (1,5625 frente a 3,75). De hecho, el procedimiento de derivación de las fórmulas [7.3] y [7.4] nos garantiza que es menor que la que se obtendría con cualquier otro modelo lineal (véase la figura 7.3).

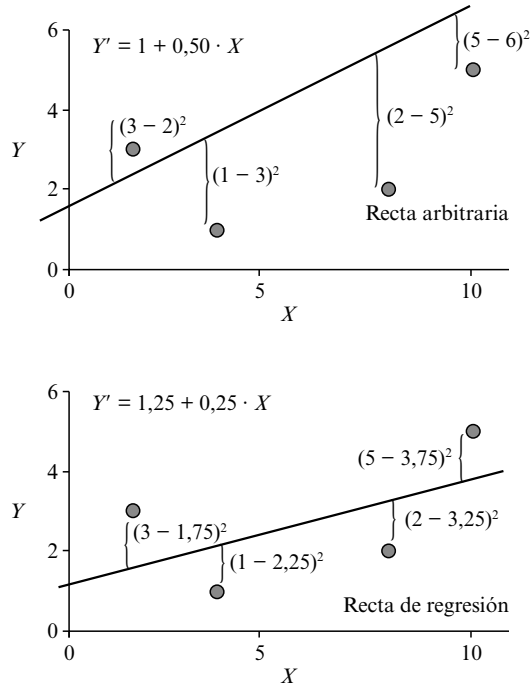


Figura 7.3.—Representación de los errores cuadráticos con respecto a una recta arbitraria (figura superior) y con respecto a la recta de regresión (figura inferior).

Hay otra fórmula alternativa a la de la pendiente en puntuaciones directas, que en ocasiones puede ser más útil que [7.3] y que se basa en otros estadísticos ya conocidos:

$$B_{yx} = r_{yx} \cdot \frac{S_y}{S_x} \qquad [7.5]$$

Esta fórmula será más práctica cuando, antes de hallar las ecuaciones de regresión, hayamos obtenido las medias y varianzas de las variables y la correlación entre ellas. Establece que *la pendiente de la ecuación de regresión es igual a la correlación entre las variables multiplicada por el cociente de sus desviaciones típicas* (la de la variable criterio en el numerador).

Todo el desarrollo que hemos hecho hasta aquí ha sido con puntuaciones directas, pero podríamos repetirlo idénticamente con otros tipos de puntuaciones.

Es bastante frecuente hacerlo en puntuaciones típicas, por razones que no vienen al caso en este momento. Para referirnos a las ecuaciones de regresión en puntuaciones típicas utilizaremos letras minúsculas. Las fórmulas que se obtienen al aplicar el procedimiento que hemos descrito a las puntaciones típicas son bien sencillas. El origen es siempre necesariamente igual a 0; es decir, la recta siempre pasa por la intersección de los ejes. Observando la ecuación [7.4] advertimos en seguida este hecho, dado que las medias de las puntuaciones típicas es, en ambas variables, igual a 0. La pendiente es igual al valor de la correlación de Pearson entre las variables. En la fórmula [7.5] se aprecia de inmediato la razón: las desviaciones típicas de las puntuaciones z son siempre iguales a 1. En resumen:

$$a_{yx} = 0 \qquad [7.6]$$

$$b_{yx} = r_{yx} \qquad [7.7]$$

Con lo que $z'_y = r_{xy} \cdot z_x$.

7.3.2. Valoración del modelo: coeficiente de determinación

Es importante separar explícitamente la cuestión de la identificación del modelo lineal de la cuestión de su valoración. El criterio de mínimos cuadrados nos garantiza que el modelo de regresión obtenido sea precisamente aquel para el que el ajuste de un conjunto dado de puntos es óptimo. Sin embargo, ese ajuste «óptimo» puede no ser muy bueno. De ese grado de ajuste dependerá su utilidad predictiva. Dicho de otra forma, el mejor de los ajustes posibles para un conjunto de puntos puede ser un ajuste muy pobre (recuérdense los ejemplos de la figura 5.2). Tal y como hemos visto en el capítulo 5, el ajuste de unos puntos a un modelo lineal se mide mediante el coeficiente de correlación de Pearson. Conviene, por tanto, que nos detengamos en la interpretación que se debe dar a ese coeficiente en el contexto de la regresión. El desarrollo de este apartado se divide en dos partes. En la primera se definirán algunos conceptos y se establecerán tres relaciones básicas; en la segunda se aplicarán esas relaciones a la valoración de los modelos lineales.

Tomaremos como punto de partida de nuestra exposición la primera de esas tres relaciones básicas, que establece lo siguiente:

La puntuación de un individuo u observación en la variable criterio es igual al pronóstico hecho para ese individuo u observación mediante el modelo de regresión, más el error que se comete al hacer ese pronóstico. Es decir:

$$\boxed{Y} = \boxed{Y'} + \boxed{(Y - Y')} \qquad [7.8]$$

Puntuación empírica	Pronóstico	Error en el pronóstico

Podemos aplicar a esta expresión las propiedades que conocemos para las combinaciones lineales de variables (capítulo 6), puesto que Y es una suma de dos variables: el pronóstico y el error. En concreto, la varianza de las puntuaciones del criterio (S_y^2) será igual a la suma de la varianza de los pronósticos más la varianza de los errores, más dos veces la covarianza entre ambos (véase la fórmula [6.3]):

$$S_y^2 = S_{y'}^2 + S_{(y-y')}^2 + 2 \cdot S_{y'(y-y')} \qquad [7.9]$$

Veamos a qué es igual cada uno de estos tres elementos:

a) $S_{y'}^2$ no es más que la varianza de los pronósticos hechos con la recta.

b) Para mostrar a qué es igual la varianza de los errores comenzaremos por demostrar que la media de los errores es igual a cero. Representando por E a los errores ($E_i = Y_i - Y'_i$):

$$\bar{E} = \frac{\sum E_i}{N} = \frac{\sum (Y_i - Y'_i)}{N} = \frac{\sum [Y_i - (A + B \cdot X_i)]}{N} = \frac{\sum Y_i}{N} - \frac{N \cdot A}{N} - B \cdot \frac{\sum X}{N} =$$
$$= \bar{Y} - (\bar{Y} - B \cdot \bar{X}) - B \cdot \bar{X} = \bar{Y} - \bar{Y} + B \cdot \bar{X} - B \cdot \bar{X} = 0$$

Obtenemos ahora la varianza de los errores:

$$S_E^2 = \frac{\sum (E_i - \bar{E})^2}{N} = \frac{\sum (E_i - 0)^2}{N} = \frac{\sum E_i^2}{N} = \frac{\sum (Y_i - Y'_i)^2}{N}$$

En esta última expresión reconocemos lo que en el apartado anterior definíamos como el promedio de los errores cuadráticos (fórmula [7.2]). Ahora comprobamos que ese promedio es también igual a la varianza de los errores en los pronósticos. A esta varianza la representaremos a partir de aquí como $S_{y \cdot x}^2$, la llamaremos *error cuadrático medio,* y con ella nos referiremos a la varianza de los errores cometidos al pronosticar la variable Y a partir de la variable X (o mediante la ecuación de regresión de Y sobre X).

c) La covarianza entre los pronósticos y los errores es necesariamente nula, tal y como se demuestra en el apéndice del presente capítulo (demostración 1).

Por tanto, de la fórmula [7.9] pasamos a la siguiente, que constituye la segunda relación básica para la valoración de un modelo lineal simple:

$$S_y^2 = S_{y'}^2 + S_{y \cdot x}^2 \qquad [7.10]$$

Expresada en palabras, la fórmula [7.10] nos indica que la varianza del criterio se puede descomponer en dos partes aditivas, una constituida por la varianza de los pronósticos hechos mediante la recta de regresión y la otra por el error cuadrático medio, o varianza de los errores en los pronósticos.

La tercera relación básica para la valoración de un modelo lineal simple es la que conecta los elementos de la expresión [7.10] con la correlación de Pearson. La fórmula, que también demostramos en el mismo apéndice (demostración 4), es la siguiente:

$$S^2_{y \cdot x} = S^2_y \cdot (1 - r^2) \qquad [7.11]$$

Queremos destacar dos últimas cuestiones relativas a la segunda relación básica. La primera es que la obtención de $S^2_{y'}$ y $S^2_{y \cdot x}$ constituye una especie de disección de S^2_y, dado que supone calcular qué parte de esta última (la varianza del criterio) suponen $S^2_{y'}$ y $S^2_{y \cdot x}$, respectivamente. La obtención de estos factores aditivos constituye lo que se suele denominar *descomposición de la varianza del criterio*; más adelante expondremos un ejemplo numérico de ello.

La segunda cuestión se refiere a la relación entre $S^2_{y'}$ y $S^2_{y \cdot x}$, por un lado, y S^2_x por el otro. En el apéndice mostramos que los pronósticos mantienen una correlación perfecta con las puntuaciones en X (demostración 2), mientras que los errores son linealmente independientes de X (demostración 3). Por ello, la varianza de los pronósticos también recibe los nombres de varianza dependiente de X, varianza común con X, varianza determinada por X o *varianza explicada por X*, mientras que a la varianza de los errores, o error cuadrático medio, se la conoce también como varianza independiente de X, varianza no común con X, varianza no determinada por X o *varianza no explicada por X*. En este sentido, se puede decir que la descomposición de la varianza del criterio consiste en calcular qué parte de la varianza del criterio (S^2_y) depende de la variable predictora ($S^2_{y'}$) y qué parte es independiente de ella ($S^2_{y \cdot x}$).

Ya podemos abordar la cuestión de la valoración de los modelos de regresión. Para ello partiremos de la expresión [7.11], que refleja la relación que existe entre el error cuadrático medio, la varianza del criterio y la correlación de Pearson. Esta fórmula también se puede expresar de la siguiente forma:

$$r^2 = \frac{S^2_{y \cdot x}}{S^2_y} \qquad [7.12]$$

Sustituyendo en esta ecuación $S^2_{y \cdot x}$ por su valor en [7.10] llegamos a la expresión:

$$r^2 = \frac{S^2_{y'}}{S^2_y} \qquad [7.13]$$

Esta última fórmula refleja una idea de gran interés para la interpretación de r en el contexto de la regresión: *el coeficiente de correlación de Pearson elevado al cuadrado indica la proporción de varianza del criterio que queda explicada con el modelo lineal*. Es decir, si por ejemplo en un modelo lineal la varianza del criterio es 20, y al descomponerla encontramos que la varianza explicada es 5 y la no explicada 15 (necesariamente la suma de estas dos últimas tiene que ser igual a la

total, por la fórmula [7.10]), entonces las proporciones de varianza explicada y no explicada son, respectivamente:

$$\frac{5}{20} = 0{,}25 \qquad \text{y} \qquad \frac{15}{20} = 0{,}75$$

Lo que se demuestra con la fórmula [7.13] es que el primer cociente es igual a r^2. Por eso r^2 recibe el nombre específico de *coeficiente de determinación*.

> El coeficiente de determinación en un modelo de regresión simple es igual al coeficiente de correlación de Pearson al cuadrado, r^2, e indica la proporción de varianza del criterio que queda explicada por ese modelo lineal.

Cuando no hay tendencia alguna a la linealidad, entonces $S^2_{y'}/S^2_y = 0$ y $S^2_{y\cdot x}/S^2_y = 1$ (la correlación entre X e Y es igual a cero). Cuando los puntos están perfectamente en línea recta, entonces $S^2_{y'}/S^2_y = 1$ y $S^2_{y\cdot x}/S^2_y = 0$ (la correlación entre X e Y es igual a 1 o −1). La mayor parte de los modelos reales presentan valores intermedios entre estos dos extremos.

Por otra parte, el hecho de que el coeficiente de determinación se obtenga elevando r al cuadrado justifica nuestras afirmaciones del capítulo 5 acerca de que el ajuste de unos puntos a una recta no depende del signo de la correlación. Dos modelos en los que r sea, respectivamente, 0,8 y −0,8 explican la misma proporción de varianza del criterio (0,64).

Todo esto nos indica que la valoración de un modelo lineal no se puede hacer sobre el valor absoluto del promedio de los errores cuadráticos, sino sobre la cantidad que este promedio representa con respecto a la varianza total del criterio. Dado lo expuesto en el apartado 7.3.1 sobre el método de mínimos cuadrados, ante dos modelos en los que el promedio de los errores cuadráticos fueran 8 y 10, respectivamente, tenderíamos a pensar que el primero tiene un mejor ajuste. Sin embargo, no necesariamente es así. Quizá la varianza total en el primero sea 10 y en el segundo 50. En el primer caso la proporción de varianza sin explicar es 8/10 = = 0,80, mientras que en el segundo es 10/50 = 0,20. Es decir, la valoración de las varianzas explicada y no explicada se debe hacer en términos relativos, y para ello el instrumento apropiado es r^2. Podemos imaginarnos esta interpretación de las varianzas de forma gráfica, mediante diagramas de Euler. La varianza de una variable se expresa como un área; el área común entre dos variables sería la varianza explicada al predecir una a partir de la otra, mientras que el área de Y no compartida con X sería la varianza del criterio no explicada por el modelo (figura 7.4).

7.3.3. Aplicación del modelo

Una de las aplicaciones fundamentales de los modelos de regresión es la de hacer predicciones (o conjeturas, dependiendo del contexto) en la variable criterio para los casos nuevos de los que sólo se conoce el valor en la variable predictora.

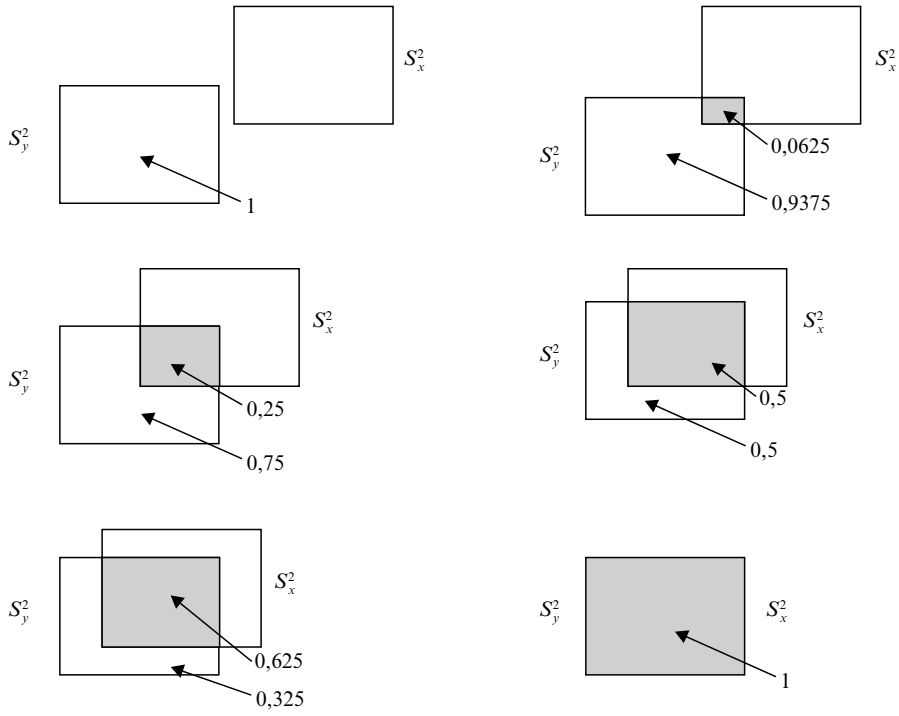


Figura 7.4.—Representación gráfica del coeficiente de determinación en términos de varianza común o varianza de *Y* (variable criterio) explicada por *X* (variable predictora). En cada caso, la zona oscura representa la proporción de la varianza de *Y* que se solapa con la varianza de *X*; es decir, representa el coeficiente de determinación (r^2).

Para el uso de la ecuación de regresión obtenida mediante [7.3] y [7.4] será necesario conocer la puntuación directa del caso para el que queremos hacer una predicción o conjetura (para emplear la ecuación en puntuaciones típicas se necesitará la puntuación típica del nuevo caso). Veamos un ejemplo numérico completo en el que se realizan las tres tareas fundamentales de la regresión, incluyendo la aplicación a la predicción en los dos tipos de puntuación.

Ejemplo. Queremos: *a*) obtener la ecuación de regresión que permite hacer predicciones en *Y* conociendo los valores en *X*. Para ello contamos con 8 pares de valores completos (en las dos primeras columnas de la tabla 7.2). También queremos *b*) valorar la capacidad predictiva del modelo obtenido y *c*) hacer con él la predicción en puntuaciones directas a un individuo con un valor de 4 en *X* y la predicción en puntuaciones típicas a un individuo con un valor de 8 en *X*.

a) *Identificación del modelo.* Obtenemos primero algunos estadísticos y los coeficientes de la ecuación de regresión de *Y* sobre *X* mediante los 8 pares de valores:

TABLA 7.2

X	Y	X^2	Y^2	XY
4	3	16	9	12
6	4	36	16	24
2	1	4	1	2
5	2	25	4	10
7	5	49	25	35
6	3	36	9	18
3	1	9	1	3
3	2	9	4	6
36	21	184	69	110

$$\bar{X} = \frac{36}{8} = 4{,}5$$

$$\bar{Y} = \frac{21}{8} = 2{,}625$$

$$S_x^2 = \frac{184}{8} - 4{,}5^2 = 2{,}75$$

$$S_y^2 = \frac{69}{8} - 2{,}625^2 = 1{,}734$$

$$B_{yx} = \frac{8 \cdot 110 - 36 \cdot 21}{8 \cdot 184 - 36^2} = 0{,}705 \qquad A_{yx} = 2{,}625 - 0{,}705 \cdot 4{,}5 = -0{,}548$$

$$r_{yx} = \frac{8 \cdot 110 - 36 \cdot 21}{\sqrt{8 \cdot 184 - 36^2} \cdot \sqrt{8 \cdot 69 - 21^2}} = 0{,}887$$

Ecuación en puntuaciones directas: $Y'_i = -0{,}548 + 0{,}705 \cdot X_i$.
Ecuación en puntuaciones típicas: $z'_y = 0{,}887 \cdot z_x$

b) *Valoración del modelo.* Para valorar la capacidad predictiva del modelo obtenemos el coeficiente de determinación, elevando al cuadrado la correlación de Pearson. El resultado nos indica que con este modelo se explica el 78,7 por 100 de la varianza del criterio:

$$r_{xy}^2 = 0{,}887^2 = 0{,}787$$

c) *Aplicación del modelo.* Para hacer las predicciones solicitadas utilizamos, respectivamente, las ecuaciones correspondientes. Para la primera predicción sustituimos directamente en la ecuación en directas:

$$Y'_i = -0{,}548 + 0{,}705 \cdot 4 = 2{,}273$$

Para la segunda predicción debemos primero obtener la puntuación típica del individuo en cuestión y luego sustituir en la ecuación de regresión para puntuaciones típicas:

$$z_x = \frac{8 - 4{,}5}{\sqrt{2{,}75}} = 2{,}11$$

$$z'_y = 0{,}887 \cdot 2{,}11 = 1{,}87$$

7.3.4. Algunas consideraciones en torno a la regresión

En este apartado hemos recogido algunas ideas adicionales que servirán de ayuda en la interpretación y la valoración de los modelos de regresión simple.

a) Aunque todos los desarrollos de los apartados anteriores han estado dirigidos a derivar las ecuaciones de regresión de Y sobre X, el uso de unas letras u otras para designar a las variables predictora y criterio es arbitrario. También podíamos haber derivado las ecuaciones de regresión de X sobre Y (aquellas que permitirían predecir X a partir de Y). Éstas serían análogas a las anteriores, pero sustituyendo en las fórmulas X por Y y viceversa. En cualquier caso, se debe tener presente que no necesariamente los coeficientes de las ecuaciones en directas son iguales y, salvo casualidades, $B_{yx} \neq B_{xy}$ y $A_{yx} \neq A_{xy}$. Por el contrario, las pendientes de las ecuaciones en puntuaciones típicas sí serán iguales, puesto que $r_{xy} = r_{yx}$.

b) Una vez construido el modelo de regresión, éste puede servirnos para predecir valores en la variable criterio a partir de la variable predictora. Pero para ello los casos nuevos deben ser homogéneos con aquellos sobre los que se construyó el modelo. Por ejemplo, si construimos una recta de regresión del peso sobre la estatura tomando los pesos y estaturas de cien hombres adultos, luego no debemos utilizar esa recta para predecir el peso de una mujer o de un niño.

c) ¿Cuándo debemos restringir nuestro análisis a la correlación y cuándo debemos hacer un análisis de regresión? A veces lo único que nos interesará será conocer la capacidad predictiva de una variable con respecto a otra. En esos casos basta con hallar el coeficiente de determinación (r^2). Sólo cuando nos interese conocer la tasa de cambio (pendiente), la predicción para $X = 0$ (origen) o la aplicación del modelo a la predicción (o conjetura) de casos nuevos para los que sólo se conoce X, tendrá sentido emprender la tarea de identificar completamente el modelo de regresión. En otras ocasiones nos interesará hacer predicciones, pero sólo si estos pronósticos merecen la suficiente confianza. Un esquema típico de trabajo es el siguiente:

1. Representar gráficamente los puntos para tener una primera visión de conjunto y descartar los modelos lineales cuando el diagrama de dispersión así lo aconseje.

2. Si el modelo lineal parece plausible a partir de la gráfica, obtener el coeficiente de determinación y, a partir de él, decidir si las predicciones hechas con el modelo lineal merecerán la suficiente confianza.
3. En caso de que la respuesta a la pregunta anterior sea negativa se puede pasar a estudiar algún otro tipo de modelo o simplemente descartar a la variable X como predictora potencial de Y. Por el contrario, si la respuesta es positiva se pasa a la identificación del modelo, calculando la pendiente y el origen de la recta de regresión.
4. Se aplica el modelo a los valores de X para los que se quiera hacer predicciones o conjeturas (Y').

d) Aun a riesgo de ser repetitivos, queremos insistir de nuevo en que el descubrimiento de relaciones lineales y su aplicación a la predicción no justifican por sí solas la conclusión de la existencia de relaciones causales entre las variables, tal y como ya hemos discutido en relación con el ejemplo del control de calidad.

e) Al utilizar continuamente los términos «pronóstico» y «predicción» en el contexto de la regresión, algunos recién llegados a la estadística tienden a pensar que ésta se aplica sólo en situaciones en las que se conoce el valor de la variable predictora, pero el de la variable criterio todavía no se ha registrado, y de lo que se trata es de anticiparse al futuro mediante predicciones más o menos precisas. Aunque ésta es una situación bastante realista, de la que, además, hemos utilizado ejemplos, hay otras situaciones en las que no es aplicable. A veces el problema consiste en la imposibilidad material de medir una de las variables, o hay limitaciones de recursos económicos, humanos, etc., que nos impiden medir la variable Y en todos los individuos. La recta de regresión se puede utilizar precisamente como instrumento para decidir a qué individuos se les mide Y, en función del valor que se conjetura. Así, en el ejemplo del control de calidad que describíamos más arriba, una vez establecido el modelo, sólo a aquellos a los que se predice un buen rendimiento se les asignaría ese puesto laboral y, por tanto, sólo de esos individuos se acabaría teniendo información sobre los valores reales de Y. Como consecuencia de esta selección, sería incorrecto recalcular la ecuación incorporando los datos de estos sujetos, dado que éstos habrían sido seleccionados precisamente con base en X y no se pueden considerar como una muestra representativa. Otro ejemplo podría ser el de algunas pruebas diseñadas para detectar precozmente problemas de dislexia. Con estas pruebas, y basándonos en ecuaciones de regresión construidas a partir de grandes bancos de datos, se podría identificar a aquellos niños que tienen una mayor probabilidad de desarrollar el problema. Sin embargo, sobre esos individuos se aplicaría un tratamiento preventivo y, por tanto, su nivel de dislexia posterior no nos serviría para valorar la auténtica capacidad predictiva de nuestra prueba de detección precoz.

f) Las ecuaciones de regresión necesariamente pasan por el punto correspondiente a las medias de ambas variables. Por una parte, la media de las puntuaciones típicas es 0, y ya hemos visto que la ecuación en este tipo de puntuaciones pasa por el punto de cruce de los ejes (0, 0). Con respecto a las puntuaciones

directas, se demuestra fácilmente que la recta pasa también por el punto $(\bar{X}, \bar{Y})$: sustituyendo el valor $\bar{X}$ en la ecuación de regresión de Y sobre X obtenemos el pronóstico asociado a ese valor:

$$Y' = A + B \cdot \bar{X} = (\bar{Y} - B \cdot \bar{X}) + B \cdot \bar{X} = \bar{Y}$$

g) Otro hecho destacable es que las puntuaciones z'_y no son verdaderas puntuaciones típicas, sino predicciones en puntuaciones típicas. La diferencia entre ambos conceptos no es trivial, pues es la que explica que se observen algunas aparentes anomalías. Por ejemplo, aunque estas puntuaciones tengan media cero, en general su varianza no será igual a 1, como ocurre con las verdaderas típicas. Por eso las puntuaciones z'_y reciben el nombre de puntuaciones *pseudotípicas*. En concreto, la varianza de estas puntuaciones es igual a r^2 y, por tanto, sólo cuando ambas variables mantengan una correlación lineal perfecta su varianza será exactamente igual a 1. Es fácil demostrar esto; dado que los pronósticos en puntuaciones típicas son una transformación lineal de las típicas de X (pues $z'_y = r \cdot z_x$), su varianza será igual a la varianza de las z_x multiplicada por el cuadrado de la constante multiplicada:

$$S^2_{z'} = r^2 \cdot S^2_z = r^2 \cdot 1 = r^2$$

h) El término *regresión* surgió en el contexto de los estudios de Francis Galton sobre la herencia. Utilizó primero el término reversión, para luego pasar a utilizar definitivamente el término regresión (Galton, 1885); de hecho, de la inicial de ese término procede el uso de r para representar al coeficiente de correlación. Uno de los fenómenos que más llamaban su atención era que en las poblaciones se observaba una relativa estabilidad en los caracteres cuantitativos. Así, las estaturas parecían mantenerse dentro de un rango bastante estable, un hecho que parecía incompatible con la observación de una variabilidad simple en el paso de una generación a otra. Veámoslo con un ejemplo numérico. Supongamos que las estaturas de los hijos de los hombres de 1,80 sufrieran una variabilidad, oscilando entre 1,70 y 1,90, pero centradas en la media de sus progenitores. Si esto ocurriera con todos los grupos formados con padres de estaturas homogéneas, la variabilidad de las estaturas de la población iría creciendo constantemente. La razón es que aquellos hijos con estaturas de 1,90 tendrían a su vez hijos con estaturas entre 1,80 y 2,00; los nietos que sean hijos de los de 2,00 oscilarían entre 1,90 y 2,10, y así sucesivamente. Es claro que las cosas no suceden así. Galton dibujó las medias que tendrían los hijos de padres con alturas homogéneas en caso de ser cierto todo lo anterior y, simultáneamente, dibujó las medias reales observadas en grupos grandes de pares padre-hijo. Mientras que la recta teórica sería la recta A de la figura 7.5, la recta observada era la B. Galton se dio cuenta de que esta desviación se podía describir como una regresión a la media, es decir, los hijos de padres con una determinada altura tenían una estatura media más cercana a la media poblacional que la de sus padres. Esto es independiente de que esa altura sea superior o inferior a la media. Es decir, si la altura media en la población es de 1,70, entonces los hijos de los padres de 1,80 tendrán una estatura

media que estará entre 1,70 y 1,80, mientras que los hijos de padres con estatura de 1,60 tendrán estaturas con una media que estará entre 1,60 y 1,70. Este fenómeno de regresión a la media dio nombre a los análisis de las relaciones entre dos variables que se practicaron, primero mediante representaciones gráficas de los valores empíricos de ambas variables (diagramas de dispersión) y después mediante la derivación de ecuaciones que cumplieran el criterio de mínimos cuadrados.

Parte del sentido original del término se mantiene, puesto que para cualquier valor de la variable predictora en puntuaciones típicas (z_x), y mientras la correlación no sea perfecta ($r^2 < 1$), el pronóstico (z'_y) será más cercano a la media (0) que el valor de la variable predictora (recuérdese la ecuación en puntuaciones típicas: $z'_y = r_{xy} \cdot z_x$). De ahí la afirmación de Hays (1988, p. 560): «Es una buena apuesta decir que un individuo estará relativamente más cerca de la media en la variable que predecimos que en la conocida». Aunque esto no se verifique en todos los individuos concretos, globalmente será así.

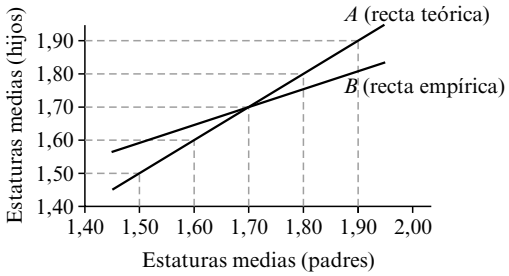


Figura 7.5.—Representación gráfica que dio origen al término regresión (véase texto).

La utilidad de los modelos de regresión se manifiesta en la gran amplitud con que se aplican. Como ejemplo del uso de las ecuaciones de regresión en psicología vamos a señalar algunas aplicaciones prácticas, tanto de regresión simple como de regresión múltiple (regresión con más de una variable predictora; véase el apéndice de este capítulo), que servirán como ilustración de lo expuesto hasta aquí:

a) En sus estudios de búsqueda visual, Treisman y Gelade (1980) midieron el tiempo (T) que tardaban sus observadores en detectar la presencia de un estímulo-objetivo en una matriz de símbolos, en función del número de elementos que componían esas matrices (M). Lógicamente, el tiempo de búsqueda era mayor cuantos más elementos se incluían en la matriz. La ecuación que relacionaba esas variables, así como su bondad de ajuste, fueron las siguientes:

$$T' = 514 + 25{,}1 \cdot M \qquad r^2 = 0{,}966$$

b) Martin (1977) pensó que el rendimiento en una tarea, cuando se ejecuta como tarea secundaria (R_S), y, por tanto, simultáneamente a otra, guar-

daría una relación directa con el rendimiento en esa misma tarea al ser realizada aisladamente (R_a). Encontró de ese modo la siguiente ecuación de regresión:

$$R'_S = 0{,}12 + 0{,}93 \cdot Ra \qquad r^2 = 0{,}846$$

c) García-Jimenez, Alvarado y Jiménez-Blanco (2000) estudiaron la relación entre el rendimiento académico de un grupo de estudiantes de primer curso de psicología en el examen final de una asignatura de metodología *(Rend)* y un conjunto de variables potencialmente predictoras, como son la calificación media obtenida en el bachillerato *(Bach),* el grado de participación en clase *(Par)* y la asistencia a clase durante el curso *(Asis).* La ecuación de regresión múltiple encontrada fue:

$$Rend' = 2{,}522 + 0{,}814 \cdot Bach + 1{,}125 \cdot Par + 0{,}084 \cdot Asis \qquad R^2 = 0{,}419$$

d) Del estudio sobre la satisfacción con la vivienda (*SV*) de Martín-Baró (1985) se desprende la siguiente ecuación de regresión en puntuaciones típicas, basada en el hacinamiento (*H*), el tipo de vivienda (*TV*), la densidad espacial (*DE*) y la ocupación (*O*):

$$z'_{SV} = 0{,}47 \cdot z_H - 0{,}27 \cdot z_{TV} + 0{,}20 \cdot z_{DE} - 0{,}10 \cdot z_O \qquad R^2 = 0{,}502$$

En esta ecuación hay coeficientes positivos y negativos, pues algunas variables mantienen relaciones directas con la variable criterio y otras mantienen relaciones inversas.

PROBLEMAS Y EJERCICIOS

1. Obtenga las ecuaciones de regresión, en puntuaciones directas, de Y sobre X y de X sobre Y, a partir de los siguientes pares de valores:

X	*Y*
0	1
2	3
2	2
3	4
6	7

2. Se ha obtenido la recta que permite pronosticar el rendimiento en la asignatura de Matemáticas, M, a partir de la capacidad de abstracción, CA, en estudiantes de primer año de secundaria. Sabiendo que la recta es: $M' = -10 + 0{,}30 \cdot CA$, obtenga los pronósticos en M a partir de las tres puntuaciones siguientes en CA: 40, 50 y 60.

3. Siguiendo con el ejercicio anterior, una vez finalizado el curso se ha obtenido el rendimiento en la asignatura de Matemáticas de tres alumnos. Teniendo en cuenta los datos de la siguiente tabla, obtenga el error en el pronóstico de cada uno de ellos.

CA	*M*
40	4
52	6
65	9

4. A partir de los pares de valores en las variables X e Y que se muestran en la tabla siguiente, obtenga:

a) S_{xy}, S_x, S_y y r_{xy}.
b) La recta de regresión en puntuaciones directas de Y sobre X.
c) La recta de regresión en puntuaciones típicas de Y sobre X.

X	*Y*
1	6
2	5
4	5
5	3
6	2

5. A partir de los pares de valores mostrados en la tabla que aparece más abajo, obtenga la recta de regresión de Y sobre X en puntuaciones típicas, los pronósticos y la varianza de los pronósticos.

X	Y
2	1
3	6
3	5
4	7
5	8

6. Siguiendo con el ejercicio anterior, obtenga la recta de regresión en puntuaciones directas de Y sobre X y los pronósticos.

7. Se ha calculado la recta de regresión de Y sobre X, obteniéndose un valor de la pendiente mayor que cero. Si la varianza de los errores en los pronósticos es 120 y la varianza de la variable criterio es igual a 160, calcule la proporción de varianza de Y explicada por X y la correlación de Pearson entre X e Y.

8. Se ha realizado una investigación sobre depresión, D, y rendimiento laboral, RL, en una muestra de trabajadores de una empresa. En dicha muestra se obtuvo que la varianza de D fue de 25 y la de RL de 49; además, la recta de regresión que permite pronosticar RL a partir de D fue: $RL' = 5 - 0{,}56 \cdot D$. Obtenga:

a) La proporción de varianza explicada.
b) El error cuadrático medio.

9. Se ha estudiado la relación existente entre la percepción de la complejidad de una tarea, PC, y la ansiedad al resolverla, A. La varianza de PC ha sido igual a 40 y la de A 20, mientras que la correlación entre PC y A fue de 0,3. En el caso de que se utilizara la variable PC como criterio y la variable A como predictora, ¿cuál sería la varianza de los errores que se cometen al realizar las predicciones? ¿Y la varianza de los pronósticos?

10. ¿Qué correlación es preciso que exista entre rendimiento académico, RA, y horas de estudio, HE, para que se pueda afirmar que un 12,25 por 100 de la varianza en rendimiento académico está explicado por las horas de estudio, sabiendo que existe una relación lineal directa entre RA y HE?

11. Calcule y comente la recta de regresión que permite predecir Y a partir de X con los siete pares de puntuaciones siguientes:

X	*Y*
6	8
6	6
6	6
6	4
6	3
6	1
6	0

12. Se ha desarrollado una investigación en la que se ha observado que el 36 por 100 de la variabilidad observada en la variable creatividad, *C*, queda explicada por las puntuaciones obtenidas en un test de personalidad, *A*. Sabiendo que la varianza de *C* ha sido igual a 20 y la de *A* igual a 18, obtenga:

a) La varianza explicada.
b) La varianza no explicada.
c) La correlación de Pearson entre *C* y *A*.

13. Decida razonadamente, a partir de la matriz de correlaciones expuesta a continuación, cuál es la mejor predictora de la variable *W*. Con la información que se dispone, obtenga la correspondiente recta de regresión y la proporción de varianza no explicada.

$$\mathbf{R} = \begin{array}{c|cccc} & U & V & W & X & Y \\ U & & 0{,}30 & 0{,}45 & 0{,}70 & 0{,}10 \\ V & & & -0{,}90 & 0{,}08 & 0{,}12 \\ W & & & & 0{,}65 & 0{,}80 \\ X & & & & & 0{,}13 \\ Y & & & & & \end{array}$$

14. Se ha realizado una investigación sobre ansiedad laboral, *AL*, en un grupo de controladores aéreos. Se encontró que un 64 por 100 de la varianza de *AL* queda sin explicar por el grado percibido de responsabilidad en el puesto de trabajo, *GR*; además, se observó que un 25 por 100 de la varianza de *AL* no queda explicada por las puntuaciones obtenidas en el STAI, *ST*. Sabiendo que la varianza de *AL* es 60, obtenga:

a) La descomposición de la varianza de *AL* si se utiliza como variable predictora la variable *GR*.
b) La descomposición de la varianza de *AL* si se utiliza como predictora la variable *ST*.
c) Decida razonadamente cuál de las dos variables, *GR* o *ST*, predice mejor la ansiedad.

15. Se está utilizando X como variable predictora, tanto para predecir a la variable Y como a la variable U. Al construir la recta de Y sobre X se obtiene una varianza explicada igual a 100, siendo la varianza de Y igual a 200. Además, con la recta de U sobre X se obtiene una varianza explicada igual a 60, siendo la varianza de U igual a 80. ¿En cuál de las dos rectas es mejor predictora la variable X?

16. Con la tabla de datos que aparece más abajo, obtenga:

a) La matriz de varianzas-covarianzas, así como el vector de medias.
b) La matriz de correlaciones.
c) La recta de regresión en puntuaciones directas de X sobre Y, así como la de X sobre U.
d) El diagrama de dispersión y la recta de regresión de X sobre Y.
e) ¿Cuál de las dos rectas obtenidas en el apartado *c*) predice mejor a la variable X?
f) Descomponga la varianza criterio de las dos modelos obtenidos en el apartado *c*).

X	*Y*	*U*
0	8	1
1	7	4
1	6	3
2	3	2
3	2	3

17. Mediante los datos obtenidos en un experimento, se ha podido construir la recta que permite pronosticar el tiempo de respuesta, *TR*, a partir del grado de complejidad de una tarea auditiva, *GC*. Se ha observado que la varianza explicada es igual a 180. Sabiendo que la correlación entre *TR* y *GC* es igual a 0,60, obtener:

a) La proporción de varianza explicada.
b) La proporción de varianza no explicada.
c) La descomposición de la varianza del criterio.

18. Diga en qué casos se cumple que:

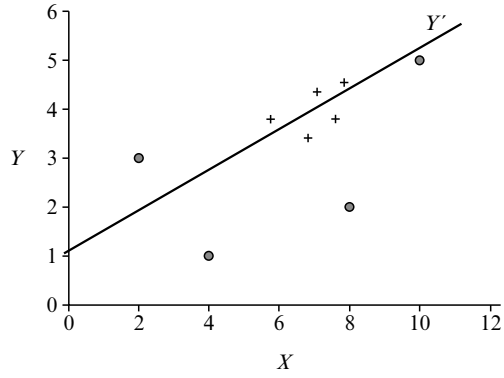
a) La varianza explicada es igual a la varianza del criterio.
b) La varianza no explicada es igual a la varianza del criterio.

19. A partir de los datos de una muestra de cien sujetos, hemos encontrado que la recta de regresión que permite predecir la *creatividad* (*C*) a partir de la *inteligencia* (*I*) es: $C' = 30 + 1{,}3 \cdot I$; y a partir de la *extraversión* (*E*) es: $C' = 20 - 0{,}36 \cdot E$. Sabiendo, además, que las varianzas de las variables son: $S_I^2 = 20$, $S_C^2 = 60$ y $S_E^2 = 30$, responda a las siguientes cuestiones:

a) Teniendo en cuenta que Julia ha obtenido unas puntuaciones típicas de 0,4 en *C*, 0,15 en *I* y −0,8 en *E*, si hacemos el pronóstico en *creatividad* en puntuaciones típicas para Julia utilizando por un lado la *inteligencia* y por otro la *extraversión*, ¿con cuál de esas variables predictoras se comete un mayor error en el pronóstico?

b) De un sujeto elegido al azar, ¿con cuál de esos modelos lineales esperaríamos a priori tener un mayor error en su pronóstico?

20. En la figura inferior aparece la representación gráfica del diagrama de dispersión y la recta de regresión de *Y* sobre *X* de una muestra formada por varones (+) y mujeres (•):



A partir del gráfico anterior, y sin hacer ningún cálculo, responda a las siguientes cuestiones:

a) ¿De cuántos varones y mujeres se compone la muestra?

b) ¿Cuántos varones recibirían con ese modelo pronósticos superiores a sus puntuaciones empíricas en *Y*?

c) ¿Cuántas mujeres recibirían con ese modelo pronósticos inferiores a sus puntuaciones empíricas en *Y*?

d) ¿Qué pronóstico se haría a un sujeto con un valor 0 en *X*?

21. Deseamos estudiar el tipo de relación que existe entre el número de alternativas (*NA*) y el tiempo de respuesta (*TR*) en una tarea de tiempos de reacción de elección. Pasamos a un sujeto varios ensayos con distinto número de alternativas y medimos el *TR* en décimas de segundo, obteniendo los resultados que aparecen en la tabla siguiente.

NA	2	3	4	1	5	3	8	6	4	5
TR	2,5	3	2,5	2	4	2,5	6	5	3,5	3,5

a) Obtenga la recta de regresión que permite predecir el *TR* a partir del número de alternativas y comente la confianza que pueda merecer ese modelo lineal.
b) Descomponga la varianza del criterio del modelo lineal obtenido.
c) ¿Qué *TR* predecimos que se obtendrá con nueve alternativas?
d) Elabore la representación gráfica de la recta de regresión superpuesta en el diagrama de dispersión.

22. En la tabla siguiente se presentan los valores medidos en un test de *inteligencia* (*X*) y un test de *razonamiento abstracto* (*Y*) en una muestra de cuatro sujetos.

X	2	1	5	4
Y	1	3	6	3

Si transformamos las puntuaciones de *X* en las puntuaciones de la escala derivada de *CI* (donde la media es 100 y la desviación típica 15):

a) ¿Qué cociente intelectual le correspondería al tercer sujeto de la muestra?
b) ¿Qué puntuación de *CI* le pronosticaríamos a partir de su valor en *Y*? ¿Cuál sería el error cometido en el pronóstico?
c) Descomponga la varianza del criterio del modelo de regresión de *X* sobre *Y*.

23. Hemos medido a diez sujetos en las variables *capacidad de concentración* (*C*) y *aprovechamiento académico* (*A*).

C	5	7	7	5	6	8	8	12	3	9
A	3	3	2	0	3	4	4	6	0	5

Responda a las siguientes cuestiones:

a) ¿Podría decirse que hay relación lineal entre las dos variables?
b) ¿Utilizaría la variable *C* para predecir el *aprovechamiento académico*?
c) ¿Qué nivel de *aprovechamiento académico* pronosticaría a un sujeto con una puntuación de 6 en *C*?
d) Elabore la representación gráfica de la recta de regresión superpuesta sobre el diagrama de dispersión.

24. Supongamos que a cada sujeto del ejercicio anterior le asignamos como puntuación la diferencia entre su puntuación en *capacidad de concentración* (*C*)

y su puntuación en *aprovechamiento académico* (*A*). ¿Cuánto valdrán la media y la varianza de estas puntuaciones?

25. Treisman y Gelade (1980) estudiaron el tiempo de respuesta (*TR*) medio que se tarda en reaccionar ante una matriz de estímulos en función del número de elementos de la matriz (*N*). Trabajaron con matrices de 1, 5, 15 y 30 elementos. Tras calcular el *TR* medio invertido en las respuestas para cada uno de los tamaños, clasificaron los ensayos en dos tipos: *ensayos positivos,* si la matriz contenía en elemento que había que detectar, y *ensayos negativos* si no lo contenía. Obtuvieron las ecuaciones de regresión siguientes:

$$\text{Para ensayos positivos:} \quad TR' = 448 + 3{,}1 \cdot N$$

$$\text{Para ensayos negativos:} \quad TR' = 514 + 25{,}1 \cdot N$$

En el primer caso quedaba explicado el 67,9 por 100 de la varianza del criterio, y en el segundo el 96,6 por 100. Según su teoría de la atención basada en la integración de caracteres, el ajuste al modelo lineal debería ser bueno para los ensayos negativos, pero no para los positivos. A continuación, conteste a las siguientes cuestiones:

a) Si ante una matriz de diez elementos que no contiene el objetivo se tarda 753 milisegundos, ¿en cuánto nos habríamos equivocado en caso de predecir ese tiempo mediante el modelo lineal encontrado?
b) ¿Cuánto es el tiempo medio que se predice que se incrementará el *TR* en ensayos positivos por cada tres elementos añadidos a la matriz?
c) ¿Qué podemos decir de la teoría de Treisman y Gelade?

26. Unos investigadores del hospital universitario Ramón y Cajal de Madrid publicaron en 2007 un estudio sobre la relación entre el consumo de sal diario (*CS*) y la tensión arterial (*TA*). Más concretamente, administraron a una serie de voluntarios distintas dosis de sal en su dieta (medida en gramos al día) y midieron su tensión arterial (en milímetros de mercurio o mmHg) un tiempo después, y encontraron que el modelo lineal de mejor ajuste en esos datos era:

$$TA' = 86{,}371 + 6{,}335 \cdot CS \quad \text{(en puntuaciones directas)}$$

$$z_{TA'} = 0{,}967 \cdot z_{CS} \quad \text{(en puntuaciones típicas)}$$

A partir de los datos anteriores, responda a las siguientes cuestiones:

a) ¿Qué tensión arterial pronosticaríamos a un sujeto que consuma una dosis de 1,8 gramos de sal en su dieta diaria?
b) ¿Qué confianza dan los pronósticos ofrecidos con ese modelo lineal?
c) Sabiendo que la varianza de *TA* es 55,22, descomponga la varianza del criterio.

27. Supongamos que hemos encontrado que la recta de regresión de Y sobre X siguiendo el criterio de mínimos cuadrados es: $Y' = 10 + 3{,}5 \cdot X$. Obtenga la ecuación de regresión si hacemos las siguientes transformaciones:

a) Multiplicamos los valores de Y por 2.
b) Sumamos 5 a los valores de X.
c) Sumamos 3 a los valores de Y y multiplicamos por 2 a los de X.

28. Basándonos en el enunciado del ejercicio 6 del capítulo 5 (estudio sobre la relación entre la motivación de logro con diferentes facetas de la satisfacción laboral), conteste a las siguientes cuestiones:

a) Se desea hacer pronósticos sobre *MC*. ¿Qué variable seleccionaría como la mejor predictora?
b) ¿Cuál es el porcentaje de varianza explicada en la regresión de *SH* sobre *SR*?
c) ¿Qué bondad ofrecería el modelo de regresión de *OP* sobre *MC*?
d) Sabiendo que la varianza de *SR* es 36,5, ¿cuál es la varianza explicada en la regresión de *SR* sobre *MC*?

29. Basándonos en el enunciado del ejercicio 7 del capítulo 5 (estudio sobre la relación entre *X*: *rendimiento académico* e *Y*: *tiempo dedicado al ocio*), y considerando sólo la muestra de estudiantes de Valencia, conteste a las siguientes cuestiones:

a) Obtenga la ecuación de regresión que permite pronosticar el *rendimiento académico* a partir del *tiempo dedicado al ocio.*
b) ¿Cuál es el porcentaje de varianza explicada por el modelo lineal obtenido en el apartado anterior?
c) Descomponga la varianza del criterio.

30. Los pronósticos hechos a los sujetos *FT* y *MC* con un modelo de regresión lineal han sido 7,8 y 6,5, siendo sus puntuaciones en la variable predictora 45 y 53, respectivamente. A partir de los datos anteriores, diga cuál ha sido el modelo lineal utilizado para hacer esos pronósticos.

31. ¿Qué posible rango de valores podría tener la puntuación en *capacidad de concentración* de un sujeto al que se predijo un *aprovechamiento académico* superior a la media en el ejercicio 23 del presente capítulo?

32. Disponemos de las puntuaciones obtenidas por ocho sujetos en las variables *inteligencia emocional* (*I*) y *creatividad* (*C*) en la tabla siguiente.

I	4	5	5	3	6	4	5	7
C	4	5	7	4	6	4	6	8

Un psicólogo ha utilizado un modelo lineal entre estas variables para predecir *creatividad* a partir de la *inteligencia emocional*. De hecho, a un sujeto le ha pronosticado en *creatividad* una puntuación directa de 6,5. Según lo anterior, responda a las siguientes cuestiones:

a) Calcule la recta de regresión en directas utilizada para el pronóstico mencionado, y represente gráficamente el diagrama de dispersión con la recta de regresión superpuesta.

b) ¿Qué puntuación diferencial tiene el mismo sujeto en *inteligencia emocional*?

33. Para hacer un estudio sobre la relación entre los trastornos de la personalidad y la satisfacción general del individuo, administramos a una muestra de 300 sujetos que se encuentran en situación de paro una batería de tests de *agorafobia* (*A*), *hipocondría* (*H*), *depresión* (*D*), *esquizofrenia* (*E*) y *madurez emocional* (*M*). También evaluamos la *satisfacción vital general* (*V*) autoinformada por esas personas. A continuación, se presentan parte de las matrices de varianzas y covarianzas y de correlaciones entre esas variables, así como el vector con las medias:

S =

	A	*H*	*D*	*E*	*M*	*V*
A	36	19,5	29,4	−7,2	−35,1	−24,3
H		25	17,5	( )	−36	−24,75
D			49	2,10	−12,16	−40,95
E				36	−5,40	−5,40
M					81	( )
V						81

Medias	50	60	30	50	40	70

R =

	A	*H*	*D*	*E*	*M*	*V*
A		0,65	( )	−0,20	−0,65	−0,45
H			0,50	−0,30	−0,80	( )
D				0,05	−0,20	−0,65
E					−0,10	−0,10
M						0,60
V						

Responda a las siguientes cuestiones:

a) Complete los valores que faltan en las matrices.
b) El profesor Martínez mantiene que, de entre estas variables, la mejor predictora de la *satisfacción vital* es la *depresión*. ¿Tiene razón el profesor Martínez?
c) Uno de los participantes obtiene las siguientes puntuaciones directas: $A = 40$, $H = 63$, $D = 40$, $E = 38$, $M = 30$ y $V = 78$. Si hubiéramos realizado un pronóstico en V para este participante basándonos únicamente en la *hipocondría*, ¿en cuánto nos hubiéramos equivocado?
d) Tome el modelo lineal propuesto por el profesor Martínez en el apartado *b*) y descomponga la varianza del criterio.
e) Decidimos trabajar con la variable *neuroticismo* (*N*), cuya puntuación se obtiene sumando sus puntuaciones en *agorafobia*, *hipocondría* y *depresión*. Para la muestra de este problema, ¿cuánto valdrán la media y la varianza en *neuroticismo*?

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. $Y'_i = 0{,}719 + 1{,}031 \cdot X_i$; $\quad X'_i = -0{,}575 + 0{,}934 \cdot Y_i$

2.

CA	*M*
40	2
50	5
60	8

3. Llamando *E* al error:

CA	*E*
40	2
52	0,4
65	−0,5

4. *a*) $S_{xy} = -2{,}52$; $\quad S_x = 1{,}855$; $\quad S_y = 1{,}470$; $\quad r_{xy} = -0{,}924$.
b) $Y'_i = 6{,}835 - 0{,}732 \cdot X_i$.
c) $z'_y = -0{,}924 \cdot z_x$.

(Obsérvese que, para un mismo conjunto de pares de valores, el signo de la covarianza, el de la correlación de Pearson y el de la pendiente de la recta siempre coinciden).

5. Recta de regresión en puntuaciones típicas: $z'_y = 0{,}909 \cdot z_x$.
Pronósticos en puntuaciones típicas:

X	z_x	z'_y
2	−1,373	−1,248
3	−0,392	−0,356
3	−0,392	−0,356
4	0,588	0,534
5	1,569	1,426

Varianza de los pronósticos, $S^2_{z'} = 0{,}826$ (observe que es menor que 1; hay que recordar que los pronósticos en puntuaciones típicas no son autenticas puntuaciones típicas sino *pseudotípicas,* ya que estas últimas siempre tienen una varianza menor o igual a 1) (además, se puede comprobar que $S^2_{z'} = r^2_{xy}$).

6. Recta de regresión en puntuaciones directas: $Y'_i = -1{,}924 + 2{,}154 \cdot X_i$.
Pronósticos:

X	Y'
2	2,384
3	4,538
3	4,538
4	6,692
5	8,846

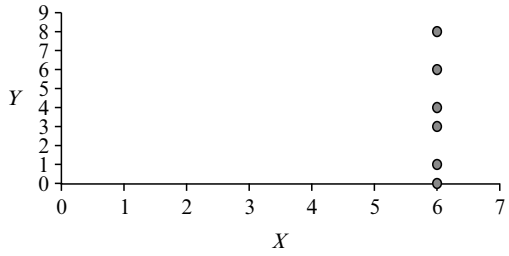
7. La proporción de varianza explicada es: $\frac{S^2_{y'}}{S^2_y} = 0{,}25$. La correlación es $r_{xy} = 0{,}5$.

8. *a*) Proporción de varianza explicada, $r^2_{DRL} = 0{,}16$.
b) Error cuadrático medio, $S^2_{RL \cdot D} = 41{,}16$.

9. Varianza de los errores: $S^2_{PC.A} = 36{,}4$. Varianza de los pronósticos: $S^2_{PC'} = 3{,}6$.

10. $r_{RAHE} = 0{,}35$.

11. Al ser $S_x = 0$, no se puede calcular ni la pendiente B_{yx} ni la correlación. La gráfica que representa a estos datos sería una recta que es paralela al eje Y.



12. *a*) $S^2_{C'} = 7{,}2$.
b) $S^2_{C.A} = 12{,}8$.
c) $r_{CA} = \pm 0{,}60$.

13. La mejor es *V*, porque $r^2_{wv} = 0{,}81$ (mayor que las demás). Recta: $z'_w = -0{,}90 \cdot z_v$. Proporción de varianza no explicada: 0,19.

14. *a*) 60 = 21,6 + 38,4.
b) 60 = 45 + 15.
c) El mejor predictor de *AL* es la variable *ST*.

15. La variable *X* predice mejor a la variable *U* que a la *Y*, a la vista de las proporciones de varianza explicadas.

16. *a*)

		X	*Y*	*U*
S =	*X*	1,04	−2,28	0,36
	Y		5,36	−0,32
	U			1,04

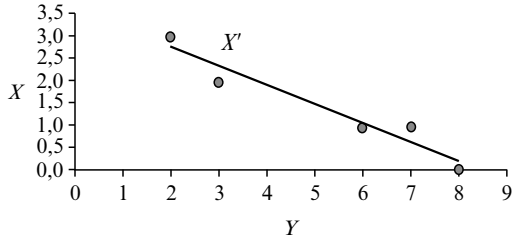
Medias	1,4	5,2	2,6

b)

		X	*Y*	*U*
R =	*X*		−0,966	0,346
	Y			−0,136
	U			

c) Recta de *X* sobre *Y*: $X'_i = 3{,}615 - 0{,}426 \cdot Y_i$; Recta de *X* sobre *U*: $X'_i = 0{,}5 + 0{,}346 \cdot U_i$.

d)



e) La variable Y es mejor predictora, ya que $r^2_{xy} > r^2_{xu}$.

f) Recta de X sobre Y: 1,04 = 0,970 + 0,07; recta de X sobre U: 1,04 = = 0,125 + 0,915.

17. *a*) Proporción de varianza explicada, $r^2_{TRGC} = 0{,}36$.

b) Proporción de varianza no explicada, 0,64.

c) 500 = 180 + 320.

18. Teniendo en cuenta la fórmula [7.10]:

a) Cuando $r^2_{xy} = 1$.

b) Cuando $r^2_{xy} = 0$.

19. *a*) Se consigue un pronóstico más ajustado con la *extraversión* (menor error en el pronóstico).

b) Esperaríamos un mayor error con el segundo modelo (regresión de *creatividad* sobre *extraversión*), puesto que se explica un menor porcentaje de la varianza del criterio (6,25 por 100 frente al 56,25 por 100).

20. *a*) Se compone de cinco varones y cuatro mujeres.

b) Dos varones.

c) Una mujer.

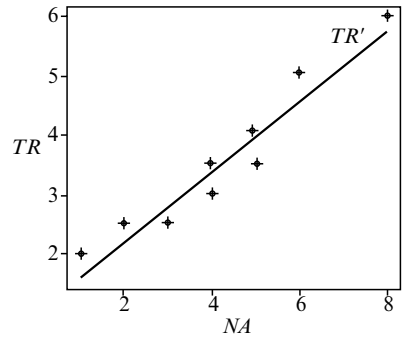
d) Con un valor 0 en X se pronosticaría un valor 1 en Y.

21. *a*) $TR'_i = 1{,}06 + 0{,}58 \cdot NA_i$ y los pronósticos probablemente serán bastante fiables, puesto que con este modelo se explica el 88,5 por 100 de la varianza del criterio.

b) 1,4225 = 1,3110 + 0,1115.

c) 6,28.

d)



22. *a*) Teniendo en cuenta la definición de *CI* del apartado 4.4, el valor es 119,05.

b) La regresión de *CI* sobre *Y* se calcula sobre los siguientes datos:

CI	90,55	80,95	119,05	109,45
Y	1	3	6	3

El pronóstico es 116,44. El error cometido en el pronóstico sería 2,61; en este caso el modelo infraestimaría la puntuación en *CI* del sujeto.

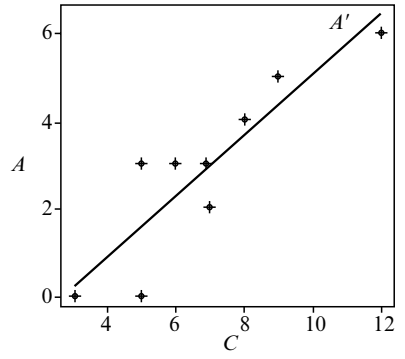
c) 2,5 = 1,255 + 1,245.

23. *a*) Sí, puesto que $r_{CA} = 0{,}894$ (véase también el gráfico del apartado *d*).

b) Sí, porque explica el 79,92 por 100 de la varianza del criterio.

c) $A'_i = -1{,}875 + 0{,}696 \cdot C_i$. Por tanto, a un sujeto con $C = 6$ se le pronosticaría una puntuación 2,301 en *aprovechamiento académico*.

d)



24. Definiendo *T* como $T_i = C_i - A_i$, la media de *T* es 4 y la varianza 1,2.

25. *a*) Nos equivocaríamos en 12 milisegundos.
b) Se incrementará en 9,3 milisegundos.
c) Los datos parecen ir en la línea de lo esperado. El ajuste en los ensayos negativos es muy bueno, aunque hay que tener cautela porque el modelo se basa sólo en 4 puntos. El modelo no es malo en los ensayos positivos, aunque esto podría deberse a una fluctuación aleatoria, por basarse sólo en 4 puntos.

26. *a*) El pronóstico será de 97,77 mmHg.
b) La confianza es alta, pues el modelo explica un 93,51 por 100 de la varianza del criterio.
c) 55,22 = 51,64 + 3,58.

27. *a*) $Y'_i = 20 + 7\ X_i$.
b) $Y'_i = -7{,}5 + 3{,}5\ X_i$.
c) $Y'_i = 13 + 1{,}75 \cdot X_i$.

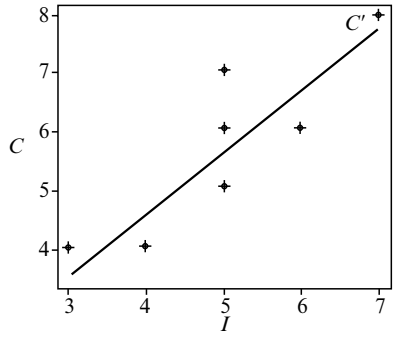
28. *a*) Seleccionaría *SR,* pues es la que explica mayor proporción de varianza de *MC* (0,5625).
b) 12,25 por 100.
c) El 20,25 por 100 de la varianza del criterio.
d) $S^2_{SR'} = 20{,}53$.

29. *a*) $X'_i = 15{,}89 - 0{,}405 \cdot Y_i$.
b) El 67,73 por 100.
c) 5,20 = 3,52 + 1,68.

30. El modelo lineal utilizado ha sido: $Y'_i = 15{,}1125 - 0{,}1625 \cdot X_i$.

31. Podemos afirmar que la puntuación de ese sujeto en *capacidad de concentración* (*X*) es superior a la media.

32. *a*) La ecuación de regresión es: $C'_i = 0{,}345 + 1{,}057 \cdot I_i$ y el gráfico:



b) Si la puntuación directa era $C = 6{,}5$, su puntuación diferencial es: $c = 6{,}5 - 5{,}5 = 1$.

33. *a*) Los valores que faltan son $S_{HE} = -9$, $S_{MV} = 48{,}6$, $r_{AD} = 0{,}70$ y $r_{HV} = -0{,}55$.

b) Sí, porque es la que más varianza de V explica (el 42,25 por 100).

c) El error hubiera sido de $V - V' = 10{,}97$.

d) $81 = 34{,}22 + 46{,}78$.

e) Si $N_i = A_i + H_i + D_i$, entonces la media de N es 140 y la varianza 242,8.

APÉNDICE

Algunas demostraciones

Realizamos las siguientes demostraciones con puntuaciones diferenciales, dado que de esta forma se simplifican mucho los desarrollos. En concreto, vamos a mostrar que, al trabajar con ese tipo de puntuaciones, la pendiente es la misma que en puntuaciones directas, mientras que el origen es necesariamente cero. Es fácil ver que la pendiente no cambia, puesto que en la fórmula [7.5] vemos que los elementos que conducen a la obtención de la pendiente no cambian al trabajar con diferenciales; en concreto, trabajar en diferenciales significa restar a los valores una constante (la media, $\overline{X}$), y hemos visto en capítulos anteriores que al restar una constante ni la correlación ni la varianza se alteran. Lo segundo también es fácil de comprobar, puesto que si en la fórmula del origen, [7.4], sustituimos las medias por su valor (que en diferenciales es siempre cero), comprobamos que el origen también será siempre cero:

$$A_{yx} = \overline{y} - B_{yx} \cdot \overline{x} = 0 - B_{yx} \cdot 0 = 0$$

En resumen, al trabajar con puntuaciones diferenciales la ecuación de regresión queda:

$$y'_i = B_{yx} \cdot x_i$$

Demostración 1: *los pronósticos y los errores en los pronósticos son linealmente independientes*. Para ello basta con demostrar que su covarianza es nula, puesto que si la covarianza es nula también lo será la correlación. Lo hacemos en puntuaciones diferenciales:

$$S_{y'(y-y')} = \frac{\sum y'(y - y')}{N} = \frac{\sum B \cdot x \cdot (y - B \cdot x)}{N} = B \cdot \frac{\sum xy}{N} - B^2 \cdot \frac{\sum x^2}{N} =$$

$$= B \cdot r \cdot S_y \cdot S_x - B^2 \cdot S_x^2 = r \cdot \frac{S_y}{S_x} \cdot r \cdot S_y \cdot S_x - r^2 \cdot \frac{S_y^2}{S_x^2} \cdot S_x^2 = r^2 \cdot S_y^2 - r^2 \cdot S_y^2 = 0$$

Demostración 2: *la correlación entre las puntuaciones en la variable predictora y los pronósticos es perfecta y del mismo signo que la pendiente de la ecuación*. De nuevo trabajaremos con puntuaciones diferenciales:

$$r_{xy'} = \frac{\sum xy'}{N \cdot S_x \cdot S_{y'}} = \frac{\sum x \cdot (B \cdot x)}{N \cdot S_x \cdot (|B| \cdot S_x)} = \frac{B}{|B| \cdot S_x^2} \cdot \frac{\sum x^2}{N} = \frac{B}{|B|} \cdot \frac{S_x^2}{S_x^2}$$

Por tanto, $r_{xy'}$ tiene que ser ± 1.

Demostración 3: *las puntuaciones en la variable predictora y los errores en los pronósticos son linealmente independientes*. Para comprobarlo, basta demostrar

(en diferenciales) que la covarianza es nula, puesto que si la covarianza lo es, también lo será su correlación:

$$S_{x(y-y')} = \frac{\sum x(y - y')}{N} = \frac{\sum x \cdot (y - B \cdot x)}{N} = \frac{\sum xy}{N} - B \cdot \frac{\sum x^2}{N} =$$

$$= r \cdot S_y \cdot S_x - B \cdot S_x^2 = r \cdot S_x \cdot S_y - r \cdot \frac{S_y}{S_x} \cdot S_x^2 = r \cdot S_x \cdot S_y - r \cdot S_x \cdot S_y = 0$$

Demostración 4: *relación entre la varianza de los errores y la correlación de Pearson* (en diferenciales):

$$S_{y \cdot x}^2 = \frac{\sum (y - y')^2}{N} = \frac{\sum (y - B \cdot x)^2}{N} = \frac{\sum (y^2 + B^2 \cdot x^2 - 2 \cdot y \cdot B \cdot x)}{N} =$$

$$= \frac{\sum y^2}{N} + B^2 \cdot \frac{\sum x^2}{N} - 2 \cdot B \cdot \frac{\sum xy}{N}$$

El primer quebrado de esta expresión no es más que la varianza de Y; el segundo es la pendiente elevada al cuadrado, multiplicada por la varianza de X, mientras que en el tercero aparece de nuevo la pendiente multiplicada por la covarianza entre X e Y. Sustituimos la pendiente según la fórmula [7.5] y la covarianza según la fórmula [5.5]:

$$S_{y \cdot x}^2 = S_y^2 + \left(r \cdot \frac{S_y}{S_x}\right)^2 \cdot S_x^2 - 2 \cdot \left(r \cdot \frac{S_y}{S_x}\right) \cdot (r \cdot S_x \cdot S_y) =$$

$$= S_y^2 + r^2 \cdot S_y^2 - 2 \cdot r^2 \cdot S_y^2 = S_y^2 - r^2 \cdot S_y^2$$

Por tanto:

$$S_{y \cdot x}^2 = S_y^2 \cdot (1 - r^2)$$

Regresión múltiple

Es evidente que, en general, si en lugar de hacer los pronósticos mediante una variable predictora utilizamos varias variables predictoras, la capacidad predictiva del modelo se incrementará. Si hay terceras variables que explican parte de la varianza del criterio que la primera variable predictora deja sin explicar, entonces al incluir esas variables en el modelo mejorará la bondad de ajuste y la capacidad predictiva. Cuando se emplea más de una variable predictora en un modelo lineal se está utilizando un modelo de *regresión múltiple*. Si representamos por Y a la variable que queremos predecir o explicar (variable criterio) y por U, V y X a tres variables que sospechamos que mantienen una relación lineal con ella y en las que podemos fundamentar las predicciones (variables predictoras), la forma de la ecuación de regresión múltiple será:

$$Y' = A + B_u \cdot U + B_v \cdot V + B_x \cdot X$$

De nuevo A representa el origen, mientras que B_u, B_v y B_x representan las pendientes asociadas a cada variable predictora. Cuando el número de variables predictoras se limita a dos, se puede hacer una representación en un espacio tridimensional; en ese caso la ecuación de regresión ya no representa una línea, sino un plano de regresión, que tiene una pendiente asociada a cada uno de los dos ejes de las variables predictoras. Con más de dos predictoras no es posible hacer una extensión de la representación en forma de diagrama de dispersión.

No es este el lugar apropiado para exponer en detalle las fórmulas y procedimientos de la regresión múltiple, que con frecuencia exceden el nivel de complejidad matemática que queremos mantener a lo largo de este libro. Sin embargo, vamos a exponer brevemente algunas ideas, siempre en analogía con la regresión simple y como extensión de esta última. Remitimos a los lectores interesados en profundizar sus conocimientos de regresión a otros textos más específicos (Ato y Vallejo, 2007; Cohen y Cohen, 1983; Etxeberría, 1999; Hays, 1988; Peña, 1987; Tatsuoka, 1976).

En la regresión múltiple se deben distinguir también las tareas de identificación, valoración y aplicación del modelo, pero a ellas hay que añadir la de adoptar decisiones respecto a cuántas y cuáles serán las variables predictoras introducidas en el modelo. Será frecuente que el investigador cuente con un conjunto de variables potencialmente explicativas de la varianza de la variable criterio. Sin embargo, deberá acabar seleccionando una parte de ellas para su modelo final. Es precisamente esta parte de la tarea la que no podemos exponer sin haber expuesto previamente los conceptos fundamentales de la estadística inferencial. Por ello hacemos una exposición breve, concisa y no técnica de la regresión múltiple.

Las fórmulas que permiten la *identificación* de los coeficientes de la ecuación de regresión múltiple son más complejas que las de la ecuación de regresión simple, pero se llega a ellas aplicando también el criterio de mínimos cuadrados: serán aquellos que minimizan el promedio de los errores (al cuadrado) cometidos al predecir mediante la ecuación.

La *valoración* del modelo de regresión se hace también mediante la proporción de varianza del criterio que queda explicada por aquél. Aquí se trata del cociente entre la varianza de los pronósticos hechos mediante la ecuación de regresión múltiple y la varianza del criterio. La nomenclatura, aplicada al ejemplo anterior, es la siguiente: $S^2_{y\cdot uvx}$ representa la varianza de los errores cometidos al predecir Y a partir de U, V y X mediante la ecuación de regresión múltiple, mientras que $S^2_{y'}$ representa la varianza de los pronósticos hechos mediante la ecuación de regresión múltiple. La fórmula será, por tanto:

$$S^2_y = S^2_{y'} + S^2_{y\cdot uvx}$$

Por su parte, el coeficiente de determinación se representa por $R^2_{y\cdot uvx}$. Se puede expresar como el cociente entre las varianzas explicada y total, o como el complementario del cociente entre la no explicada y la total:

$$R^2_{y\cdot uvx} = \frac{S^2_{y'}}{S^2_y} \qquad R^2_{y\cdot uvx} = 1 - \frac{S^2_{y\cdot uvx}}{S^2_y}$$

Para la *aplicación* de un modelo de regresión múltiple a la predicción basta con sustituir los valores de las variables predictoras en las ecuaciones correspondientes. Naturalmente, también se puede establecer una ecuación de regresión múltiple en puntuaciones típicas (de hecho, en regresión múltiple se hará incluso con más frecuencia que en regresión simple).

Aunque ya hemos advertido que en el caso de la regresión múltiple es más difícil aportar ideas para el trabajo práctico sin recurrir a la estadística inferencial, no queremos dejar de exponer algunas, aun sin demostración.

La primera idea importante es que la inclusión de nuevas predictoras a un modelo puede incrementar o no la proporción de varianza explicada, pero nunca reducirla. Es decir, si una nueva variable predictora no tiene relación alguna con el criterio, o correlaciona perfectamente con alguna de las variables predictoras ya incluidas (un problema que en este contexto se llama *colinealidad*), entonces el modelo no mejorará en su capacidad predictiva, pero la varianza explicada será como mínimo la que ya existía antes de incluir esa variable. Esto se entenderá mejor si de nuevo representamos gráficamente las varianzas como superficies, e interpretamos los solapamientos entre éstas como aquella parte de la varianza que es común entre ellas (figura 7A.1). Podemos expresar esta idea de la siguiente forma:

$$R^2_{y\cdot u} \leq R^2_{y\cdot uv} \leq R^2_{y\cdot uvx} \leq R^2_{y\cdot uvx\ldots}$$

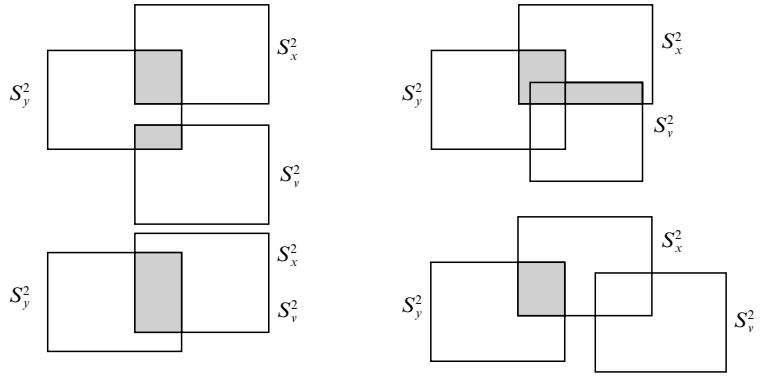


Figura 7A.1.—Representación gráfica del coeficiente de determinación en términos de varianza común o varianza de Y (variable criterio) explicada por X y V (variables predictoras). En cada caso, la zona oscura representa la proporción de la varianza de Y que se solapa con las variables predictoras; es decir, representa el coeficiente de determinación (R^2), pudiendo apreciar el efecto que tiene añadir a X la variable V como predictora de Y. En general, la varianza explicada de Y se incrementará si hay alguna parte de S^2_y que es común con S^2_x y esa parte común no es a la vez común con S^2_x, pues en ese caso ya quedaría explicada en el modelo de regresión simple. Así, en el ejemplo superior izquierdo toda la variación común entre Y y V pasa a engrosar la varianza explicada de Y. En el caso superior derecho también hay un cierto incremento, pero no toda la variación común entre Y y V, pues parte de ella ya era explicada por X. En los casos inferiores la incorporación de V al modelo no incrementa la varianza explicada. En el de la izquierda, porque entre las variables predictoras hay correlación perfecta ($r_{xv} = 1$) y, por tanto, explican la misma parte de Y. En el segundo, porque Y y V son linealmente independientes ($r_{yv} = 0$).

Otra cuestión diferente que nos podemos plantear es si el incremento de varianza explicada al incluir una nueva variable predictora merece la pena. En algunos casos, la medición de variables predictoras tiene un alto coste (económico, de tiempo, etc.), y sólo merecerá la pena hacerlo si la proporción de varianza explicada se incrementa significativamente con respecto a la que ya había antes de incluir esa variable. No podemos aportar criterios para tomar estas decisiones restringiéndonos exclusivamente a la estadística descriptiva, así que remitimos al lector a los textos de estadística inferencial.

Organización y descripción de datos con más de una variable

8

En este capítulo vamos a exponer procedimientos para organizar y analizar conjuntamente la información concerniente a dos variables. Aunque ya hemos tratado esta cuestión para el caso en que las dos variables son cuantitativas (diagrama de dispersión, covarianza, correlación y regresión), en este capítulo vamos a abordar otras situaciones en las que las variables involucradas están medidas en otros tipos de escalas. Nos detendremos especialmente en el caso en que se estudia la relación entre dos variables cualitativas. Terminaremos con algunas ideas respecto al caso de tres variables.

8.1. EL CASO DE DOS VARIABLES CUALITATIVAS

8.1.1. Organización de los datos

Cuando queremos estudiar la relación entre dos variables cualitativas es conveniente empezar, al igual que cuando nos referíamos a una única variable, construyendo una distribución de frecuencias. En este caso se trata de una distribución de frecuencias conjuntas, conocida también como *tabla de contingencia*. Una tabla de contingencia es un cuadro de doble entrada, organizada de forma que las modalidades de una variable ocupan las filas, las de la otra variable ocupan las columnas, y en cada casilla se incluye la frecuencia conjunta de ambas modalidades.

> Se llama *frecuencia conjunta* de dos modalidades de dos variables, n_{ij}, al número de veces que aparecen combinadamente ambas modalidades; es decir, al número de casos que adoptan, simultáneamente, la modalidad i en la primera variable y la modalidad j en la segunda variable.

A lo largo de este capítulo emplearemos la nomenclatura que aparece en la tabla 8.1, que representa la tabla de contingencia de las variables cualitativas X_i (con I categorías) e Y_i (con J categorías); por tanto, la tabla tiene I filas y J columnas:

TABLA 8.1

Nomenclatura para la representación de las distribuciones de frecuencias conjuntas o tablas de contingencia

		Y_j				
		$j = 1$	$j = 2$	...	J	
X_i	$i = 1$	n_{11}	n_{12}	...	n_{1J}	n_{1+}
	$i = 2$	n_{21}	n_{22}	...	n_{2J}	n_{2+}
	⋮	⋮	⋮	...	⋮	⋮
	I	n_{I1}	n_{I2}	...	n_{IJ}	n_{I+}
		n_{+1}	n_{+2}	...	n_{+J}	N

Supongamos, por ejemplo, que queremos estudiar la comorbilidad, es decir, la tendencia de los trastornos a presentarse conjuntamente. Para ello analizamos 400 casos clínicos y anotamos, por un lado, cuál es el trastorno principal diagnosticado a cada uno y, por otro, si ese trastorno se presenta junto con un trastorno afectivo. La distribución de frecuencias conjuntas de las variables trastorno principal y presencia/ausencia de trastornos afectivos podría ser la que aparece en la tabla 8.2 (datos no reales).

TABLA 8.2

Distribución de frecuencias conjuntas de las variables trastorno principal y presencia comórbida de trastorno afectivo

		Trastorno principal					
		Esquizofrenia	Alcoholismo	Trastorno de alimentación	Trastorno de personalidad	Trastorno obsesivo	
Trastorno afectivo	SÍ	50	78	32	48	20	228
	NO	50	42	48	12	20	172
		100	120	80	60	40	400

En la tabla 8.2 aparecen los siguientes elementos:

a) En la casilla de la esquina inferior derecha aparece *N*, que es el total de observaciones incluidas en la tabla.
b) En cada casilla interior aparece el número de casos con presencia o ausencia de trastorno afectivo (fila *i*) y que han sido diagnosticados en cada

trastorno principal (columna *j*); esto es, la frecuencia conjunta para las modalidades respectivas, n_{ij}. El conjunto de estas frecuencias constituye la *distribución de frecuencias conjuntas*.

c) En los laterales derecho e inferior de la tabla se añaden una fila y una columna con las sumas de las frecuencias por filas (n_{i+}) y columnas (n_{+j}). Éstas no son otra cosa que las distribuciones de frecuencias simples de cada variable (la presencia/ausencia de trastorno afectivo en el lateral derecho y el trastorno principal en el lateral inferior), y reciben el nombre de *distribuciones marginales*.

d) En cada fila (o columna) aparecen las llamadas *distribuciones condicionales*. Son las distribuciones de frecuencias en una variable, pero condicionada (o restringida) a que sean casos que adoptan un valor específico en la otra.

Las distribuciones conjuntas se pueden expresar también como frecuencias relativas (tabla 8.3), dividiendo cada frecuencia conjunta absoluta por el tamaño total de la muestra (400 en este caso): $p_{ij} = n_{ij}/N$.

TABLA 8.3

Distribución de frecuencias relativas conjuntas de las variables trastorno principal y presencia comórbida de trastorno afectivo

		Trastorno principal					
		Esquizofrenia	Alcoholismo	Trastorno de alimentación	Trastorno de personalidad	Trastorno obsesivo	
Trastorno afectivo	SÍ	0,125	0,195	0,080	0,120	0,050	0,570
	NO	0,125	0,105	0,120	0,030	0,050	0,430
		0,250	0,300	0,200	0,150	0,100	1,000

Las distribuciones de frecuencias simples de este ejemplo, que hemos denominado distribuciones marginales, son las que aparecen en la tabla 8.4, y se calculan mediante $p_{+j} = n_{+j}/N$ y $p_{i+} = n_{i+}/N$, respectivamente.

Como ejemplo de distribuciones condicionales hemos extraído la distribución de los trastornos principales para los que presentan un trastorno afectivo comórbido, y la presencia/ausencia de trastorno afectivo de los que tienen trastorno de alimentación como trastorno principal (tabla 8.5). Emplearemos una barra vertical (|) para indicar que se trata de una frecuencia condicional; se debe leer «dado que» o «condicionado a que». Para calcular las frecuencias relativas condicionales se divide la frecuencia conjunta absoluta por el marginal correspondiente de su fila o columna, de forma que la distribución condicional completa en frecuencias relativas vuelve a sumar 1. Las frecuencias relativas condicionales se calculan mediante $p_{i|j}=n_{ij}/n_{+j}$ y $p_{j|i}=n_{ij}/n_{i+}$, respectivamente.

TABLA 8.4

Distribuciones de frecuencias marginales de las variables a) trastorno principal y b) presencia comórbida de trastorno afectivo

a)

Trastorno principal	n_{+j}	p_i
Esquizofrenia	100	0,250
Alcoholismo	120	0,300
Tr. de alimentación	80	0,200
Tr. de personalidad	60	0,150
Trastorno obsesivo	40	0,100
	400	1,000

b)

Trastorno afectivo	n_{+j}	p_i
Sí	228	0,570
No	172	0,430
	400	1,000

TABLA 8.5

Distribuciones de frecuencias condicionales: a) la de la variable trastorno principal, condicionada a trastorno afectivo, presente, y b) la de la variable presencia comórbida de trastorno afectivo, condicionada a trastorno principal: trastorno de alimentación

a) **Trastorno activo: presente**

Trastorno principal	$n_{j\|i=1}$	$p_{j\|i=1}$
Esquizofrenia	50	0,219
Alcoholismo	78	0,342
Tr. de alimentación	32	0,140
Tr. de personalidad	48	0,211
Trastorno obsesivo	20	0,088
	228	1,000

b) **Trastorno principal: trastorno de alimentación**

Trastorno afectivo	$n_{i\|j=3}$	$p_{i\|j=3}$
Sí	32	0,400
No	48	0,600
	80	1,000

Es frecuente mostrar la tabla completa con todas las distribuciones condicionales en una variable, para cada modalidad de la otra. Esto se puede hacer para la variable de las filas, condicionada a los valores de las columnas (tabla 8.6*a*), o viceversa (tabla 8.6*b*).

Las tablas de contingencia permiten apreciar ciertas relaciones que son difíciles de advertir desde las distribuciones simples de las variables. Así, en el ejemplo anterior vemos en la distribución marginal de la tabla 8.3 que la variable trastorno principal tiene como categoría modal el alcoholismo (30 por 100), mientras que la categoría trastorno obsesivo es la categoría de menor presencia (10 por 100). Sin embargo, esta distribución no es homogénea para los casos con presencia o ausencia de trastorno afectivo. En la distribución

TABLA 8.6

Distribuciones de frecuencias condicionales completas de las variables a) trastorno principal y b) presencia de trastorno afectivo

a)

		Trastorno principal					
		Esquizofrenia	Alcoholismo	Trastorno de alimentación	Trastorno de personalidad	Trastorno obsesivo	
Trastorno afectivo	SÍ	0,219	0,342	0,140	0,211	0,088	1,000
	NO	0,291	0,244	0,279	0,070	0,116	1,000

b)

		Trastorno principal				
		Esquizofrenia	Alcoholismo	Trastorno de alimentación	Trastorno de personalidad	Trastorno obsesivo
Trastorno afectivo	SÍ	0,500	0,650	0,400	0,800	0,500
	NO	0,500	0,350	0,600	0,200	0,500
		1,000	1,000	1,000	1,000	1,000

condicional de la segunda fila de la tabla 8.6*a*) vemos que, dentro de los que no presentan trastorno afectivo, es la esquizofrenia el trastorno principal de mayor frecuencia (29,1 por 100), mientras que el de menor frecuencia no es el trastorno obsesivo (11,6 por 100), sino el trastorno de personalidad (7 por 100). Frecuentemente, las distribuciones simples de variables esconden tras su apariencia importantes variaciones para los subgrupos que se generan al dividir la población total en las categorías que se definen desde una segunda variable.

8.1.2. Representaciones gráficas

Con respecto a las representaciones gráficas, vamos a ver dos tipos: diagramas de barras conjuntos y rectángulos partidos. Un *diagrama de barras conjunto* se confecciona en esencia como el diagrama de barras sencillo, pero disponiendo barras de diferentes colores (o tonos de gris) para cada categoría de la segunda variable. Así, la representación de la distribución conjunta del trastorno principal y la presencia de trastorno afectivo, en frecuencias absolutas, sería la que aparece en la figura 8.1.

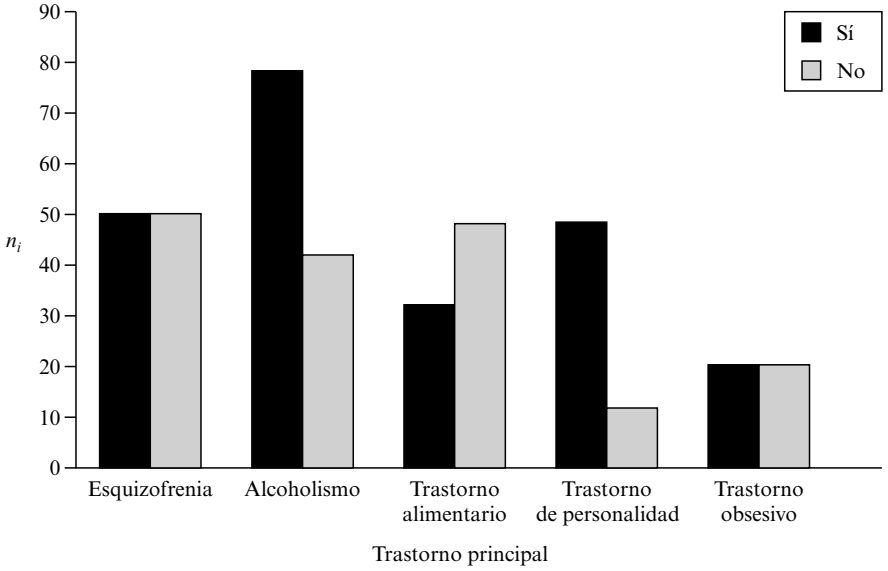


Figura 8.1.—Diagrama de barras conjunto de las variables trastorno principal y presencia de trastorno afectivo (SÍ/NO).

Si nos detenemos a observar sólo las barras de un color (negro o gris) de la figura 8.1, tenemos una representación de la distribución condicional de la variable trastorno principal para los que presentan (o no presentan) un trastorno afectivo comórbido. Así, entre los que presentan un trastorno afectivo (barras negras) lo más frecuente es el alcoholismo y lo menos frecuente es el trastorno obsesivo; por el contrario, entre los que no lo presentan (barras grises), lo más frecuente es la esquizofrenia y lo menos frecuente es el trastorno de personalidad. Las barras de cada color son una representación de la distribución condicional del trastorno principal, condicionada al valor «presente» o «ausente» de la variable trastorno afectivo. También podemos fijarnos en las distribuciones condicionales contrarias, mirando las dos columnas dentro de cada categoría del trastorno principal. Así, dentro del trastorno principal alcoholismo hay más casos con comorbilidad de trastorno afectivo que sin él, mientras que dentro de los trastornos de alimentación ocurre lo contrario.

En el *diagrama de rectángulos partidos* (figura 8.2) se construyen barras, que aquí llamaremos rectángulos, para cada categoría de una de las variables. Estos rectángulos son todos de las mismas dimensiones, pero cada uno se subdivide en parcelas en función de las frecuencias relativas en la otra variable. Es una representación muy apropiada para mostrar las distribuciones condicionales. A continuación se presentan las del ejemplo de la distribución de la presencia/ausencia de trastorno afectivo, condicionada al trastorno principal, expresado en porcentajes.

Esta representación gráfica también se puede construir invirtiendo los papeles de las variables implicadas, según muestra la figura 8.3.

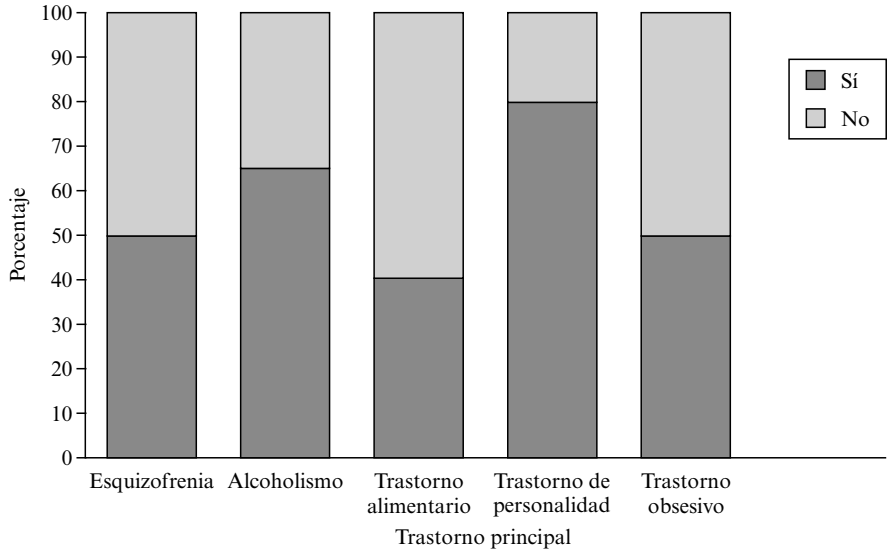


Figura 8.2.—Diagrama de rectángulos partidos de las variables trastorno principal (abscisas) y presencia de trastorno afectivo (ordenadas).

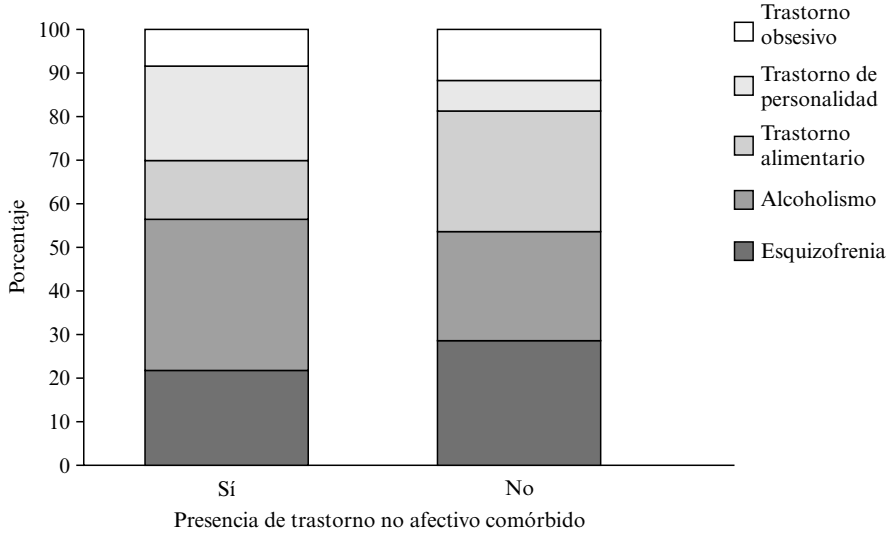


Figura 8.3.—Diagrama de rectángulos partidos de las variables trastorno principal (ordenadas) y presencia de trastorno afectivo (abscisas).

8.1.3. Valoración de la asociación: coeficiente de contingencia

Igual que en capítulos anteriores exponíamos índices para valorar el grado de asociación entre dos variables cuantitativas (covarianza y correlación), en el caso de dos variables cualitativas también se puede valorar la magnitud de la asociación. El caso extremo de ausencia de asociación es aquel en el que las distribuciones condicionales de una variable son iguales para todas las modalidades de la otra. Eso significaría que los porcentajes de cada categoría son los mismos, con independencia del valor que adopta en la otra. Cuando ocurre esto se dice que las variables son independientes. El índice que vamos a exponer se llama *coeficiente de contingencia* y mide el grado en que los datos se desvían de esa situación de independencia. Por tanto, cuanto mayor sea el valor del índice, mayor es el grado de asociación entre las variables. Su cálculo exige la obtención previa del índice conocido como ji-cuadrado (por la letra con la que se representa: X^2). El índice X^2 mide la discrepancia entre las frecuencias de la tabla (n_{ij}) y las frecuencias que se esperaría obtener en el caso de que las variables fueran independientes (m_{ij}). La frecuencia esperada de cada casilla, en condiciones de independencia, se puede estimar mediante la expresión:

$$m_{ij} = \frac{n_{i+} \cdot n_{+j}}{N} \qquad [8.1]$$

En la fórmula [8.1], m_{ij} es la frecuencia esperada para la casilla de la fila i y la columna j. Por ejemplo, la frecuencia esperada (si las variables fueran independientes) de la casilla correspondiente a la presencia de trastorno afectivo (fila 1 de la tabla 8.2) y al trastorno principal esquizofrenia (columna 1 de la tabla 8.2) sería:

$$m_{ij} = \frac{228 \cdot 100}{400} = 57$$

En la tabla 8.7 se incluyen las m_{ij} de todas las casillas de nuestro ejemplo. Observándolas con atención y comparándolas con las de la tabla 8.2, vemos que en algunas hay grandes discrepancias entre las frecuencias observadas y las esperadas (por ejemplo, en las casillas de la columna trastornos de la personalidad), mientras que en otras los valores observados son muy parecidos a los esperados (por ejemplo, en las casillas de la columna trastornos obsesivos).

El índice X^2 se define como el sumatorio, para todas las casillas, de los cocientes entre las diferencias entre la frecuencia observada y la esperada, elevadas al cuadrado y divididas por la frecuencia esperada:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad [8.2]$$

TABLA 8.7

Frecuencias esperadas estimadas (m_{ij}) en condiciones de independencia, del ejemplo con las variables trastorno principal y presencia de trastorno afectivo

		Trastorno principal				
		Esquizofrenia	Alcoholismo	Trastorno de alimentación	Trastorno de personalidad	Trastorno obsesivo
Trastorno afectivo	SÍ	57	68,4	45,6	34,2	22,8
	NO	43	51,6	34,4	25,8	17,2

En nuestro ejemplo sería:

$$X^2 = \frac{(50 - 57)^2}{57} + ... + \frac{(20 - 17{,}2)^2}{17{,}2} = 0{,}8596 + ... + 0{,}4558 = 28{,}315$$

Conocido el valor de X^2, ya podemos calcular el coeficiente de contingencia (C), cuya fórmula es:

$$C = \sqrt{\frac{X^2}{X^2 + N}} \qquad [8.3]$$

En nuestro ejemplo nos daría:

$$C = \sqrt{\frac{28{,}315}{28{,}315 + 400}} = 0{,}257$$

Naturalmente, el lector se estará preguntando cómo interpretar este valor. Para ello hay que tener en cuenta que, cuando hay independencia completa entre las variables, las frecuencias observadas y esperadas coinciden, por lo que se obtendrá $X^2 = 0$ y el coeficiente de contingencia también será $C = 0$. Por el contrario, cuanto mayor sea la asociación entre las variables mayor será la discrepancia entre las frecuencias observadas y esperadas, y, por tanto, mayores serán los valores de X^2 y C. Sin embargo, este índice no tiene un máximo fijo con el que compararlo; se podría calcular un máximo en algunas condiciones. En cualquier caso, la estructura de la fórmula garantiza que el valor de C nunca será mayor de 1. En el capítulo 15, en el contexto de la estadística inferencial, discutiremos una forma más adecuada de interpretar el resultado obtenido con este índice.

8.1.4. Dos variables dicotómicas: coeficiente Phi

Si bien el coeficiente de contingencia se puede emplear siempre que se trate de dos variables cualitativas, en el caso particular en que se trate de dos variables

dicotómicas (variables que sólo admiten dos modalidades o categorías), se puede emplear también el coeficiente *Phi*, que se representa por la letra griega φ. Supongamos que definimos la tabla de contingencia de la tabla 8.8, con dos categorías en cada variable, codificadas como 0 y 1 (las frecuencias conjuntas se representan por las letras *a*, *b*, *c* y *d*).

TABLA 8.8

Esquema de una tabla de contingencia entre dos variables dicotómicas

		Y		
		$Y_1 = 0$	$Y_2 = 1$	
X	$X_1 = 1$	a	b	$(a + b)$
	$X_2 = 0$	c	d	$(c + d)$
		$(a + c)$	$(b + d)$	N

La fórmula del coeficiente *Phi* es:

$$\varphi = \frac{c \cdot b - a \cdot d}{\sqrt{(a + b) \cdot (c + d) \cdot (a + c) \cdot (b + d)}} \qquad [8.4]$$

Sus valores oscilan entre −1 y +1. Con respecto a la interpretación de su magnitud, es similar a la de la correlación de Pearson: los valores ±1 representan grados de asociación máxima, mientras que el valor 0 refleja independencia entre las variables. Sin embargo, para interpretar el sentido de la asociación habrá que mirar la tabla e identificar las combinaciones de categorías más frecuentes. Veámoslo con un ejemplo. Supongamos que disponemos de una tabla de contingencia (tabla 8.9), en la que se resumen las frecuencias conjuntas de las variables fumador (1: SÍ; 0: NO) y sexo (1: hombre; 0: mujer), de una determinada población.

TABLA 8.9

Tabla de contingencia de las variables dicotómicas sexo (X) y consumo de tabaco (Y)

		Y		
		NO (0)	SÍ (1)	
X	Hombre (1)	80	20	100
	Mujer (0)	50	100	150
		130	120	250

El valor del coeficiente será:

$$\varphi = \frac{50 \cdot 20 - 80 \cdot 100}{\sqrt{100 \cdot 150 \cdot 130 \cdot 120}} = \frac{-7.000}{15.297} = -0{,}458$$

El valor indica (a nivel descriptivo) una asociación moderada entre las variables. En cuanto al sentido, que el signo sea positivo o negativo es algo arbitrario, ya que la codificación como 0 o 1 de las dos categorías de cada variable es arbitraria. En nuestro caso, dentro de los hombres hay más no fumadores que fumadores, mientras que entre las mujeres hay más fumadoras que no fumadoras. Por tanto, concluimos que hay una tendencia moderada de asociación, que consiste en una mayor presencia del tabaquismo entre las mujeres.

8.2. EL CASO DE UNA VARIABLE CUALITATIVA Y OTRA CUANTITATIVA

8.2.1. Organización y representación de los datos

Cuando se quiere realizar una descripción conjunta de una variable cualitativa y otra cuantitativa, la opción más eficaz suele ser la descripción de la variable cuantitativa, condicionada a cada categoría de la variable cualitativa. Supongamos, por ejemplo, que hemos medido el nivel de ansiedad-rasgo mediante el test STAI y el tabaquismo (fumadores o no). Ya vimos en el tercer capítulo que una forma bastante completa de describir una variable cuantitativa consiste en informar de una medida de tendencia central y otra de variabilidad (habitualmente la media y la desviación típica). Pues bien, una forma de describir una variable cuantitativa conjuntamente con una cualitativa consiste en exponer una tabla con los estadísticos descriptivos de la variable cuantitativa (en nuestro ejemplo, la ansiedad), para las submuestras que adoptan cada modalidad de la variable cualitativa (en nuestro ejemplo, que sea o no fumador). Además, como muestra la tabla 8.10, se suelen acompañar los estadísticos de la muestra total, como en el presente ejemplo (entre paréntesis se incluye el tamaño de cada submuestra).

TABLA 8.10

Estadísticos descriptivos de la variable ansiedad-rasgo, condicionados al consumo de tabaco

	Ansiedad-rasgo	
Consumo de tabaco	**Media**	**Desv. típica**
Fumadores (N_F = 200)	18,6	5
No fumadores (N_{NF} = 300)	12,3	4
Total (N = 500)	14,82	5,397

La representación gráfica de la tendencia central se hace habitualmente mediante diagramas de barras que reflejan las medias parciales (figura 8.4) o mediante puntos que conforman un polígono. La variabilidad se puede representar simultáneamente incorporando «bigotes» de longitud proporcional a las desviaciones típicas. Así, en los datos anteriores la representación en forma de diagrama de barras quedaría como en la siguiente figura, en la que se aprecia que los fumadores muestran un nivel medio de ansiedad-rasgo mayor que los no fumadores, y que su desviación típica es también mayor.

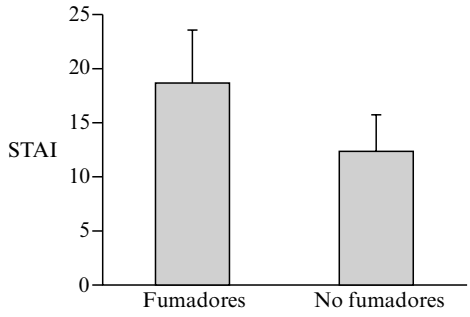


Figura 8.4.—Diagrama de barras que representa la relación entre una variable dicotómica (tabaquismo, en abscisas) y otra cuantitativa (ansiedad, en ordenadas).

8.2.2. Valoración de la asociación: coeficiente biserial-puntual

Hay varias estrategias e índices que permiten valorar la asociación entre una variable cualitativa y otra cuantitativa. Aquí vamos sólo a describir un índice, el *coeficiente biserial-puntual,* que se emplea cuando estamos en esa situación y la variable cualitativa es de naturaleza dicotómica. El ejemplo de la ansiedad y el tabaquismo de la tabla 8.10 es de este tipo, por lo que podemos aprovecharlo para ilustrar el cálculo de este estadístico.

En la fórmula del coeficiente, que se representa por r_{BP}, aparecen las medias en la variable cuantitativa de los dos grupos formados por la variable cualitativa, la desviación típica de la muestra total en la variable cuantitativa y las proporciones de la muestra total que adoptan las dos modalidades (P y Q). En concreto:

$$r_{BP} = \frac{\bar{X}_P - \bar{X}_Q}{S_X} \cdot \sqrt{P \cdot Q} \qquad [8.5]$$

En el ejemplo anterior las proporciones de las dos modalidades son 0,40 para los fumadores (P = 200/500) y 0,60 para los no fumadores (Q = 300/500, o también Q = 1; P = 1 − 0,40). Sustituyendo los estadísticos disponibles, obtenemos:

$$r_{BP} = \frac{18{,}6 - 12{,}3}{5{,}397} \cdot \sqrt{0{,}40 \cdot 0{,}60} = 0{,}572$$

No es casualidad que se represente por la misma letra que la correlación de Pearson. En realidad, el coeficiente biserial-puntual es una adaptación de aquel a estas circunstancias. Su interpretación es la misma. Respecto a su intensidad, oscila entre ±1, y el valor 0 refleja independencia entre las variables. Respecto a su sentido, hay que mirar qué categoría de la variable dicotómica tiene asociada una media mayor en la variable cuantitativa.

8.3. OTROS ÍNDICES DE ASOCIACIÓN PARA DOS VARIABLES

Si en algo es rica la estadística es en índices y coeficientes. Se han propuesto literalmente cientos de ellos, tratando de que en cada circunstancia se refleje la asociación de la manera más eficaz. Así, cuando se trata de una variable cuantitativa y otra cualitativa con más de dos modalidades, cuando se trata de variables que en su naturaleza no son dicotómicas, sino dicotomizadas, cuando se atiende a las propiedades ordinales de los datos, etc., disponemos de coeficientes específicos cuyas fórmulas y propiedades exceden la amplitud que queremos dar a este texto. En este capítulo hemos descrito unos pocos, pero animamos al lector interesado a profundizar en otras fuentes. Hemos evitado intencionadamente los estadísticos específicos de variables ordinales, dado que casi no hemos hablado de este tipo de variables hasta aquí. Sin embargo, hemos incluido en el apéndice del capítulo 8 el coeficiente más utilizado para ese tipo de variables (el coeficiente de Spearman). Para otros casos y coeficientes remitimos al lector a otras fuentes (Howell, 2009; Pardo y San Martín, 2010; Sánchez Carrión, 1999; Solanas, Salafranca, Fauquet y Núñez, 2005).

8.4. DESCRIPCIÓN CONJUNTA DE TRES VARIABLES

Cuando se quieren estudiar tres variables conjuntamente, se pueden presentar de nuevo multitud de circunstancias diferentes. En cada caso habrá que emplear la estrategia adecuada. Vamos al menos a indicar lo que hacer cuando se quiere representar una variable dependiente cuantitativa en función de dos independientes cualitativas. En estos casos la mejor opción es emplear polígonos de frecuencias, como en el caso de una sola variable independiente, pero añadiendo un polígono para cada nivel de la segunda independiente. Ésta es, además, la forma de estudiar gráficamente la interacción entre dos variables (véase en León y Montero, 2003). Vamos a ilustrarlo continuando con el ejemplo del tabaquismo, añadiendo ahora las medias en ansiedad de cada grupo de tabaquismo, pero separadas en función del sexo. Supongamos que esas medias son las que aparecen en la tabla 8.11; la representación gráfica es la de la figura 8.5.

Ya vimos anteriormente que los fumadores muestran un nivel medio superior en ansiedad-rasgo. Al estudiar conjuntamente la relación, vemos que el gráfico apunta a que ese efecto no es independiente del sexo, sino que el nivel de ansiedad es mayor en hombres que en mujeres. La diferencia en ansiedad entre los hombres fumadores y no fumadores es mayor que entre las mujeres fumadoras y no fuma-

TABLA 8.11

Medias en ansiedad rasgo, en función del sexo y el tabaquismo

	Ansiedad-rasgo	
Consumo de tabaco	**Hombres**	**Mujeres**
Fumadores	20,0	17,0
No fumadores	11,0	14,8
	15,2	14,6

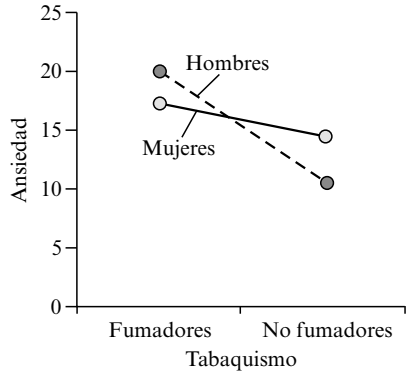


Figura 8.5.—Polígono de frecuencias que relaciona el tabaquismo con la ansiedad, separando según la variable sexo.

doras. Al dividir los casos en función del sexo (la otra variable «independiente»), vemos que esta tendencia no es idéntica en ambos sexos. Ambos grupos parecen mostrar la misma tendencia (mayor ansiedad de los fumadores), pero es más acusada en los hombres que en las mujeres.

PROBLEMAS Y EJERCICIOS

1. Hemos obtenido información referente a la *ciudad de procedencia* (M: Madrid, B: Barcelona, V: Valencia y S: Sevilla) y *situación laboral* (D: desempleado y A: en activo) en una muestra de 20 graduados en filosofía y letras de los últimos dos cursos. Los resultados se muestran en la tabla siguiente. A partir de estos datos, construya una distribución de frecuencias conjuntas y elabore una representación gráfica apropiada.

Ciudad	M	B	M	M	V	S	M	S	M	V	V	M	M	S	M	M	V	B	V	S
Situación	D	A	A	D	D	A	D	D	A	D	A	D	D	D	A	D	A	D	A	D

2. Obtenga, con los datos del ejercicio anterior, las tres distribuciones de frecuencias relativas (conjuntas y condicionales). A continuación, conteste a las siguientes cuestiones:

a) ¿Qué porcentaje de graduados son de Sevilla y están en activo?
b) De los que son de Madrid, ¿qué porcentaje están desempleados?
c) ¿Qué porcentaje de graduados desempleados son de Valencia?

3. Supongamos que evaluamos a 50 trabajadores (20 varones y 30 mujeres) en la variable *nivel de estrés laboral* (bajo, medio y alto). Sabiendo que la mitad de los trabajadores dicen que tienen un nivel de estrés medio, que hay 10 mujeres y 10 varones que tienen un nivel de estrés alto y 2 mujeres con un nivel de estrés bajo, conteste a las siguientes preguntas:

a) Elabore la distribución de frecuencias absolutas conjuntas para las variables *sexo* y *nivel de estrés laboral*.
b) ¿Qué porcentaje de sujetos son varones y tienen un nivel de estrés bajo?
c) De las mujeres, ¿qué porcentaje tienen un nivel de estrés alto?
d) De los que tienen un nivel de estrés alto, ¿qué porcentaje son varones?

4. A partir de las siguientes distribuciones de frecuencias relativas condicionadas de la variable *nivel de ingresos* (bajo, medio y alto) sobre la variable *estado civil* (casado/a, soltero/a y divorciado/a) que aparece en la tabla siguiente, y sabiendo que los datos se obtuvieron sobre una muestra compuesta por 200 casados, 300 solteros y 150 divorciados, obtenga la correspondiente tabla de contingencia con las frecuencias absolutas.

		Nivel de ingresos			
		Bajo	Medio	Alto	
Estado civil	Casado/a	0,70	0,25	0,05	1,00
	Soltero/a	0,55	0,25	0,20	1,00
	Divorciado/a	0,60	0,20	0,20	1,00

5. Obtenga los índices de asociación adecuados para las variables del ejercicio 1 e interprete los resultados obtenidos.

6. El director de un centro escolar está interesado en averiguar los factores que afectan al *rendimiento escolar* en los estudiantes de secundaria. Para ello ha seleccionado una muestra de 10 estudiantes (5 varones y 5 mujeres) y les ha preguntado sobre la *dificultad percibida de la asignatura* (A: alta, M: media y B: baja). Los resultados obtenidos han sido los siguientes:

Sujeto	Sexo	Dificultad de la asignatura	Rendimiento
1	Mujer	A	3
2	Varón	B	6
3	Varón	M	2
4	Varón	A	3
5	Mujer	A	5
6	Varón	A	2
7	Mujer	M	6
8	Varón	B	7
9	Mujer	M	7
10	Mujer	B	8

a) Elabore la distribución de frecuencias conjunta para las variables *sexo* y *dificultad de la asignatura* y confeccione la representación gráfica entre ambas que considere más adecuada. Una vez obtenida, responda a las siguientes cuestiones: ¿qué porcentaje de alumnos son varones y consideran la asignatura de dificultad alta?; de los alumnos varones, ¿qué porcentaje consideran la asignatura de dificultad baja?; de los alumnos que consideran la asignatura de dificultad alta, ¿qué porcentaje son mujeres?

b) Calcule un índice de asociación adecuado para las variables *sexo* y *rendimiento*.

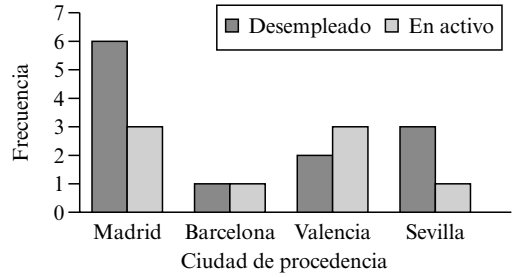
c) Describa la relación entre las variables *dificultad* y *rendimiento* de la forma que le parezca más adecuada y elabore la correspondiente representación gráfica.

d) Represente en una sola gráfica la relación entre las variables *sexo*, *dificultad* y *rendimiento*.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1.

		Ciudad			
		Madrid	Barcelona	Valencia	Sevilla	
Situación laboral	Desempleado (D)	6	1	2	3	12
	En activo (A)	3	1	3	1	8
		9	2	5	4	20



2.

		Ciudad			
		Madrid	Barcelona	Valencia	Sevilla	
Situación laboral	Desempleado	0,30	0,05	0,10	0,15	0,60
	En activo	0,15	0,05	0,15	0,05	0,40
		0,45	0,10	0,25	0,20	1,00

		Ciudad			
		Madrid	Barcelona	Valencia	Sevilla
Situación laboral	Desempleado	0,67	0,50	0,40	0,75
	En activo	0,33	0,50	0,60	0,25
		1,00	1,00	1,00	1,00

		Ciudad			
		Madrid	Barcelona	Valencia	Sevilla	
Situación laboral	Desempleado	0,50	0,08	0,17	0,25	1,00
	En activo	0,375	0,125	0,375	0,125	1,00

a) El 5 por 100.
b) El 67 por 100.
c) El 17 por 100.

3. *a*)

		Nivel de estrés laboral			
		Bajo	Medio	Alto	
Sexo	Varón	3	7	10	20
	Mujer	2	18	10	30
		5	25	20	50

b) El 6 por 100.
c) El 33,3 por 100.
d) El 50 por 100.

4.

		Nivel de ingresos			
		Bajo	Medio	Alto	
Estado civil	Casado/a	140	50	10	200
	Soltero/a	165	75	60	300
	Divorciado/a	90	30	30	150
		395	155	100	650

5. Frecuencias esperadas:

		Ciudad				
		Madrid	Barcelona	Valencia	Sevilla	
Situación laboral	Parado	5,4	1,2	3	2,4	12
	En activo	3,6	0,8	2	1,6	8
		9	2	5	4	20

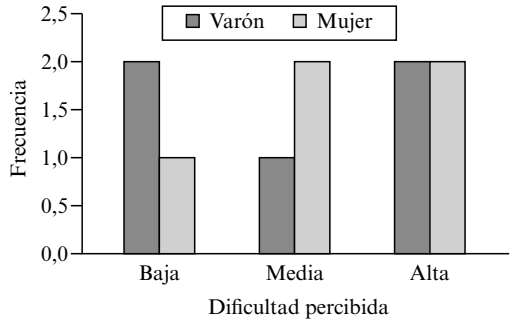
$$X^2 = 1{,}458; \quad C = 0{,}261$$

Dado que el valor del coeficiente de contingencia no es cero, pero está próximo a ese valor, se puede concluir que las variables *situación laboral* y *ciudad de procedencia* no son independientes, pero su asociación es moderada o baja.

6. *a*)

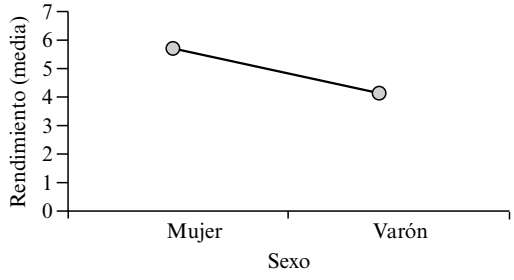
Dificultad de la asignatura

Sexo	Bajo	Medio	Alto	
Varón	2	1	2	5
Mujer	1	2	2	5
	3	3	4	10



— Porcentaje de alumnos varones y que consideran la asignatura de dificultad alta: 20 por 100.
— De los varones, porcentaje que consideran la asignatura de dificultad baja: 40 por 100.
— De los que consideran la asignatura de dificultad alta, porcentaje de mujeres: 50 por 100.

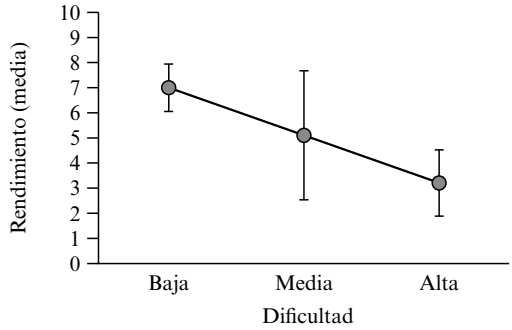
b) $r_{BP} = 0{,}42$. Gráficamente:



Por tanto, existe una relación moderada entre las variables *sexo* y *rendimiento*, siendo éste mayor en promedio en las mujeres.

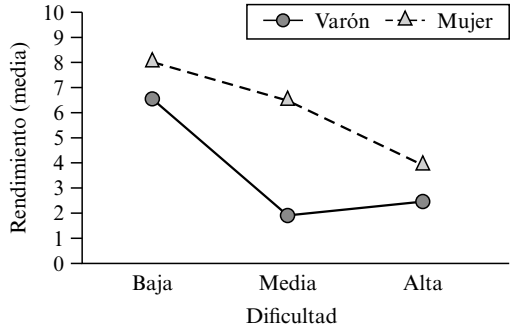
c)

Dificultad de la asignatura	Rendimiento	
	Media	Desv. típ.
Baja	7,00	1,00
Media	5,00	2,65
Alta	3,25	1,26



d)

Dificultad de la asignatura	Sexo	Rendimiento
Baja	Mujer	8,00
	Varón	6,50
Media	Mujer	6,50
	Varón	2,00
Alta	Mujer	4,00
	Varón	2,50



APÉNDICE

El coeficiente de correlación de Spearman

Son muchas las circunstancias en las que se prefiere indagar en las propiedades ordinales de los datos, en lugar de interpretarlas como auténticas magnitudes. Es obligado hacerlo cuando, las variables están medidas a nivel ordinal (capítulo 1), pero también se hace cuando siendo de nivel cuantitativo, no cumplen algunas de las condiciones o supuestos que exigirían las técnicas estadísticas apropiadas para variables cuantitativas. Aunque dejamos esta cuestión para otra ocasión, al menos queremos describir el índice de asociación más utilizado con variables ordinales. Se trata del coeficiente de correlación de Spearman.

Consiste en aplicar la fórmula del coeficiente de Pearson a los órdenes de los valores. Por tanto, primero es necesario convertir los valores registrados en órdenes. Supongamos que hemos obtenido los cinco pares de valores en las variables X e Y que aparecen en las dos primeras columnas de la tabla siguiente. Convertimos esos valores en órdenes, asignando en cada variable un 1 al valor más pequeño, un 2 al siguiente, y así sucesivamente hasta llegar al valor mayor, al que en este caso le asignamos un 5 (por haber sólo 5 pares de valores). Los órdenes (O_X y O_Y) se incluyen en las siguientes columnas de la tabla. A continuación obtenemos las diferencias de órdenes dentro de cada par y elevamos esas diferencias al cuadrado (elementos necesarios para sustituir en la fórmula que mostramos más abajo).

X	Y	O_X	O_Y	$(O_X - O_Y)^2$
12	4	3	2	1
9	2	1	1	0
10	8	2	4	4
13	10	4	5	1
18	6	5	3	4
				10

La fórmula del coeficiente es:

$$r_S = 1 - \frac{6 \cdot \sum (O_X - O_Y)^2}{N \cdot (N^2 - 1)}$$

Sustituyendo en la fórmula obtenemos:

$$r_S = 1 - \frac{6 \cdot 10}{5 \cdot (25 - 1)} = 0{,}50$$

Como se trata de una aplicación del coeficiente de Pearson a los órdenes, su interpretación es la misma que la de aquel: sus valores oscilan entre ±1, y el valor

0 indica independencia entre las variables. En el ejemplo anterior, hemos obtenido una relación moderada y directa entre las variables.

Aunque por ahora nos limitaremos a una valoración descriptiva del coeficiente de Spearman, contamos con procedimientos más sofisticados para evaluar el significado de los coeficientes. En el capítulo 15 veremos cómo se aplicarían a r_{xy}; su aplicación a r_s se puede consultar en Howell (2009).

PARTE TERCERA
Probabilidad

9 Introducción a la probabilidad

9.1. INTRODUCCIÓN

En la vida cotidiana hay muchas situaciones cuyo resultado no podemos predecir con seguridad. Lo mismo ocurre en el estudio del comportamiento. ¿Qué responderá un individuo ante una de las láminas del test de Rorscharch?, ¿cuánto tardará en dar la primera asociación libre a la palabra «dolor»?, ¿qué tarjetas levantará el próximo participante en la tarea de las cuatro tarjetas de Wason?, ¿se rebelará un individuo con ciertas características al encontrarse en la situación de presión de grupo de Milgram?, ¿qué puntuación obtendrá en *impulsividad* al aplicarle el test de Barrat?

Aunque no podamos predecir con total seguridad la respuesta a estos interrogantes, podríamos tratar de valorar las opciones de que el resultado sea uno concreto de entre los posibles. Vamos a exponer con mayor detalle tres ejemplos que utilizaremos a lo largo del capítulo. El primero es un estudio de condicionamiento llevado a cabo con adolescentes. Cada vez que el adolescente elija una de las dos cajas que tiene en la pantalla, recibirá una cantidad mayor de puntos que si elije la otra. Aunque las cajas se distinguen en muchas características, la única relevante es que tenga un logotipo a la izquierda o a la derecha. Sin embargo, antes de comenzar el proceso de reforzamiento tenemos que registrar las elecciones que hacen las personas en ausencia de ese efecto (es la línea base). Observamos y anotamos la primera elección de tres adolescentes (I, caja con el logotipo a la izquierda; D, caja con el logotipo a la derecha).

El segundo ejemplo consiste en elegir al azar a uno de los diez miembros de una colonia de titís de cabeza blanca para realizar una sesión de observación en una sala enriquecida. Algunos datos de estos titís aparecen en el cuadro 9.1. Para el tercer ejemplo empleamos una tarea de asociación libre; le pedimos a un artista que nos diga las palabras que le sugiera el desencadenante «tropezar», y anotamos el número de palabras que nos dice en asociación libre. Aprovecharemos estos ejemplos para proponer una terminología, definir el concepto de probabilidad y exponer algunos teoremas. Para ello utilizaremos la teoría de conjuntos (Jáñez, 1989).

CUADRO 9.1

Los diez miembros de la colonia de titís de cabeza blanca. La columna «nacimiento» se refiere a si éste se produjo en libertad, en cautividad en la propia colonia, o en cautividad en otra colonia

Nombre	Sexo	Nacimiento	Edad (meses)
Montri	Hembra	Libertad	12
Chulina	Hembra	Cautividad, colonia	8
Pinta	Hembra	Cautividad, externa	18
Pardi	Macho	Cautividad, externa	26
Pancho	Macho	Libertad	42
Pachi	Macho	Libertad	36
Leyre	Hembra	Cautividad, externa	16
Celia	Hembra	Cautividad, externa	22
Rasan	Macho	Cautividad, colonia	45
Tangue	Hembra	Cautividad, colonia	28

9.2. DEFINICIONES

El azar tiene que ver con aquellos eventos cuyo resultado no podemos predecir con certeza; llamaremos a estos eventos *experimentos aleatorios*. Por ejemplo, cuando elegimos al azar a un individuo de la población para preguntarle su opinión sobre una cuestión de actualidad, ésta puede ser favorable o desfavorable. La observación de las cajas que eligen nuestros tres adolescentes en el estudio sobre condicionamiento, la extracción de un individuo de nuestra colonia o la observación del número de palabras asociadas por nuestro artista son también experimentos aleatorios.

Esto no quiere decir que, por ejemplo, la opinión del encuestado sea algo aleatorio. Seguramente tiene una opinión formada sobre la cuestión que se pregunta, y su respuesta no depende del azar. Lo que depende del azar en este contexto es el procedimiento de elección de un individuo, y sólo uno, de la población; es éste el que da sentido al término «aleatorio». La incertidumbre se refiere a si el individuo extraído será de los que tienen opinión favorable o de los que están en contra.

Todo experimento aleatorio tiene dos o más resultados posibles, que nosotros llamaremos *sucesos elementales*. En un experimento que tuviera sólo un resultado posible no habría incertidumbre; estrictamente hablando, no podríamos hablar de experimento aleatorio. La realización de un experimento aleatorio da lugar a un suceso elemental de entre los posibles. Se puede decir que se ha observado una muestra o porción de lo que hubiera podido ser observado, o que lo observado es un subconjunto de lo potencialmente observable. Por eso, al conjunto de todos los resultados posibles de un experimento aleatorio, o sucesos elementales, se le llama *espacio muestral,* el cual se representa por **E**. Los espacios muestrales de nuestros tres ejemplos estarían formados, respectivamente, por las ocho posibles

combinaciones de elecciones de cajas, los diez miembros de la colonia y los posibles números de palabras de la respuesta del artista:

Cajas elegidas, **E** = {III, IID, IDI, DII, DDI, DID, IDD, DDD}

Colonia de titís, **E** = {Montri, Chulina, ..., Tangue}

Palabras asociadas, **E** = {1, 2, 3, ...}

Al hecho de que la realización del experimento aleatorio genere un suceso elemental nos referiremos como *verificación* del suceso. Sobre los espacios muestrales, como conjuntos que son, se pueden definir subconjuntos, que denominaremos *sucesos*, y que representaremos por letras mayúsculas. Todos los subconjuntos que podrían ser definidos sobre **E** forman una clase, sobre la que definiremos algunas operaciones. Pero antes vamos a ver algunos ejemplos de sucesos definidos sobre los espacios muestrales de nuestros tres ejemplos. Así, sobre el ejemplo de las cajas podemos definir los sucesos A = {III, IDI}, B = {Ø}, C = {III, DDD}, etc.; sobre la colonia de titís podemos definir los sucesos A = = {Montri, Pancho, Pachi}, B = {Pardi, Pancho, Pachi, Rasan}, etc.; y sobre el estudio de asociación libre podemos definir los sucesos A = {1, 2, 3}, B = {5}, etc. Éstos no son más que ejemplos de los muchos sucesos que se podrían definir sobre estos espacios muestrales. Aunque para definir un suceso basta con definir un subconjunto cualquiera de **E**, normalmente los sucesos con los que trabajaremos no se constituirán de forma arbitraria, sino con los sucesos elementales que cumplen alguna condición relevante para nosotros. Así, podríamos definir un suceso de nuestro experimento aleatorio de las cajas tomando como base la característica de que al menos uno de los adolescentes elija la caja de la izquierda.

Los siguientes son ejemplos de sucesos definidos sobre nuestros experimentos aleatorios siguiendo alguna regla:

Cajas, A: «Al menos un adolescente elije la de la izquierda»
A = {III, IID, IDI, DII, IDD, DID, DDI}

B: «Los adolescentes eligen la misma caja»
B = {III, DDD}

Colonia, A: «Haber nacido en libertad»
A = {Montri, Pancho, Pachi}

B: «Ser hembra»
B = {Montri, Chulina, Pinta, Leyre, Celia, Tangue}

C: «Tener más de 24 meses»
C = {Pardi, Pancho, Pachi, Rasan, Tangue}

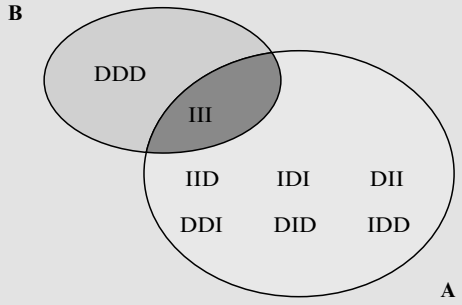
Palabras asociadas, A: «El artista genera menos de cuatro palabras asociadas»
A = {1, 2, 3}

Es habitual representar gráficamente los espacios muestrales mediante diagramas de Venn, tal y como aparece en el cuadro 9.2.

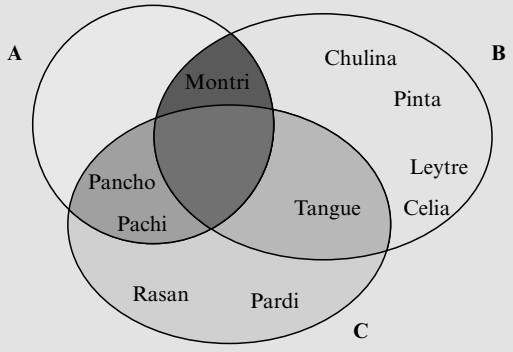
CUADRO 9.2

Representación de los espacios muestrales de nuestros tres ejemplos, mediante diagramas de Venn, y de algunos sucesos definidos sobre ellos

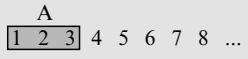
a) Espacio muestral de la elección entre dos cajas, por tres adolescentes, con los sucesos A y B (véase texto). Como se puede apreciar, la intersección está compuesta por el único suceso elemental común entre esos dos subconjuntos (III).



b) Espacio muestral de la elección de un miembro de la colonia de titís, con la representación de los sucesos A, B y C (véase texto).



c) Espacio muestral del experimento aleatorio del estudio sobre asociación de palabras. En él aparecen puntos suspensivos para indicar que el espacio está incompleto; también se representa el suceso A.



Un suceso se verifica cuando el experimento aleatorio genera uno de los sucesos elementales que integran el subconjunto que lo define. Así, si los tres adolescentes eligen la caja con el logotipo a la derecha, se verifica el suceso B definido sobre el primer ejemplo, mientras que si los tres eligen la de la izquierda se verifican tanto el A como el B.

En algunas ocasiones se definen sucesos a partir de subconjuntos vacíos, como por ejemplo el suceso definido sobre nuestra colonia mediante la regla D: «Tener más de 50 meses», cuyo subconjunto está vacío:

$$D = \{\emptyset\}$$

Estos sucesos reciben el nombre de *suceso imposible*. En otras ocasiones se definen sucesos cuyo subconjunto está formado por todos los elementos del espacio muestral, como por ejemplo el suceso *F*, definido sobre nuestra colonia mediante la regla «tener más de 6 meses»; este tipo de sucesos reciben el nombre de *suceso seguro*.

Vamos a definir varias operaciones sobre sucesos:

a) Llamaremos *unión* de dos sucesos al subconjunto de **E** formado por los sucesos elementales que pertenecen al menos a uno de ellos; se representa por el signo $\cup$. Así, la unión de los sucesos A y C, definidos sobre nuestro segundo ejemplo, estaría formada por los sucesos elementales que pertenecen a A, a C o a ambos a la vez. Es decir:

$$A \cup C: \quad \{\text{Montri, Pancho, Pachi, Pardi, Rasan, Tangue}\}$$

b) Llamaremos *intersección* de dos sucesos al subconjunto de **E** formado por los sucesos elementales que pertenecen simultáneamente a ambos sucesos; se representa por el signo $\cap$. Así, la intersección de esos mismos sucesos, $A \cap C$, estará formada por los sucesos elementales que pertenezcan a ambos subconjuntos, es decir:

$$A \cap C: \quad \{\text{Pancho, Pachi}\}$$

Cuando la intersección de dos sucesos es un subconjunto vacío se dice que son sucesos *incompatibles* o *excluyentes*.

c) La *diferencia de dos sucesos* es el subconjunto de **E** integrado por los sucesos elementales que pertenecen al primero, pero no al segundo. Así, en los sucesos del experimento de la colonia de titís:

$$C - A: \quad \{\text{Pardi, Rasan, Tangue}\}$$

d) Llamaremos *complementario* de un suceso al subconjunto de **E** integrado por los sucesos elementales no incluidos en ese suceso. Los representaremos por un apóstrofe ($'$) junto a la letra que designa el suceso. Así, A' representa al com-

plementario del suceso A; en el ejemplo de las cajas está integrado por los sucesos elementales:

$$A' = \{DDD\}$$

Representaremos por N al número de sucesos elementales que integran el espacio muestral y por n_A al número de sucesos elementales que constituyen el suceso A. Igualmente, $n_{A \cap B}$ y $n_{A \cup B}$ representarán los tamaños de los subconjuntos que constituyen la intersección y la unión, respectivamente, entre A y B.

Vamos a reunir los términos definidos hasta aquí:

Un *experimento aleatorio* es toda acción cuyo resultado no se puede predecir con certeza.

Cada uno de los resultados posibles de un experimento aleatorio se llama *suceso elemental;* su conjunto constituye el *espacio muestral* (**E**) del experimento aleatorio.

La *verificación* de un suceso elemental es la observación de ese suceso elemental al realizar el experimento aleatorio.

Un *suceso* es cualquier subconjunto de los elementos de un espacio muestral.

Dos sucesos son *incompatibles* o *excluyentes* si no se pueden verificar simultáneamente, por no tener elementos comunes.

El *complementario* de un suceso es el subconjunto de sucesos elementales del espacio muestral que no forman parte de ese suceso.

La *intersección* de dos sucesos es el subconjunto de elementos del espacio muestral que, simultáneamente, están incluidos en los subconjuntos de ambos sucesos.

La *unión* de dos sucesos es el subconjunto de elementos del espacio muestral que están incluidos en al menos uno de esos sucesos.

Algunas de estas definiciones se pueden extender con facilidad a casos que involucran a más de dos sucesos. Así, se puede hablar de la unión o la intersección de k sucesos mediante la extensión simple de las definiciones anteriores. La primera estaría constituida por los sucesos elementales que están incluidos en al menos uno de los subconjuntos integrantes de esos k sucesos. La segunda estaría constituida por los sucesos elementales que participan, simultáneamente, en los k subconjuntos integrantes. Igualmente, se puede hablar del complementario de la unión o de la intersección de sucesos, como por ejemplo $(A \cup B)'$ o $(A \cap B)'$.

Los espacios muestrales se clasifican en espacios muestrales finitos e infinitos; a su vez, los infinitos se subdividen en numerables y no numerables. Veamos las características de cada uno de ellos:

a) *Espacios muestrales finitos*. Un espacio muestral es finito si tiene un número de sucesos elementales finito, como por ejemplo nuestro estudio de las cajas o la extracción de un individuo de nuestra colonia de titís.
b) *Espacios muestrales infinitos numerables*. Tiene infinitos sucesos elementales, pero éstos se pueden poner en correspondencia biunívoca con los números naturales, como por ejemplo nuestro estudio sobre asociación de palabras.
c) *Espacios muestrales infinitos no numerables*. Tiene infinitos sucesos elementales, pero éstos no se pueden poner en correspondencia biunívoca con los números naturales, como por ejemplo el tiempo invertido en realizar una tarea.

9.3. DEFINICIÓN DE PROBABILIDAD

Ya estamos en disposición de abordar la definición del propio concepto de probabilidad. Este concepto hace referencia a cómo al estudiar la repetición de un experimento aleatorio un número grande de veces, éste comienza a tener resultados globalmente previsibles y a mostrarse sujeto a ciertas leyes. La probabilidad es un concepto ideal, ya que se refiere a las frecuencias con las que ocurrirían las cosas en el caso hipotético de que los eventos se repitiesen un número infinitamente grande de veces y en las mismas condiciones. Estas frecuencias no garantizan nada acerca de lo que ocurrirá en ninguna de las repeticiones individuales concretas. Sin embargo, una forma racional de actuar se basa en la idea de que si sabemos que en la mayoría de las repeticiones futuras se daría uno concreto de entre los resultados posibles, parece más racional predecir que la próxima repetición individual se resolverá precisamente con ese resultado. En general, la confianza puesta en cada uno de los resultados posibles en la próxima realización del evento será proporcional a la frecuencia de repeticiones de cada una de esas alternativas en el futuro. La asignación de números (o probabilidades) a esos grados de confianza depositados en la obtención de cada resultado es la clave del concepto de probabilidad.

> La probabilidad de un suceso es un número que cuantifica en términos relativos las opciones de verificación de ese suceso.

Las opciones se cuantifican en términos relativos para que las probabilidades sean magnitudes comparables. Se podría hacer con cualquier máximo arbitrario, pero desde sus orígenes se ha hecho en tantos por uno. Es decir, un suceso sin opción alguna tendría una probabilidad igual a 0, mientras que un suceso con todas las opciones tendría una probabilidad igual a 1. Cualquier suceso con un número de opciones intermedio entre esos dos tendrá como probabilidad asociada un número intermedio cuya magnitud represente cuantitativamente sus opciones de verificación. Los sucesos posibles, pero raros, tendrán valores de probabi-

lidad cercanos a 0; los que son casi seguros, aunque no del todo, tendrán valores de probabilidad cercanos a 1. No obstante, a veces se utilizan vulgarizadamente números expresados en tantos por cien para indicar probabilidades. Aunque el sentido que se pretende dar es el mismo, estrictamente hablando esos valores no son probabilidades, sino porcentajes de posibilidades que expresan cuántas de cada cien veces se espera que ocurra el suceso.

Lamentablemente, es muy difícil proponer una definición de probabilidad en la que no intervenga el propio término que se pretende definir o un sinónimo suyo, como «posibilidades» u «opciones». Por ello, las definiciones de probabilidad se suelen reducir a la descripción operativa de los enfoques utilizados para su determinación. Hay varios enfoques alternativos, de los que nosotros vamos a exponer dos: el enfoque clásico o *a priori* y el enfoque frecuencialista o *a posteriori*. Por razones obvias, excluimos el enfoque formal, matemático o axiomático (Amón, 1996).

9.3.1. Enfoque clásico o a priori

La aplicación del *enfoque clásico* o *a priori* exige la aceptación del llamado principio de indiferencia, que establece que, al realizar un experimento aleatorio, todos los elementos del espacio muestral tienen las mismas opciones de ser verificados.

Desde el enfoque clásico o a priori, que exige asumir el principio de indiferencia, se define la probabilidad de un suceso como la frecuencia relativa de ese suceso en el espacio muestral. Es decir, si en nuestra colonia de titís hay seis hembras y el procedimiento de extracción nos permite asumir el principio de indiferencia, entonces la probabilidad del suceso B, definido anteriormente sobre este espacio muestral y que representaremos por P(B), será igual a:

$$P(B) = \frac{n_B}{N} = \frac{6}{10} = 0{,}60$$

Dicho en otras palabras, y siguiendo los términos de Laplace (1812), desde este enfoque la probabilidad de un suceso es igual al cociente entre el número de casos favorables y posibles:

$$\text{Probabilidad de un suceso} = \frac{\text{Número de casos favorables}}{\text{Número de casos posibles}}$$

Esta forma de definir la probabilidad se puede entender como un reparto equitativo de la «masa de opciones» entre los elementos del espacio muestral. Al repartirla entre N elementos, a cada uno le corresponde una proporción de $1/N$. La probabilidad de un suceso no será más que la suma de las opciones de los elementos que lo integran; es decir, $1/N$ sumado n_B veces, que es igual a n_B/N.

En muchas ocasiones existe una dificultad práctica para computar el número de casos favorables y el número de casos posibles. Los procedimientos habitual-

mente utilizados para determinar estas cantidades reciben los nombres de técnicas de contar o combinatoria (Amón, 1996).

De la forma de definir la probabilidad desde este enfoque se deducen algunas consecuencias y propiedades:

a) La probabilidad de un suceso es un valor comprendido entre 0 y 1:

$$0 \leq P(A) \leq 1$$

b) Un suceso que no contiene ningún suceso elemental tiene una probabilidad igual a 0; por ello recibe el nombre de *suceso imposible*. Si A es un suceso de este tipo, entonces:

$$P(A) = \frac{0}{N} = 0$$

c) Un suceso que contiene todos los sucesos elementales del espacio muestral ($n_A = N$) tiene una probabilidad igual a 1; por ello recibe el nombre de *suceso seguro*. Si A es un suceso seguro, entonces:

$$P(A) = \frac{n_A}{N} = \frac{N}{N} = 1$$

d) La suma de las probabilidades de un suceso y su complementario es igual a 1. Es decir:

$$P(A) + P(A') = \frac{n_A}{N} + \frac{n_{A'}}{N} = \frac{n_A}{N} + \frac{N - n_A}{N} = \frac{n_A - n_A + N}{N} = 1$$

y como consecuencia:

$$P(A') = 1 - P(A)$$

9.3.2. Enfoque frecuencialista o a posteriori

No siempre se puede aplicar el enfoque clásico. Ejemplo de ello podrían ser las opciones varón/mujer para el bebé del próximo parto en una clínica, el número de preguntas que acertará un estudiante en un examen tipo test, la autoubicación política de un encuestado, etc. El procedimiento para determinar las probabilidades asociadas a estos eventos no puede ser el que prescribe el enfoque clásico, porque se desconocen n_A y/o N, o no se puede asumir el principio de indiferencia.

En estos casos se puede aplicar el *enfoque frecuencialista* o *a posteriori,* según el cual la probabilidad se determinaría mediante una operación ideal de repetición sistemática del experimento aleatorio y de conteo del número de veces que se

verifican los sucesos. Las opciones de verificación de un suceso se manifestarían en el número de veces que se repite éste al realizar una y otra vez el experimento aleatorio. Sin embargo, para estar seguros de que las veces que se verifica el suceso representan proporcionalmente su probabilidad, el número de veces que se realiza el experimento debe ser infinitamente grande. Desde el enfoque frecuencialista, la probabilidad de un suceso se define como el límite de la frecuencia relativa de apariciones de ese suceso cuando el número de repeticiones del experimento aleatorio tiende a infinito. Si representamos aquí por n_A al número de veces que se produce un resultado con el que se verifica el suceso A al repetir el experimento N veces, la probabilidad de ese suceso se define como el número al que tiende el cociente entre n_A y N cuando N tiende a infinito:

$$P(A) = \lim_{N \to \infty} \frac{n_A}{N}$$

Adviértase que la probabilidad nada dice sobre los hechos individuales, sino sobre las opciones «a la larga». No obstante, la convergencia entre el cociente n_A/N y el valor de probabilidad, P(A), es bastante rápida. Por ejemplo, se puede demostrar que es muy improbable que la proporción de caras al lanzar una moneda imparcial 100 veces quede fuera del intervalo 0,40-0,60, mientras que si el número de lanzamientos asciende a 1.000 es casi seguro que estaría entre 0,47 y 0,53; si se lanza 10.000 veces, ese cociente difícilmente será un valor fuera del intervalo 0,49-0,51. En esta idea se basa el teorema de Bernouilli (1713), que exponemos a continuación:

> Si la probabilidad de un suceso A es P(A) y se realizan N ensayos, independientemente y bajo las mismas condiciones, entonces la probabilidad de que la frecuencia relativa de aparición de A difiera de P(A) en una cantidad arbitrariamente pequeña, ε (siendo $\varepsilon > 0$), se acerca a cero a medida que crece el número de ensayos.
>
> Es decir:
>
> $$P\left(\left|\frac{n_A}{N} - P(A)\right| \geq \varepsilon\right) \to 0$$
>
> $$\text{Si} \quad N \to \infty$$

La diferencia fundamental entre el enfoque frecuencialista y el enfoque clásico es que mientras en éste N es el tamaño del espacio muestral, en aquel representa el número de repeticiones del experimento aleatorio. De la definición del enfoque frecuencialista se deducen las mismas consecuencias y propiedades que exponíamos en conexión con el enfoque clásico.

Hay que hacer notar que la probabilidad de un suceso es un número concreto. Otra cuestión diferente es la de cuál es el procedimiento más apropiado para cal-

cularla o para definirla operativamente. Si: *a*) se tiene un conocimiento exhaustivo del espacio muestral, y *b*) se puede asumir el principio de indiferencia, entonces la probabilidad de un suceso se puede obtener mediante la definición del enfoque clásico. Cuando no se cumple alguna de esas dos condiciones la probabilidad se define según el enfoque a posteriori, pero no se puede calcular con exactitud, puesto que ningún experimento aleatorio se puede repetir un número infinito de veces. Se pueden hacer estimaciones más o menos precisas repitiendo el experimento un número muy grande de veces, pero nunca se podrá deducir con total precisión la probabilidad del suceso a partir del cociente entre las verificaciones del suceso y el número (finito) de repeticiones del experimento.

En resumen, la probabilidad es un valor ideal relacionado con las expectativas «a la larga»; sus leyes sólo se cumplen cuando el número de repeticiones tiende a infinito. En realidad, el enfoque frecuencialista es más universal que el clásico, al que en alguna medida incluye. Si los elementos de un espacio muestral tienen las mismas opciones de ser observados (principio de indiferencia), entonces las probabilidades se pueden obtener dividiendo los casos favorables por los posibles, pero tanto si se cumple esta condición como si no se cumple, las probabilidades de los eventos siempre se pueden definir desde el enfoque frecuencialista.

Como hemos visto, la probabilidad de un evento definido sobre un experimento aleatorio es un número que indica la frecuencia relativa esperada de ocurrencias de ese evento cuando el número de repeticiones del experimento aleatorio, realizado en condiciones idénticas, tiende a infinito. No obstante, estos valores no garantizan nada acerca de los resultados que se observarán en repeticiones concretas del experimento. Así, si se extrae al azar una bola de una urna con 90 bolas negras y 10 blancas es más prudente apostar que saldrá una negra. Aunque no se puede garantizar que la próxima bola lo sea, lo que se quiere indicar con el número que se asocia al suceso «bola negra», que llamamos probabilidad, es que si se repitiese la extracción un número infinito de veces, en 90 de cada 100 saldría negra. Por tanto, es más razonable confiar en que la próxima extracción dará el resultado que en un futuro indefinidamente largo sería mayoritario, y no en otro que en ese futuro sería minoritario.

9.4. PROBABILIDAD CONDICIONAL

Volvamos al ejemplo de la colonia de titís y determinemos la probabilidad del suceso A: «El individuo extraído nació en libertad». Dado que tres de los diez individuos cumplen la condición especificada, según el enfoque clásico la probabilidad de este suceso será:

$$\mathrm{P(A)} = \frac{n_A}{N} = \frac{3}{10} = 0{,}30$$

Pero supongamos ahora que las bolas con las que hacemos la extracción aleatoria son de dos colores: blancas las de las hembras y negras las de los machos. Extraemos una bola al azar y resulta ser de color blanco. Sabiendo, por tanto,

que el individuo seleccionado es hembra, ¿cuánto vale ahora la probabilidad de que haya nacido en libertad? Está claro que la probabilidad del suceso podría ser diferente. Si sabemos que la bola es blanca, está claro que la incertidumbre ya no abarca a los diez elementos del espacio muestral, sino a los seis elementos que son hembras. Es decir, el espacio muestral ha quedado reducido a seis sucesos elementales y los casos favorables serán sólo aquellos que, siendo mujeres, hayan nacido en libertad. En la tabla siguiente aparecen las frecuencias absolutas de verificación y no verificación de los sucesos A y B del ejemplo de la colonia de titís.

		B	**B′**	
		Hembra	Macho	
A	Libertad	1	2	3
A′	Cautividad	5	2	7
		6	4	10

Como se puede apreciar, en el grupo total hay tres titís de diez que nacieron en libertad, pero dentro de las hembras hay sólo una de las seis que cumple esa condición. La probabilidad de ese suceso, tal y como lo hemos descrito, será:

$$\frac{\text{Número de casos favorables}}{\text{Número de casos posibles}} = \frac{1}{6} = 0{,}167$$

De lo que estamos hablando aquí es de la probabilidad de verificación del suceso A, sabiendo que se verifica el suceso B. Éste es un caso especial de probabilidad que se llama *probabilidad condicional;* se representa por $P(A|B)$ y se lee «probabilidad de A, dado B». Es fácil demostrar que esa probabilidad se puede obtener a partir de las probabilidades de la intersección de esos sucesos y de la condición impuesta. En términos generales:

$$P(A|B) = \frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/N}{n_B/N} = \frac{P(A\cap B)}{P(B)}$$

Podemos, por tanto, definir la probabilidad condicional de la siguiente forma:

La probabilidad de un suceso, A, dada la verificación de otro suceso, B, se llama probabilidad condicional de A dado B, y es igual a la probabilidad de su intersección dividida por la probabilidad de la condición.

Es decir:

$$P(A|B) = \frac{P(A\cap B)}{P(B)} \qquad [9.1]$$

9.5. TEOREMAS BÁSICOS

Hay dos teoremas fundamentales cuya aplicación es constante en el trabajo con experimentos aleatorios y que ayudarán a entender mejor el propio concepto de probabilidad. El primero se refiere a la probabilidad de la unión de dos sucesos, $P(A \cup B)$ y el segundo a la condición de independencia.

9.5.1. Teorema de la adición

Empezaremos por definirlo y luego lo explicaremos e ilustraremos con algún ejemplo.

Según el *teorema de la adición*, la probabilidad de la unión de dos sucesos es igual a la suma de sus probabilidades menos la probabilidad de su intersección:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad [9.2]$$

A veces los recién llegados a la probabilidad se sorprenden al observar la formulación de este teorema, especialmente el hecho de que se reste la probabilidad de la intersección. Merece la pena detenerse brevemente en este punto. Supongamos que disponemos de una urna con diez bolas numeradas del 1 al 10 y que extraemos una de ellas mediante un procedimiento aleatorio que garantice el principio de indiferencia. Definimos dos sucesos mediante las siguientes reglas, A: «La bola extraída tiene un número par», y B: «La bola extraída tiene un número múltiplo de 3». La probabilidad de la unión de esos sucesos se podría obtener, en principio, aplicando directamente la definición del enfoque clásico: dividiendo el número de casos favorables por el número de casos posibles. El número de bolas que cumplen al menos una de esas condiciones es igual a 7, mientras que tres bolas (las que tienen los números 1, 5 y 7) no cumplen ninguna de ellas. En consecuencia, la probabilidad de la unión de esos sucesos sería:

$$P(A \cup B) = \frac{7}{10} = 0{,}70$$

Pero, por otra parte, las probabilidades simples de los sucesos A y B son:

$$P(A) = \frac{5}{10} = 0{,}50 \qquad P(B) = \frac{3}{10} = 0{,}30$$

Si la probabilidad de la unión de dos sucesos se obtuviese mediante la suma simple de sus probabilidades, sería igual a 0,80 en lugar de 0,70. Esta aparente

paradoja se resuelve fácilmente al observar la figura 9.1, en la que se aprecia que la simple suma supone contabilizar dos veces aquellos sucesos elementales que verifican simultáneamente las condiciones que definen los dos sucesos (en este caso, el elemento 6). Hay que restar el número de elementos de la intersección para que cada uno de ellos sea computado sólo una vez en el cálculo de la probabilidad. Una consecuencia de esto es que la probabilidad de la unión de dos sucesos incompatibles queda reducida a la suma de sus probabilidades, dado que su intersección es un suceso imposible.

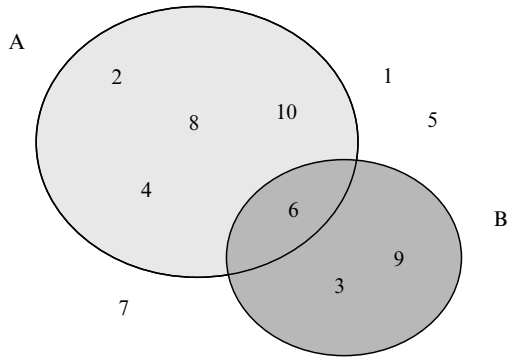


Figura 9.1.—Espacio muestral del experimento aleatorio realizado al extraer una de diez bolas numeradas del 1 al 10. El espacio A incluye los elementos que verifican el suceso «la bola extraída tiene un número par», mientras que el B incluye los elementos que verifican el suceso «la bola extraída tiene un número múltiplo de 3». En su intersección aparece el número 6, que es el único que verifica ambas condiciones simultáneamente.

9.5.2. Teorema del producto

El segundo teorema se refiere a la probabilidad de la intersección de dos sucesos en un caso especial, de alto interés para nosotros y que exige la definición previa del concepto de *independencia.* Observemos la siguiente tabla de doble entrada, definida sobre nuestro experimento aleatorio de extracción de un miembro de la colonia de titís, con los sucesos A: «Nacimiento en la colonia», y B: «Hembra»:

		B	**B′**	
		Hembra	Macho	
A	Colonia	3	2	5
A′	Otros	3	2	5
		6	4	10

Ahora las probabilidades de los sucesos, de su intersección y la probabilidad condicional serán:

$$P(A) = 5/10 = 0{,}50$$

$$P(B) = 6/10 = 0{,}60$$

$$P(A \cap B) = 3/10 = 0{,}30$$

$$P(A|B) = P(A \cap B)/P(B) = 0{,}30/0{,}60 = 0{,}50$$

En este caso se da la coincidencia de que la probabilidad de A dado B es igual a la probabilidad simple de A (0,50). Eso significa que la probabilidad del suceso A no se ve alterada por el hecho de que se verifique B. La razón es que la frecuencia relativa de los elementos que cumplen la condición de A en el espacio muestral total es la misma que la frecuencia relativa de los que la cumplen en el espacio muestral restringido, formado sólo por los que cumplen la condición definida en B. Cuando se cumple esta igualdad entre la probabilidad simple de un suceso y la condicional de éste con respecto a otro suceso, se dice que los sucesos son independientes. Dada la importancia de este concepto, lo resaltamos en el siguiente recuadro:

Dos sucesos, A y B, son *independientes* si la verificación de uno no altera la probabilidad del otro. Es decir,

Si $P(A|B) = P(A)$

entonces, A y B son sucesos independientes

Por otro lado, si A es independiente de B, entonces B es independiente de A. Para demostrarlo basta señalar que $P(A \cap B) = P(B \cap A)$ y sustituir en [9.1]:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A/B)} = \frac{P(A \cap B)}{P(A \cap B)/P(B)} = P(B)$$

La consecuencia fundamental de esto es que cuando dos sucesos son independientes, la probabilidad condicional se puede reemplazar por la probabilidad simple, y al sustituir en la fórmula [9.1] se obtiene la expresión que define el segundo teorema:

Según el *teorema del producto*, la probabilidad de verificación simultánea de dos sucesos independientes es igual al producto de sus respectivas probabilidades simples. Es decir, si A y B son sucesos independientes, entonces:

$$P(A \cap B) = P(A) \cdot P(B) \qquad [9.3]$$

El concepto de independencia entre sucesos se puede aplicar también a sucesos de diferentes experimentos aleatorios, e incluso se puede extender a los experimentos completos. Supongamos que lanzamos una moneda y luego tiramos un dado. Parece claro que las probabilidades asociadas a cada resultado del dado no se verán alteradas por el resultado del lanzamiento de la moneda. Cuando la condición que define la independencia de sucesos se cumple para cualesquiera dos sucesos definidos sobre esos experimentos aleatorios, se dice que los experimentos (y no sólo los sucesos) son independientes. Por tanto:

> Dos experimentos aleatorios son independientes si se cumple la condición de independencia de sucesos para cualquier par de sucesos, A y B, definidos sobre sus espacios muestrales respectivos.

Un caso particular de aplicación muy frecuente es el de la repetición de un mismo experimento aleatorio de forma tal que las probabilidades asociadas a cada resultado no dependan de los resultados obtenidos previamente. Por ejemplo, si lanzamos una moneda varias veces, la probabilidad de cara o cruz en cada lanzamiento es independiente de lo que haya salido en los lanzamientos anteriores. Si definimos algún suceso sobre nuestra experiencia de extracción de un miembro de la colonia de titís, las probabilidades asociadas a esos sucesos al hacer varias extracciones sucesivas no se verán alteradas si después de cada extracción reponemos de nuevo al individuo extraído. Por tanto, el concepto de independencia se puede extender a más de dos experimentos aleatorios, y las probabilidades asociadas se obtendrían extendiendo la fórmula [9.3] de la siguiente forma: si A_1, A_2, ..., A_k son sucesos definidos, respectivamente, sobre experimentos independientes, entonces:

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_k) \qquad [9.4]$$

PROBLEMAS Y EJERCICIOS

1. Se aplicó un test de capacidad lógica y una prueba de matemáticas a una muestra de 400 estudiantes de secundaria. Se consideró que un estudiante tenía capacidad lógica alta si en el test tenía un resultado superior a 20 puntos. De los 400 estudiantes, 100 tenían capacidad lógica alta; además, del total de la muestra, 120 superaron la prueba de matemáticas, de los cuales 40 tenían capacidad alta. Se definen los sucesos A: *Tener capacidad lógica alta*, y B: *Superar la prueba de matemáticas*. Si se extrae un estudiante al azar, determine:

a) Si los sucesos A y B son independientes.
b) La probabilidad de que el estudiante tenga capacidad lógica alta y no haya superado la prueba.
c) Sabiendo que el estudiante tiene capacidad lógica alta, la probabilidad de que no haya superado la prueba.

2. En una población de estudiantes de bachillerato, la probabilidad de que deseen cursar estudios de psicología (suceso A) es 0,40 y la probabilidad de que opten por estudios de educación (suceso B) es 0,25. Sabiendo que la probabilidad de que deseen cursar ambos es 0,10, calcule la probabilidad de extraer al azar un estudiante que desee cursar:

a) Al menos uno de los dos estudios.
b) *Sólo* estudios de psicología.
c) *Sólo* uno de los dos.

3. En el barco de pesca de altura *Costa del Azar,* la asignación de horarios de comida entre la tripulación se realiza siguiendo un procedimiento aleatorio. Se lanza una moneda y un dado; si sale cara (suceso A) y número par (suceso B) se asigna el primer turno de comida; si sale cruz y número impar se asigna el segundo turno; en otro caso, se asigna el tercer turno. ¿Cuáles son las probabilidades de asignación de cada turno?

4. En un centro psiquiátrico hay un total de 200 residentes, de los cuales 80 sufren un trastorno depresivo mayor y el resto otros cuadros clínicos. Se seleccionan al azar dos historiales clínicos. Calcule las siguientes probabilidades, tanto si se hacen con reposición como sin reposición:

a) La de que ambos se correspondan a residentes con depresión mayor.
b) La de que ambos se correspondan a residentes con otros trastornos.
c) La de que ambos se correspondan a residentes con el mismo trastorno.
d) La de que al menos uno se corresponda a residentes con depresión mayor.

5. Se ha realizado una encuesta sobre las preferencias que tienen los habitantes de la Comunidad de Madrid sobre el lugar de vacaciones y el mes de las mismas (julio o agosto). Con respecto al lugar, se observó que: el 60 por 100 de la muestra

prefiere un sitio de costa nacional (opción A); el 20 por 100 una casa rural del interior de España (opción B); el 15 por 100 salir al extranjero (opción C); y el resto quedarse en su residencia habitual (opción D). Sobre el mes: de aquellos que se decantaron por la opción A, la cuarta parte prefiere el mes de julio; de los que respondieron la opción B, la quinta parte eligió el mes de julio; de los que eligieron la opción C, la tercera parte eligió el mes de julio; y de los que eligieron la opción D, la mitad eligió julio. Si se extrae al azar a uno de los encuestados, calcule:

a) La probabilidad de que haya elegido la opción B o que prefiera tomar sus vacaciones en agosto.
b) Sabiendo que prefiere irse de vacaciones en julio, la probabilidad de que haya elegido la opción D.
c) Sabiendo que ha elegido la opción D, la probabilidad de que prefiera irse en julio.
d) Sabiendo que ha elegido la opción C, la probabilidad de que prefiera irse en agosto.
e) La probabilidad de que prefiera irse en julio o elija la opción A.

6. El cuestionario AUDIT, que evalúa si existen problemas de alcoholemia, detecta al 90 por 100 de personas que sí poseen dicho problema, mientras que en los casos en los que las personas no sufren dicho problema el AUDIT da respuesta positiva en el 20 por 100. Sabiendo que la probabilidad de encontrar a una persona con problemas de alcoholemia y que dé positivo en el AUDIT es igual a 0,12, calcule:

a) La probabilidad de extraer al azar a una persona con problemas de alcoholemia.
b) La probabilidad de que el AUDIT ofrezca una respuesta negativa, sabiendo que una persona no sufre de problemas de alcoholemia.

7. Se está realizando un estudio sobre empleo en la población europea. Dos de los criterios utilizados han sido tener estudios superiores y estar más de dos años sin empleo. Del total de la población, el 60 por 100 tienen estudios superiores (suceso A); además, del mismo total, el 20 por 100 han estado más de dos años sin empleo (suceso B). Sabiendo que el 65 por 100 de la población satisface al menos uno de los dos criterios, diga cuál es la probabilidad de extraer al azar una persona que satisfaga simultáneamente ambos criterios. ¿Son independientes los sucesos *tener estudios superiores* y *estar más de dos años sin empleo*?

8. Siguiendo con el problema anterior, y asumiendo que es $P(A \cap B)$ diferente del valor señalado en ese ejercicio y desconocida:

a) Si los sucesos A y B fueran incompatibles o excluyentes, ¿cuál es la probabilidad de extraer al azar una persona que satisfaga al menos uno de esos criterios?
b) Si los sucesos A y B fueran independientes, ¿cuál es la probabilidad de extraer al azar una persona que satisfaga al menos uno de esos criterios?

c) Si los sucesos A y B fueran independientes, ¿cuál es la probabilidad de que una persona tenga estudios superiores, sabiendo que ha estado más de dos años sin empleo?

9. ¿Cuál es la probabilidad de que la suma de los resultados de dos dados sea par, sabiendo que no se ha repetido el mismo número?

10. Se está realizando un estudio social en un grupo de bachillerato. Del total de los estudiantes, 20 son españoles, 10 son de origen latinoamericano, 5 orientales y 5 de otros países. Además, 18 son mujeres, de las cuales 10 son de origen español, 2 de origen latinoamericano, 4 orientales y 2 de otros países. Si se extrae al azar un estudiante, calcule:

a) La probabilidad de que sea mujer o sea de origen latino.
b) La probabilidad de que sea mujer sabiendo que es de origen oriental.

11. Se está realizando una investigación que trata de estudiar si existe independencia entre el nivel formativo y el nivel laboral alcanzado en la población europea. Se observó que la probabilidad de extraer al azar una persona con titulación universitaria es igual a 0,70 y la probabilidad de tener una titulación universitaria y ocupar un puesto alto es igual a 0,30. En el caso de que existiera independencia entre nivel formativo y puesto laboral, ¿qué valor de probabilidad cabría pronosticar al suceso *ocupar un puesto alto*?

12. Un equipo de baloncesto está disputando un concurso de tiros libres. Un primer jugador tiene una probabilidad de 0,90 de encestar el balón, mientras que la probabilidad de un segundo jugador es de 0,80. Si cada uno de los jugadores hace un solo lanzamiento, calcule las siguientes probabilidades:

a) La de que encesten ambos.
b) La de que enceste al menos uno de los dos.

13. La probabilidad de resolver correctamente dos tareas A y B de atención visual es 0,3249. Halle las probabilidades de resolver cada una de las tareas por separado, sabiendo que son equiprobables y que la resolución correcta de las tareas A y B es independiente.

14. En una población de 400 estudiantes de psicología, hay 150 varones y 250 mujeres. Si extraemos dos estudiantes al azar, ¿se modifica la probabilidad de que los estudiantes extraídos sean mujeres al hacerlo sin reposición, en lugar de hacerlo con reposición?

15. El gabinete psicológico de un instituto madrileño ha realizado un estudio de orientación vocacional entre los estudiantes del último curso de bachillerato, aconsejando a cada estudiante el tipo de estudios universitarios que mejor se ajusta a su perfil. Algunos estudiantes siguieron los consejos y otros no. En concreto, las frecuencias de orientación y de seguimiento obtenidas fueron las siguientes:

		Carrera elegida					
		D	*E*	*F*	*C*	*M*	*P*
Carrera aconsejada	(*D*) Derecho	10	4	0	0	1	0
	(*E*) Económicas	8	20	4	4	0	2
	(*F*) Filosofía y letras	0	10	15	3	1	1
	(*C*) Ciencias	1	8	5	15	2	5
	(*M*) Medicina	2	6	4	10	13	8
	(*P*) Psicología	1	2	0	4	9	22

Si extraemos a uno de los estudiantes al azar, diga cuánto valen las siguientes probabilidades:

a) La de que al estudiante extraído se le haya aconsejado Económicas.
b) La de que el estudiante extraído haya elegido Medicina.
c) La de que el estudiante extraído haya seguido el consejo recibido.
d) La de que el estudiante extraído haya elegido Filosofía y Letras, sabiendo que se le aconsejó Económicas.
e) La de que esté estudiando Medicina o Psicología.
f) La de que haya seguido el consejo recibido, sabiendo que es estudiante de Derecho.
g) La de que sea un estudiante de Medicina y recibiera el consejo de estudiar esa carrera.

16. En un estudio clínico realizado en una población de pacientes con síntomas de ansiedad, se ha encontrado que la probabilidad de que se den trastornos en el sueño (A) es 0,40, la probabilidad de que se den trastornos de tipo depresivos (B) es 0,60 y la probabilidad de que se den ambos 0,30. Si extraemos un paciente de dicha población al azar, ¿cuál es la probabilidad de que tenga al menos uno de los trastornos?

17. Los estudios sobre toma de decisiones indican que el riesgo asumido es distinto si las decisiones se toman en solitario que si se toman en grupo. En concreto, parece que las decisiones adoptadas en grupo tienden a ser más arriesgadas. Tras varios años de pasar tareas de decisión en experimentos, se sabe que en las tareas de lápiz y papel resueltas en solitario el 40 por 100 de los participantes adoptan la alternativa más arriesgada, mientras que en los casos en los que se pasan las tareas a grupos de cuatro participantes, en la mitad de ellos se adoptó la decisión menos arriesgada. ¿Apoyan estos datos la tesis de la dependencia entre la forma de adoptar las decisiones y el nivel de riesgo de las alternativas elegidas?

18. Una empresa consultora ha realizado una encuesta en la provincia de Valladolid, encontrando que el 60 por 100 de los encuestados siguen las noticias por

la televisión (T), el 30 por 100 las leen en el periódico (P) y el 10 por 100 las oye en la radio (R). Teniendo en cuenta que el 25 por 100 lee las noticias en el periódico, las ve en la televisión y las escucha la radio y que el 80 por 100 de los que leen las noticias en el periódico también las ve en la televisión, si extraemos un sujeto al azar:

a) ¿Cuál es la probabilidad de que lea las noticias en el periódico, las vea en la televisión y las escuche en la radio?
b) ¿Cuál es la probabilidad de que vea las noticias por la televisión y también las lea en el periódico?

19. Halle la probabilidad de que la suma de los resultados de lanzar dos dados sea igual a 10.

20. Sabemos que los tests A, B, C y D permiten detectar disfunciones cerebrales en los pacientes que han sufrido un traumatismo con probabilidades de 0,95, 0,80, 0,75 y 0,50, respectivamente, y que la probabilidad de que se detecte la disfunción con cualquiera de esos tests es independiente de la de que se detecte con los restantes tests. Según lo anterior, y asumiendo que basta que un test dé positivo para considerar que el problema se ha detectado, responda a las siguientes cuestiones:

a) ¿Cuál es la probabilidad de detectar el problema en un paciente si se le administran los tests A y B?
b) ¿Cuál es la probabilidad de que no se le detecte si se le pasan los tests B y C?
c) Si consideramos como fiable una detección sólo si el paciente da positivo en dos tests, ¿cuál es la probabilidad de que detectemos la disfunción si pasamos los tests C y D?
d) Si consideramos que se ha detectado la disfunción cuando al menos uno de los tests da positivo, ¿cuál es la máxima probabilidad de detección que podríamos conseguir con esos tests y cuál sería el procedimiento para ello?

21. En una facultad, que cuenta con 2.000 estudiantes matriculados, sabemos que 400 tienen el grupo sanguíneo O, 700 el A, 700 el B y 200 el AB. Sabiendo que las reglas de donación son las que se recogen en la tabla siguiente (donde el asterisco indica que el grupo de la fila es donante potencial del grupo de la columna):

	O	A	B	AB
O	*	*	*	*
A		*		*
B			*	*
AB				*

A continuación, responda a las cuestiones siguientes:

a) Si un donante tuviera que hacer una transfusión a una persona del grupo A, ¿cuál es la probabilidad de que un estudiante escogido al azar sea donante potencial?
b) ¿Y si la transfusión es a una persona del grupo O?
c) ¿Y si es a una persona del grupo AB?
d) Si el donante es del grupo B, ¿cuál es la probabilidad de que pueda donar sangre a un estudiante de la facultad elegido al azar?
e) ¿Y si es del grupo AB?

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. La tabla de frecuencias conjuntas definida por los sucesos es la siguiente:

	Supera la prueba (B)	No supera la prueba (B′)	
Alta (A)	40	60	100
Baja (A′)	80	220	300
	120	280	400

a) Los sucesos A y B no son independientes.
b) 0,15.
c) 0,60.

2. *a*) 0,55.
b) 0,30.
c) 0,45.

3. La probabilidad de estar en el primer turno es 0,25; la del segundo 0,25; la del tercero es 0,50.

4. *a*) Con reposición, 0,16; sin reposición, 0,1588.
b) Con reposición, 0,36, y sin reposición 0,3588.
c) Con reposición, 0,52, y sin reposición 0,5176.
d) Con reposición, 0,64, y sin reposición, 0,6412.

5. *a*) 0,775.
b) 0,0943.
c) 0,50.

d) 0,667.
e) 0,715.

6. *a*) 0,133.
b) 0,80.

7. La probabilidad es 0,15. Los sucesos A y B no son independientes.

8. *a*) 0,80.
b) 0,68.
c) 0,60.

9. La probabilidad es 0,40.

10. *a*) 0,65
b) 0,80.

11. El valor pronosticado es 0,429.

12. *a*) 0,72
b) 0,98.

13. Como se cumple el teorema del producto y los sucesos son equiprobables, $P(A) = P(B) = 0{,}57$.

14. Sí; con reposición es 0,3906 y sin reposición es 0,3900.

15. *a*) 0,19.
b) 0,13.
c) 0,475.
d) 0,105.
e) 0,32.
f) 0,455.
g) 0,065.

16. 0,70.

17. Sí; no son independientes, por el teorema del producto.

18. *a*) 0,25.
b) 0,24.

19. 0,0833.

20. *a*) 0,99.
b) 0,05.
c) 0,375.
d) La máxima probabilidad de detección se consigue pasando los cuatro tests. Con ello, la probabilidad de detección es 0,99875.

21. *a*) 0,55.
b) 0,20.
c) 1.
d) 0,45.
e) 0,10.

10 Variables aleatorias

10.1. INTRODUCCIÓN

En el capítulo anterior quedó establecido que los espacios muestrales son conjuntos de sucesos elementales, y que éstos a veces, pero no siempre, son números. Es muy útil representar a los sucesos elementales por números, mediante lo que se denominan variables aleatorias. En este capítulo abordaremos su estudio, primero definiéndolas y luego exponiendo sus características, distinguiendo entre variables aleatorias discretas y continuas. Concluiremos introduciendo el concepto de distribución de probabilidad y retomando el de muestreo, al que ya dedicamos unas líneas en el capítulo 1. No obstante, antes de abordar la definición de las variables aleatorias vamos a recordar brevemente el concepto de función.

Una función es cualquier conjunto de pares ordenados de elementos en los cuales no se repite el primer elemento. Así, el siguiente conjunto de pares ordenados es una función:

(España, Europa) (Italia, Europa) (Argentina, América)...

Como veremos a continuación, las variables aleatorias son funciones que cumplen ciertos requisitos.

10.2. DEFINICIÓN Y TIPOS DE VARIABLES ALEATORIAS

Comenzaremos por definir qué es una *variable aleatoria*.

> Una *variable aleatoria* es una función que asocia un número real, y sólo uno, a cada suceso elemental del espacio muestral de un experimento aleatorio.

Siguiendo con el ejemplo de la colonia de titís del capítulo anterior (cuadro 9.1), extraemos a uno de sus miembros al azar y anotamos su edad. La va-

riable se compondría a partir del emparejamiento de cada suceso elemental y la edad de cada miembro de la colonia:

(Montri, 12) (Chulina, 8) ... (Tangue, 28)

Cualquier emparejamiento de este tipo definido sobre un espacio muestral en el que no haya dos pares en los que se repita el mismo primer elemento, en el que todos los sucesos elementales estén incluidos en algún par y en el que el segundo elemento de cada par sea un número real, es una variable aleatoria. Dicho de otra forma, una variable aleatoria es una aplicación X: $\mathbf{E} \rightarrow \mathbb{R}$.

También podríamos definir, sobre el estudio de condicionamiento con adolescentes, la siguiente variable aleatoria, que refleja el número de adolescentes que eligen la caja con el logotipo a la izquierda (figura 10.1):

(III, 3) (IID, 2) (IDI, 2) (DII, 2) (IDD, 1) (DID, 1) (DDI, 1) (DDD, 0)

Representaremos a las variables aleatorias por letras mayúsculas, aunque en muchas ocasiones, para referirnos a un valor cualquiera utilizaremos la letra minúscula y un subíndice que designe a ese valor concreto.

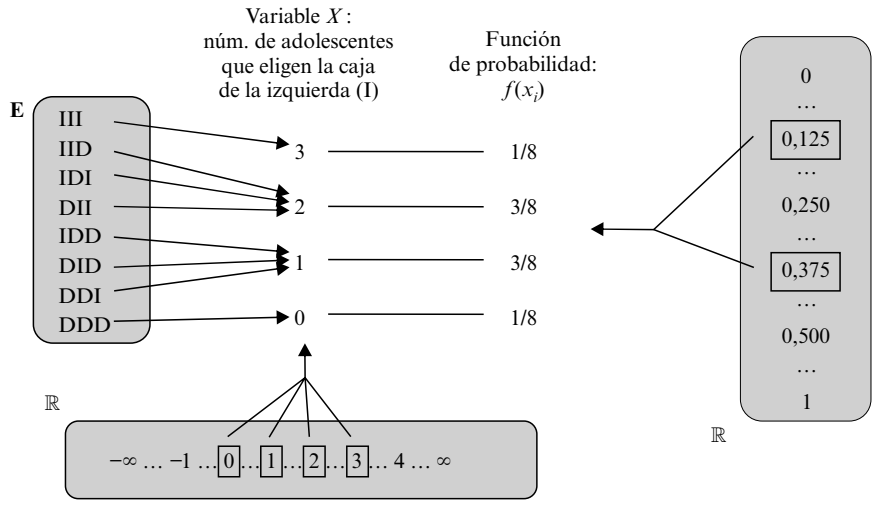


Figura 10.1.—Esquema de la definición de una variable aleatoria.

Al igual que hacíamos con las variables en estadística descriptiva, distinguiremos entre variables aleatorias discretas y variables aleatorias continuas. Las primeras son las que se definen sobre espacios muestrales finitos o infinitos pero numerables, mientras que las segundas son las que se definen sobre espacios muestrales infinitos no numerables. Describiremos sus características en las secciones subsiguientes.

Tanto en las variables aleatorias discretas como en las continuas aparecen ciertos conceptos paralelos a los que expusimos en estadística descriptiva; los

definiremos en analogía con aquéllos. En particular, definiremos primero los análogos a las frecuencias relativas y relativas acumuladas, después los análogos a la media y la varianza, y por último los de la covarianza y la correlación.

10.3. VARIABLES ALEATORIAS DISCRETAS

Ya hemos dicho que las variables aleatorias discretas son aquellas que se definen sobre espacios muestrales finitos o infinitos pero numerables. La variable aleatoria discreta adopta valores tales que se pueden encontrar dos consecutivos entre los cuales no hay valores asumibles por la variable. Ejemplo de ello es el número de adolescentes que eligen la caja con el logotipo a la izquierda, que puede adoptar sólo cuatro valores distintos (0, 1, 2, 3).

10.3.1. Función de probabilidad y función de distribución

Dos conceptos fundamentales relacionados con las variables aleatorias discretas se refieren a las funciones que asocian a cada uno de sus valores las probabilidades de que éstas adopten esos valores y a las de que adopten, como mucho, esos valores. La primera constituye la *función de probabilidad* y la segunda la *función de distribución*. Mientras que la primera se representa por la letra f minúscula, la segunda se representa por la letra F mayúscula. Por ejemplo, si X es una variable aleatoria, entonces la probabilidad de que ésta adopte un valor cualquiera, x_i, la representaremos por $f(x_i)$, y la de que adopte un valor igual o menor que él se representa por $F(x_i)$. Resaltaremos estas definiciones:

Se llama *función de probabilidad* de una variable aleatoria discreta, X, a aquella que asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor; se representa por $f(x_i)$:

$$f(x_i) = \mathrm{P}(X = x_i) \qquad [10.1]$$

Se llama *función de distribución* de una variable aleatoria discreta, X, a aquella que asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor o cualquiera inferior; se representa por $F(x_i)$:

$$F(x_i) = \mathrm{P}(X \leq x_i) \qquad [10.2]$$

Como ejemplo de ello retomaremos la variable referida al número de adolescentes que eligen la caja de la izquierda y que representaremos por X. Se trata de una variable aleatoria discreta sobre la que podemos obtener las funciones de probabilidad y de distribución (véanse las secciones *a* y *b* del cuadro 10.1).

CUADRO 10.1

Ejemplo de funciones de probabilidad y de distribución y de su representación gráfica, así como el valor esperado y la varianza de la variable aleatoria discreta X = «número de adolescentes que eligen la caja con el logotipo a la izquierda»

Utilizaremos la variable aleatoria X = «número de adolescentes que eligen la caja de la izquierda», definida sobre el estudio de condicionamiento.

a) Correspondencia entre los elementos del espacio muestral y los valores numéricos:

X: (DDD, 0); (DDI, 1); (DID, 1); (IDD, 1); (IID, 2); (IDI, 2); (DII, 2); (III, 3)

b) Funciones de probabilidad y de distribución de la variable:

X	$f(x_i)$	$F(x_i)$
0	0,125	0,125
1	0,375	0,500
2	0,375	0,875
3	0,125	1,000
	1,000	

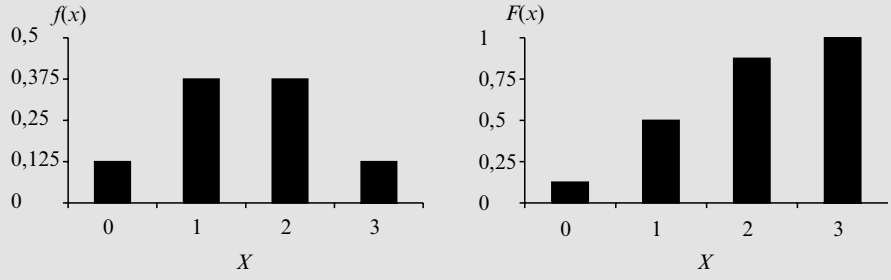
c) Cálculo del valor esperado y de la varianza:

X	$f(x_i)$	$X_i \cdot f(x_i)$	X_i^2	$X_i^2 \cdot f(x_i)$
0	0,125	0	0	0
1	0,375	0,375	1	0,375
2	0,375	0,750	4	1,500
3	0,125	0,375	9	1,125
	1,000	1,500		3,000

$E(X) = 1,5$

$\sigma^2(X) = 3 - 1,5^2 = 0,75$

d) Representación gráfica de las funciones de probabilidad y distribución:



En el cuadro 10.1 se constatan dos cosas, por otra parte obvias: que la suma de las probabilidades asociadas a todos los valores es necesariamente igual a 1, y que la probabilidad acumulada para el valor máximo es también necesariamente igual a 1:

$$\sum f(x_i) = 1 \qquad \text{y} \qquad F(x_{\text{máx}}) = 1$$

Como ya hemos indicado, la función de probabilidad y la función de distribución son el equivalente a los conceptos de frecuencia relativa y frecuencia relativa acumulada en estadística descriptiva. La diferencia es que en estadística descriptiva se trataba del cociente entre cada frecuencia absoluta y el tamaño de la muestra, mientras que aquí se trata de probabilidades.

10.3.2. Valor esperado y varianza

Otras dos características importantes de las variables aleatorias discretas son las análogas a la media y la varianza, aunque en variables aleatorias se llaman *valor esperado* y *varianza* y mantienen algunas diferencias importantes con respecto a aquellas. La media de un grupo de valores es la magnitud general observada, la cual actúa como centro de gravedad de la distribución de frecuencias (capítulo 3). Por el contrario, el valor esperado de una variable aleatoria es un valor ideal igual a la media que se obtendría en caso de que se observase un número infinito de valores de la variable aleatoria. Se representa por $E(X)$, siendo E la inicial de «esperanza matemática», que es un nombre alternativo utilizado a veces en lugar de «valor esperado», o también por la letra griega μ_x. Se define como el sumatorio de los productos de cada valor por su probabilidad. Por su parte, la varianza de una variable aleatoria, $\sigma^2(X)$, es la varianza que se obtendría sobre los valores observados en caso de que el número de observaciones creciera infinitamente.

El *valor esperado* de una variable aleatoria discreta, X, se representa por $E(X)$, o μ, y es igual a la expresión:

$$E(X) = \mu = \sum x_i \cdot f(x_i) \qquad [10.3]$$

La *varianza* de una variable aleatoria discreta, X, se representa por $\sigma^2(X)$, y es igual a la expresión:

$$\sigma^2(X) = \sum x_i^2 \cdot f(x_i) - [E(X)]^2 \qquad [10.4]$$

En la sección *c* de la tabla 10.1 hemos incluido el cálculo del valor esperado y la varianza de la variable que veníamos utilizando como ejemplo. El concepto de valor esperado se entiende mejor cuando se expone en el contexto de los juegos de azar, razón por la que recomendamos al lector detenerse en este punto y leer

el cuadro 10.2. Un caso especial que también merece la pena destacar es el de las variables aleatorias dicotómicas, que describimos en el cuadro 10.3.

CUADRO 10.2

Relación entre valor esperado y juegos de azar

Supongamos una lotería que consiste en lo siguiente: se venden 100 billetes a 2 euros cada uno. Se extraen 12 números al azar de una forma que garantice la equiprobabilidad de extracción de cada uno de los números y se distribuyen los siguientes premios: al primer número extraído se le dan 60 euros, al segundo 30 euros y a los otros 10 se les reintegra los 2 euros que pagaron por el billete. En estas condiciones se puede definir la variable aleatoria X = «euros ganados», que sólo puede adoptar cuatro valores: el que gana el primer premio $(60 - 2) = 58$ euros, el del segundo premio $(30 - 2) = 28$, los de los reintegros $(2 - 2) = 0$ y aquellos a los que no les toca nada, -2 euros. Dado que se puede asumir el principio de indiferencia, las probabilidades de estos resultados posibles se obtienen mediante el enfoque clásico o a priori. Dividimos el número de casos favorables por los cien casos posibles. Con estos valores obtenemos el valor esperado:

X	$f(x_i)$	$x_i \cdot f(x_i)$
−2	0,88	−1,76
0	0,10	0,00
28	0,01	0,28
58	0,01	0,58
	1,00	−0,90

$E(X) = -0{,}90$

Una vez expuesto lo que es el valor esperado en un juego de azar, vamos a detenernos a reflexionar sobre lo que significa en un contexto práctico. Imaginemos que participamos en el juego que se describe en el cuadro 10.2 una y otra vez, siempre en las mismas condiciones y un número muy grande de veces. A la larga, el promedio de euros ganados, o variable X, tendería a estabilizarse en torno a una pérdida de 0,90 euros. Este valor es negativo, porque la suma de los premios no es igual a lo que se recauda con la venta de los billetes. En caso de que se repartiese en premios la misma cantidad que se recauda con la venta de los billetes, el valor esperado sería igual a cero; entonces se diría que es un juego justo. También podríamos preguntarnos cuál debería ser el precio del billete, manteniendo lo demás igual, para que éste fuera un juego justo. Llamando k a ese precio, sustituimos en la fórmula del valor esperado e igualamos a cero:

$$(60 - k) \cdot 0{,}01 + (30 - k) \cdot 0{,}01 + (k - k) \cdot 0{,}10 + (0 - k) \cdot 0{,}88 = 0$$

Despejando, obtenemos $k = 1$. Es fácil comprobar que si los billetes se vendieran a 1 euro, con esa estructura de premios el valor esperado sería 0.

CUADRO 10.3

Valor esperado y varianza de variables aleatorias dicotómicas

Las *variables aleatorias dicotómicas* únicamente pueden adoptar dos valores. Es muy frecuente que se definan en experimentos aleatorios sobre los que se define un suceso, normalmente el cumplimiento de una condición concreta, que puede verificarse o no. Por ejemplo, una pregunta formulada en términos de «a favor/en contra» permite definir la siguiente función:

(A favor, 1) (En contra, 0)

Si aplicamos esta regla a las respuestas de una muestra de encuestados, tendremos una muestra de unos y ceros. Las características de esta variable aleatoria serían las siguientes. Representando por π a la probabilidad de observar el valor 1, es decir:

$$f(1) = P(X = 1) = \pi$$

entonces el valor esperado de la variable será:

$$E(X) = \sum x_i \cdot f(x_i) = 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi$$

y su varianza:

$$\sigma^2(X) = \sum x_i^2 \cdot f(x_i) - [E(X)]^2 = 1^2 \cdot \pi + 0^2 \cdot (1 - \pi) - \pi^2 = \pi \cdot (1 - \pi)$$

El valor esperado y la varianza tienen unas propiedades análogas a las que expusimos con relación a la media y la varianza en la sección de estadística descriptiva (véase el capítulo 3). Recogemos aquí algunas de esas propiedades:

El valor esperado y la varianza de una constante son iguales, respectivamente, a esa misma constante y a cero. Es decir, si a es una constante, entonces:

$$E(a) = a \quad \text{y} \quad \sigma^2(a) = 0$$

Si a los valores de una variable aleatoria se les suma una constante, su valor esperado se ve incrementado en esa constante y su varianza no se altera. Es decir:

Si $\quad Y = X + a$

entonces, $\quad E(Y) = E(X) + a \quad \text{y} \quad \sigma^2(Y) = \sigma^2(X)$

(continuación)

Si a los valores de una variable aleatoria se les multiplica por una constante, su valor esperado se ve multiplicado por esa constante, y su varianza por el cuadrado de la constante. Es decir:

Si $Y = a \cdot X$

entonces, $E(Y) = a \cdot E(X)$ y $\sigma^2(Y) = a^2 \cdot \sigma^2(X)$

Encontraremos útil extender el concepto de esperanza matemática a otras expresiones. En realidad, el valor esperado no es más que una expectativa que, al referirse a los valores naturales, representa al valor esperado de la variable como tal, pero también nos podemos preguntar por la expectativa que tendríamos acerca de los valores de X elevados al cuadrado o acerca de cualquier otra expresión. Por ejemplo, si extrajéramos un número infinito de valores de X y los eleváramos al cuadrado, ¿cuál sería el valor medio que obtendríamos? Toda esperanza matemática se obtiene sumando los productos de cada valor que puede adoptar la expresión de interés multiplicado por su probabilidad. Así, la esperanza del ejemplo que acabamos de citar, que interviene en la definición de la varianza [10.4] es:

$$E(X^2) = \sum x_i^2 \cdot f(x_i)$$

10.3.3. Relación entre dos variables aleatorias discretas

Para trabajar con dos variables aleatorias discretas hace falta definir un nuevo concepto, que se refiere a la probabilidad de que dos variables aleatorias adopten simultáneamente ciertos valores concretos (el análogo a la frecuencia conjunta, n_{ij}, del capítulo 8). Éste se llama *función de probabilidad conjunta*. Podemos definirla y representarla de la siguiente forma:

Se llama *función de probabilidad conjunta* de dos variables aleatorias discretas, X e Y, a aquella que asocia a cada par de valores de las variables, x_i e y_j, a la probabilidad de que, simultáneamente, X adopte el valor x_i e Y adopte el valor y_j. Se representa por $f(x_i,y_j)$,

$$f(x_i, y_j) = P[(X = x_i) \cap (Y = y_j)] \qquad [10.5]$$

Con este nuevo concepto ya podemos abordar los dos índices que reflejan las relaciones lineales entre dos variables aleatorias discretas. Se trata de la *covarianza* y la *correlación* de Pearson, que definimos de la siguiente forma:

Se llama *covarianza* entre dos variables aleatorias discretas, X e Y, y se representa por $\sigma(XY)$, a la expresión:

$$\sigma(XY) = E(XY) - E(X) \cdot E(Y) \qquad [10.6]$$

donde:

$$E(XY) = \sum_i \sum_j x_i \cdot y_j \cdot f(x_i, y_j)$$

Se llama *correlación* de Pearson entre dos variables aleatorias discretas, X e Y, y se representa por $\rho(XY)$, al cociente entre su covarianza y el producto de sus desviaciones típicas. Es decir:

$$\rho(XY) = \frac{\sigma(XY)}{\sigma(X) \cdot \sigma(Y)} \qquad [10.7]$$

En el capítulo 9 habíamos definido la condición de independencia entre dos sucesos como la igualdad entre cada probabilidad simple y la condicionada con respecto al otro, y habíamos extendido este concepto a experimentos aleatorios completos. Ahora lo vamos a extender a las variables aleatorias como totalidad.

Efectivamente, hay veces en que la condición de independencia se verifica para todos y cada uno de los pares de valores posibles de dos variables aleatorias, X e Y. En esos casos, en lugar de hablar de la independencia entre ciertos pares de valores (y, por tanto, en términos de sucesos), se dice que las variables aleatorias son independientes. Por tanto:

Se dice que dos variables discretas, X e Y, son independientes si para todo par de posibles valores de esas variables, X_i e Y_j, se verifica la condición de independencia de sucesos. Es decir, cuando para todo x_i e y_j se cumple:

$$f(x_i|y_j) = f(x_i) \qquad [10.8]$$

Se puede demostrar que cuando dos variables aleatorias son independientes, según la definición que expusimos unas líneas más arriba, entonces necesariamente su covarianza y su correlación son nulas. Por el contrario, encontrar una correlación igual a cero no implica que las variables sean independientes, sino únicamente que no hay relación lineal, ya que podría haber otros tipos de relación entre ellas.

En el cuadro 10.4 presentamos ejemplos de funciones de probabilidad conjunta de variables aleatorias discretas independientes y dependientes, así como del cálculo de la covarianza y la correlación.

CUADRO 10.4

Independencia de dos variables aleatorias discretas

a) Supongamos que extraemos al azar una bola de un bombo que contiene bolas rojas y blancas que tienen escrito un número entero del uno al cuatro. Si la bola es roja asignamos un uno, y si es blanca un cero (X). Igualmente, anotamos el número escrito en la bola (Y). La frecuencia relativa de cada tipo de bola (tanto colores como números) nos proporciona las probabilidades, que presentamos en una tabla de doble entrada. Cada probabilidad interior indica la probabilidad conjunta, mientras que las probabilidades marginales indican la función de probabilidad de cada variable simple.

		Y = 1	2	3	4	
X	1	0,10	0,20	0,15	0,20	0,65
	0	0,08	0,02	0,10	0,15	0,35
		0,18	0,22	0,25	0,35	1,00

Basta con que la condición de independencia no se cumpla en una de las casillas para concluir que esas variables no son independientes. En este caso, ya en la primera casilla observamos que no se verifica la condición: $0{,}10 \neq 0{,}65 \cdot 0{,}18$ y, por tanto, sin necesidad de examinar las demás casillas, ya podemos concluir que X e Y no son independientes.

b) Supongamos ahora que lanzamos una moneda imparcial y extraemos al azar una bola de las cuatro que contiene un bombo, numeradas del uno al cuatro. Asignamos un 1 si sale cara y un 0 si sale cruz (X), mientras que anotamos el número de la bola extraída (Y). La función de probabilidad conjunta es la siguiente:

		Y = 1	2	3	4	
X	1	0,125	0,125	0,125	0,125	0,500
	0	0,125	0,125	0,125	0,125	0,500
		0,250	0,250	0,250	0,250	1,000

c) Calculamos la covarianza y la correlación del primer ejemplo. En el segundo sabemos que necesariamente la covarianza y la correlación son nulas, puesto que X e Y son variables independientes:

$$E(X) = 1 \cdot 0{,}65 + 0 \cdot 0{,}35 = 0{,}65$$

$$E(Y) = 1 \cdot 0{,}18 + 2 \cdot 0{,}22 + 3 \cdot 0{,}25 + 4 \cdot 0{,}35 = 2{,}77$$

CUADRO 10.4 *(continuación)*

$$\sigma^2(X) = 1^2 \cdot 0{,}65 + 0^2 \cdot 0{,}35 - 0{,}65^2 = 0{,}2275$$

$$\sigma^2(Y) = 1^2 \cdot 0{,}18 + 2^2 \cdot 0{,}22 + 3^2 \cdot 0{,}25 + 4^2 \cdot 0{,}35 - 2{,}77^2 = 1{,}2371$$

$$E(XY) = 1 \cdot 1 \cdot 0{,}10 + 1 \cdot 2 \cdot 0{,}20 + 1 \cdot 3 \cdot 0{,}15 + 1 \cdot 4 \cdot 0{,}20 + 0 \cdot 1 \cdot 0{,}08 + \\ + 0 \cdot 2 \cdot 0{,}02 + 0 \cdot 3 \cdot 0{,}10 + 0 \cdot 4 \cdot 0{,}15 = 1{,}75$$

$$\sigma(XY) = 1{,}75 - 0{,}65 \cdot 2{,}77 = -0{,}0505$$

$$\rho(XY) = \frac{-0{,}0505}{\sqrt{0{,}2275} \cdot \sqrt{1{,}2371}} = -0{,}0952$$

Muchas veces se trabaja con variables aleatorias que son combinaciones lineales de otras variables aleatorias. Sus características se pueden deducir a partir de las características de las variables componentes. Resumimos estas relaciones a continuación:

Si $T = U + V + ... + X$

entonces, $E(T) = E(U) + E(V) + ... + E(X)$ [10.9]

Si $T = X + Y$

entonces, $\sigma^2(T) = \sigma^2(X) + \sigma^2(Y) + 2 \cdot \sigma(XY)$ [10.10]

Si $T = X - Y$

entonces, $\sigma^2(T) = \sigma^2(X) + \sigma^2(Y) - 2 \cdot \sigma(XY)$ [10.11]

10.4. VARIABLES ALEATORIAS CONTINUAS

El trabajo con variables aleatorias continuas no puede ser igual que con las discretas. Como en las discretas el número de resultados posibles es finito o infinito, pero numerable, tiene sentido hallar sumatorios como los que definen el valor esperado y la varianza. En cambio, en las variables continuas no se pueden aplicar esas fórmulas.

Podemos hacernos una mejor idea de la diferencia entre las variables aleatorias discretas y las variables aleatorias continuas, imaginándonos una flecha que gira sobre un eje. Si trabajamos con la variable que denota las posibles posiciones de parada de la flecha, la variable sería discreta en el caso de la figura 10.2*a*), mientras que en la 10.2*b*) sería continua. En el primer caso el círculo se ha dividido en

ocho sectores; la variable sólo puede asumir ocho valores distintos. En el segundo no hay segmentación de los puntos; el número de éstos sobre los que se puede detener es infinito, no numerable.

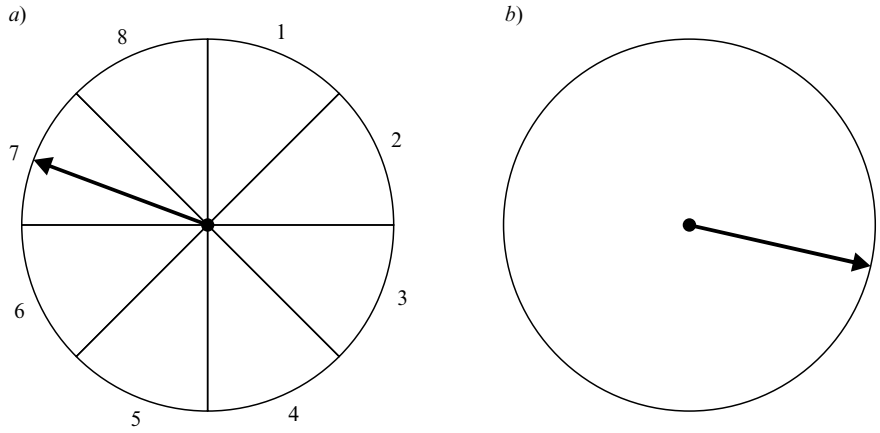


Figura 10.2.—Los puntos donde se puede detener una flecha que gira sobre un círculo constituyen una variable aleatoria continua, aunque se puede discretizar mediante la segmentación de esos puntos.

Mientras que las representaciones gráficas de las variables aleatorias discretas aparecen como un número más o menos grande de barras (véase el cuadro 10.1*d*), en las continuas aparecen como curvas que abarcan un cierto rango de valores sobre el eje de abscisas y que cubren un área igual a 1 (figura 10.3).

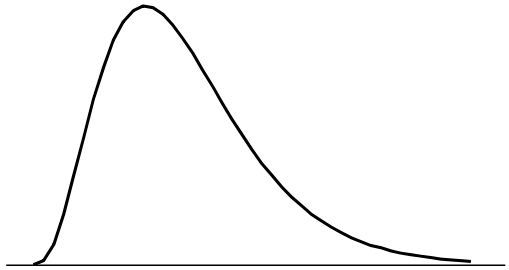


Figura 10.3.—Representación de una función de densidad.

10.4.1. Función de densidad y función de distribución

Para entender mejor cómo se asocian probabilidades a los valores de las variables aleatorias continuas utilizaremos una aproximación desde las variables discretas. La estatura es una variable continua, pero nosotros podríamos tratarla

como discreta si tomamos intervalos de 10 centímetros. Así, cada individuo estaría incluido en el grupo de los «cincuentas», o de los «sesentas», etc., tal y como aparece en la figura 10.4*a*). Supongamos ahora que dividimos cada intervalo en dos mitades, cada una incluyendo un rango de estaturas de cinco centímetros, como en la figura 10.4*b*). Podríamos continuar con este proceso de bipartición indefinidamente, de forma que, al tender a infinito el número de biparticiones, el número de rectángulos tendería a infinito, aunque la base de cada uno tendería a cero; el resultado final sería el de la figura 10.4*d*).

Por ello, en el contexto de las variables continuas no se habla de función de probabilidad, sino de *función de densidad* de probabilidad en torno a un valor. Una función de densidad asocia valores de la variable con ordenadas o alturas de la curva en cada punto. Técnicamente hablando, una función de densidad debe cumplir las dos condiciones que incluimos en el recuadro siguiente, en el que también definimos la función de distribución para variables continuas.

Se llama *función de densidad* de probabilidad de una variable aleatoria continua, X, que se representa por $f(x_i)$, a aquella que cumple las siguientes condiciones:

$$a) \quad f(x_i) \geq 0$$

$$b) \quad \int_{-\infty}^{\infty} f(x)\, dx = 1$$

Se llama *función de distribución* de una variable aleatoria continua, X, a aquella que asocia a cada valor de la variable la probabilidad de que ésta adopte como mucho ese valor; se representa por $F(x_i)$,

$$F(x_i) = \int_{-\infty}^{x_i} f(x)\, dx \qquad [10.12]$$

De estas definiciones se derivan las siguientes consecuencias:

$$a) \quad \mathrm{P}(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx \qquad (b \geq a)$$

$$b) \quad F(-\infty) = 0$$

$$c) \quad F(\infty) = 1$$

Las funciones de densidad no necesariamente presentan aspectos gráficos tan «redondos» como los de la figura 10.4*d*); las de la figura 10.5 también son representaciones gráficas de funciones de densidad.

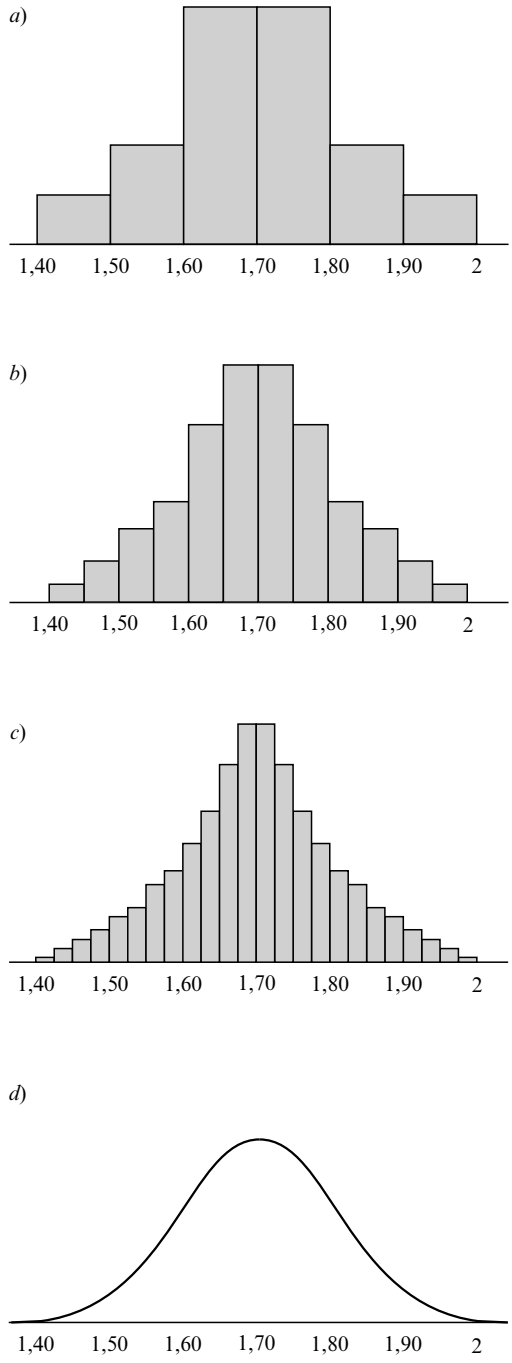


Figura 10.4.—Representación del proceso de conversión de una variable discretizada en una continua.

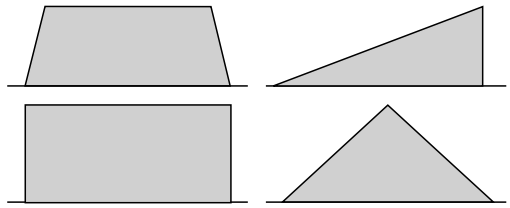


Figura 10.5.—Representación gráfica de variables aleatorias continuas.

10.4.2. Valor esperado y varianza

Al igual que las discretas, las variables aleatorias continuas tienen un valor esperado y una varianza que se pueden obtener con procedimientos análogos a los de aquéllas, pero adaptados al hecho de que la variable es continua. En concreto:

El *valor esperado* de una variable aleatoria continua se representa por $E(X)$, o μ, y se define como:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx \qquad [10.13]$$

La *varianza* de una variable aleatoria continua se representa por $\sigma^2(X)$ y se define como:

$$\sigma^2(X) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx - [E(X)]^2 \qquad [10.14]$$

El valor esperado y la varianza de las variables continuas tienen las mismas propiedades que hemos descrito con respecto a las variables discretas.

10.4.3. Relación entre dos variables aleatorias continuas

Dado que se trata de conceptos análogos a otros anteriores, nos limitaremos a proporcionar las definiciones y las fórmulas correspondientes. Sus propiedades también se mantienen.

Se llama *covarianza* entre dos variables aleatorias continuas, X e Y, y se representa por $\sigma(XY)$, a la expresión:

$$\sigma(XY) = E(XY) - E(X) \cdot E(Y) \qquad [10.15]$$

(continuación)

donde

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f(x, y)\, dx\, dy$$

Se llama *correlación* de Pearson entre dos variables aleatorias continuas, X e Y, y se representa por $\rho(XY)$, a la expresión:

$$\rho(XY) = \frac{\sigma(XY)}{\sigma(X) \cdot \sigma(Y)} \qquad [10.16]$$

Sobre variables aleatorias continuas se puede definir también la condición de independencia, análogamente a como lo hacíamos con las variables aleatorias discretas.

10.4.4. El trabajo aplicado con variables continuas

En la práctica, el trabajo con variables aleatorias continuas que se hace en estadística aplicada a las ciencias sociales consiste en hallar probabilidades, que casi siempre se reducen a uno de los tres casos siguientes y que hemos representado en la figura 10.6.

a) Calcular la probabilidad de que la observación sea como mucho igual a un determinado valor, o *probabilidad acumulada* para ese valor. En una representación gráfica aparece como el área que el valor deja a su izquierda [figura 10.6 *a*)]; es igual a:

$$\mathrm{P}(X \leq x_i) = F(x_i) = \int_{-\infty}^{x_i} f(x)\, dx$$

b) Calcular la probabilidad de que la observación sea igual o superior a un determinado valor; no es más que el complementario de la probabilidad acumulada. En una representación gráfica aparece como el área que el valor deja a su derecha [figura 10.6 *b*)]; es igual a:

$$\mathrm{P}(X \geq x_i) = 1 - \mathrm{P}(X \leq x_i) = 1 - F(x_i) = 1 - \int_{-\infty}^{x_i} f(x)\, dx$$

c) Calcular la probabilidad de que la observación esté comprendida entre dos valores cualesquiera. Se puede obtener por dos procedimientos: integrando desde el valor menor al mayor, o restando la probabilidad acumulada del menor de la probabilidad acumulada del mayor. En la representación gráfica aparece como un área limitada por la izquierda

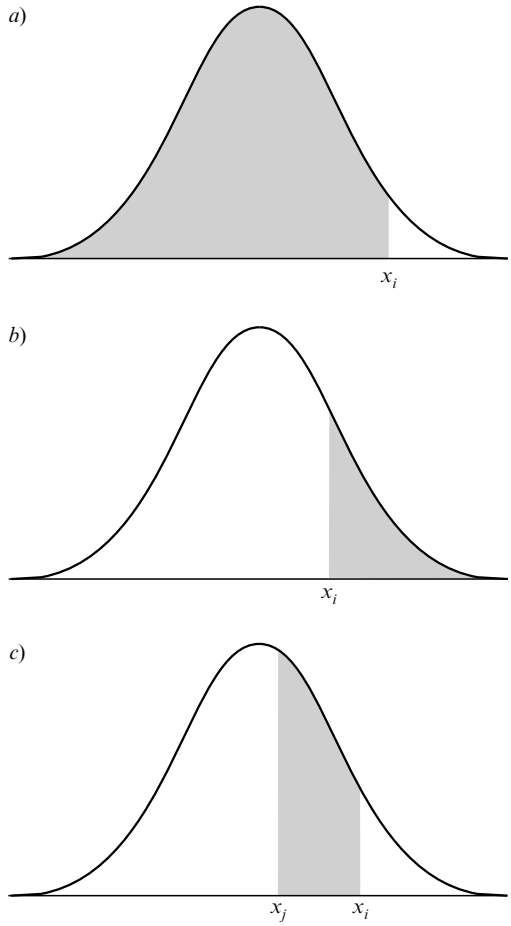


Figura 10.6.—Representación gráfica de las áreas que corresponden a las probabilidades de observar: *a*) valores menores que uno específico; *b*) mayores que uno específico, o *c*) comprendidos entre dos valores específicos.

por el valor menor y por la derecha por el valor mayor [figura 10.6 *c*)]; es igual a:

$$\mathrm{P}(x_j \leq X \leq x_i) = \int_{x_j}^{x_i} f(x)\, dx$$

o

$$\mathrm{P}(x_j \leq X \leq x_i) = F(x_i) - F(x_j) = \int_{-\infty}^{x_i} f(x)\, dx - \int_{-\infty}^{x_j} f(x)\, dx$$

Adviértase que si $x_i > x_j$, entonces necesariamente $F(x_i) \geq F(x_j)$.

10.5. DISTRIBUCIONES DE PROBABILIDAD

El lector se puede haber quedado con la idea de que el trabajo con variables aleatorias es muy laborioso, dado que exige conocer su función de probabilidad o de densidad de probabilidad. En la práctica no será necesario obtener los valores esperados, las varianzas y las demás características mediante los procedimientos que las definen según las fórmulas expuestas en este capítulo. La mayoría de los casos prácticos a los que nos enfrentamos en ciencias sociales se refieren a variables aleatorias, cuyas funciones de probabilidad o de densidad de probabilidad se ajustan a ciertos modelos teóricos. Conocer esos modelos y saberlos utilizar suele ser suficiente para resolver la mayoría de los problemas prácticos. Por ello, en los dos capítulos siguientes expondremos los modelos teóricos que con mayor frecuencia se utilizan en psicología.

10.6. MUESTREO ALEATORIO

Una de las aplicaciones de la probabilidad a la estadística consiste en el establecimiento riguroso de las condiciones necesarias para un buen muestreo. Ya vimos en el capítulo 1 que el muestreo es la actividad encaminada a la selección de un subconjunto de los elementos de una población. Hay muchos procedimientos para hacerlo, pero no todos conducen a muestras representativas. Sin embargo, había un punto débil en aquella argumentación. Dado que las poblaciones son desconocidas, ¿cómo podemos saber si una muestra es representativa de su población? Es imposible asegurarlo con certeza. Una clave para afrontar esta dificultad está en el procedimiento de extracción. Si las extracciones de elementos para componer la muestra se hacen al azar, las leyes de la probabilidad se encargan de que la tendencia general sea la de no obtener valores extremos sistemáticamente. La estadística inferencial se basará en que las muestras sobre las que se trabaje sean muestras aleatorias simples, concepto cuya definición pasamos a exponer.

Una *muestra aleatoria simple* (m.a.s.) compuesta por N elementos es una secuencia de N variables aleatorias, independientes e igualmente distribuidas. Es decir, si representamos por X_k al elemento extraído en k-ésimo lugar y por X_{k+1} al elemento extraído en el lugar siguiente a aquél, entonces:

$$P(X_{k+1} = x_i) = P(X_k = x_i) \qquad \text{para todo } k \text{ y } x_i$$

La extracción al azar no puede garantizar que la muestra extraída sea representativa, pero sí permite valorar las oscilaciones esperables en las observaciones. A efectos prácticos, basta con seguir un procedimiento que garantice que todos los elementos de la población tengan la misma probabilidad de ser extraídos y de ser incluidos en la muestra. Esto se consigue mediante extracciones simples si la

población es infinita, pero si es finita habría que hacer extracciones con reposición para que las probabilidades no se viesen alteradas por las extracciones anteriores. Por ejemplo, si extraemos un individuo de una población de 20 hombres y 20 mujeres, la probabilidad de que sea hombre es 20/40 = 0,50. Pero si queremos extraer una muestra mayor de individuos y lo hacemos sin reposición, la probabilidad de que el segundo sea también hombre dependerá de que lo haya sido el primero. En concreto, si el primer individuo es hombre, la probabilidad de que lo sea el segundo será 19/39, mientras que si el primero no lo fue esa probabilidad será 20/39. Para que la probabilidad sea constante, independientemente del resultado de las extracciones anteriores, será necesario reponer antes de hacer cada extracción los elementos extraídos previamente. En la práctica se trabaja con poblaciones infinitas, o con poblaciones que, aunque sean finitas, su tamaño es tan grande que el hecho de que se haga o no muestreo con reposición no modifica las probabilidades en cantidades apreciables; en esas condiciones se tratan como si fueran infinitas a efectos prácticos.

Un procedimiento idóneo para extraer una muestra aleatoria consistiría en tomar todos los elementos, numerarlos, introducir en una urna tantas bolas como números y proceder a la extracción. Hoy en día ya no se utiliza este procedimiento tan laborioso, sino que se emplean generadores de números aleatorios informatizados (véase el apéndice de este capítulo).

PROBLEMAS Y EJERCICIOS

1. La variable «número de pólizas vendidas por un agente de una empresa de seguros» es una variable aleatoria X que presenta la siguiente función de probabilidad:

X_i	0	1	2	3	4	5	6
$f(x_i)$	0,47	0,30	0,10	0,06	0,04	0,02	0,01

Conteste a las siguientes cuestiones:

a) Calcule el valor esperado y la varianza de X.
b) Si tomamos a uno de los agentes: ¿cuál es la probabilidad de que venda más de una póliza? ¿Y la de que venda menos de 3? ¿Y entre 1 y 4 pólizas (ambas inclusive)?
c) En el supuesto de que el director de la empresa premie con 100 puntos cada producto vendido, ¿qué representa la nueva variable (Y) y cuál es su distribución de probabilidad? ¿Cuál es el valor esperado y varianza de la variable aleatoria resultante?

2. Se enseña a una serie de ratas a recorrer un laberinto con cuatro salidas (A, B, C y D) equiprobables, donde sólo es considerada como correcta la salida A. Se introducen dos ratas en el laberinto. Si se consideran los resultados producto del azar, ¿cuál es la probabilidad de que al menos una rata salga por la salida correcta?

3. Un sujeto gana 6 euros si es capaz de detectar correctamente una señal acústica sobre un ruido de fondo, y pierde 25 euros en cualquier otro caso (omisión, detección incorrecta, etc.). Sabiendo que la probabilidad de detección de la señal es 0,99, averigüe cuál es el beneficio medio por detección.

4. La variable «número de errores cometidos en una tarea de agudeza visual» es una variable aleatoria X que presenta la siguiente función de distribución:

X_i	0	1	2	3	4	5	6
$F(x_i)$	0,45	0,70	0,80	0,89	0,92	0,98	1,00

Si se toma un sujeto al azar:

a) ¿Cuál es la probabilidad de que cometa al menos un error?
b) ¿Cuál es la probabilidad de que cometa como máximo 3 errores?
c) ¿Cuál es la probabilidad de que cometa entre 2 y 4 errores?

5. Halle el valor esperado y la varianza de la variable X, cuya función de probabilidad es la expuesta en la siguiente tabla:

X_i	5	10	15	20	25
$f(x_i)$	0,10	0,30	0,40	0,17	0,03

6. Los sujetos A, B y C extraen, por ese orden y sin reposición, una bola de una urna que contiene diez bolas numeradas del 1 al 10. Si les damos 10 euros multiplicados por el número impreso en la bola extraída, calcule:

a) El valor esperado para el primero.
b) El del segundo, si el primero ha conseguido 30 euros.
c) El del tercero, si entre los dos primeros han conseguido 50 euros.

7. Para obtener fondos con los que financiar el paso del ecuador, una promoción de estudiantes organiza una lotería con 1.000 papeletas que vende a 1 euro cada una. Si se da un primer premio de 200 euros, dos segundos de 50 euros y diez reintegros:

a) Halle el valor esperado de la variable «euros ganados».
b) Manteniendo las demás cantidades constantes, diga a cuánto habría que subir el primer premio para que éste fuese un juego justo.

8. Una prueba de razonamiento lógico consta de tres preguntas. Se sabe que la probabilidad de acertar la primera pregunta es igual a 0,60; la de acertar la segunda es igual a 0,30 y la de acertar la tercera es igual a 0,80. Sabiendo que se da un punto por cada pregunta acertada y cero en otro caso, y que la probabilidad de acertar una pregunta es independiente de acertar las otras dos, obtenga el valor esperado y la varianza de la suma de las tres puntuaciones.

9. Una aplicación del *juego justo* la encontramos en la forma de puntuar un ítem de k alternativas, donde una de las opciones es verdadera (respuesta correcta) y las otras son falsas (respuestas incorrectas). El objetivo es que el valor esperado de la puntuación del ítem sea igual a cero cuando éste se responde al azar. Los aciertos por azar se compensan penalizando las respuestas incorrectas. En este contexto, calcule cuánto tendrían que valer las respuestas incorrectas en un ítem de tres alternativas, cuando la respuesta correcta se puntúa con un uno.

10. Generalice el resultado del ejercicio anterior para un ítem de k alternativas.

11. Se está aplicando un programa de economía de fichas para mejorar el nivel de comunicación verbal en niños con trastornos del desarrollo, Para ello, y a lo largo de una sesión, se da una ficha si el niño realiza dos demandas de manera verbal; se dan dos fichas si emite tres o cuatro demandas verbales; tres fichas si

emite más de cuatro y, finalmente, ninguna ficha si sólo emite una o ninguna demanda verbal. Calcule el número de fichas esperado según las probabilidades que aparecen en la siguiente tabla.

X_i: Número de demandas	$f(x_i)$
0	0,10
1	0,13
2	0,30
3	0,25
4	0,10
5	0,06
6	0,05
7	0,01

12. Se ha observado que el número de clientes que, en promedio, van a un centro comercial en un día soleado es de 400, mientras que en un día nublado es igual a 750. El pronóstico del tiempo indica que la probabilidad de que el próximo día sea soleado es igual a 0,20. Calcule el número esperado de clientes que irán el próximo día al centro comercial.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*) $E(X) = 1{,}00$; $\sigma^2(X) = 1{,}74$.
b) $P(X > 1) = 0{,}23$; $P(X < 3) = 0{,}87$; $P(1 \leq X \leq 4) = 0{,}50$.
c) Si $Y = 100 \cdot X$, entonces: $E(Y) = 100$ y $\sigma^2(Y) = 17.400$.

2. 0,4375.

3. El beneficio medio por detección es 5,69 euros.

4. *a*) 0,55.
b) 0,89.
c) 0,22.

5. $E(X) = 13{,}65$; $\sigma^2(X) = 22{,}93$.

6. *a*) Para el sujeto A: $E(X) = 55$.
b) Sin reposición, y sabiendo que ya ha salido la bola 3, para el sujeto B: $E(X) = 57{,}78$.

c) Si entre el primero y el segundo han conseguido 50 euros, para el sujeto C: $E(X) = 62{,}50$.

7. *a*) $E(X) = -0{,}687$.

b) El primer premio tendría que ser igual a 890 euros.

8. $E(X_T) = 1{,}70$; $\sigma^2(X_T) = 0{,}61$.

9. Las respuestas erróneas se deberían penalizar con 0,5 puntos.

10. Las respuestas erróneas deberían restar $\frac{1}{k-1}$ puntos.

11. E(Núm. fichas) = 1,36.

12. E(Núm. clientes) = 680.

APÉNDICE

Generación de números aleatorios

Son muchas las circunstancias en las que se quieren obtener valores aleatorios. Son situaciones en las que actúa una o varias variables aleatorias y se pretenden obtener realizaciones de tales variables. Veamos unos ejemplos: a veces se intenta asignar aleatoriamente el orden de presentación de unos estímulos; en ocasiones se quiere extraer una muestra aleatoria simple de N elementos tomados de una población de M elementos que tenemos completamente identificados; en muchos trabajos de investigación en metodología se emplean simulaciones por ordenador de procesos aleatorios. En todos esos casos se quieren obtener valores que realmente sean aleatorios, por lo que se han ideado procedimientos para emular este carácter de «aleatorio». Una forma de hacerlo ha sido emplear realmente un bombo con diez bolas numeradas con dígitos, de 0 a 9, realizando una cantidad muy grande de extracciones. Con los valores obtenidos se construyeron tablas a las que se podía acudir a buscar una secuencia particular. Normalmente se decidía aleatoriamente en la tabla el primer número de la secuencia (llamado *semilla*). Por ejemplo, se puede determinar que la semilla sea el quinto valor de la novena fila.

Más recientemente, el procedimiento empleado reside en algoritmos que se ejecutan con medios electrónicos. La tecla *random* (RND) de las calculadoras de bolsillo, por ejemplo, nos proporciona valores basados en estos algoritmos. Hay muchos de estos algoritmos, y, siendo realistas, lo que proporcionan no son realmente valores aleatorios, sino valores pseudoaleatorios. Para que fueran aleatorios debería ocurrir que fuera imposible encontrar una secuencia regular en ellos. Las secuencias pseudoaleatorias son aquellas en las que se sabe que muestran regularidades, pero con un ciclo tan grande (a veces de millones de extracciones) que su comportamiento a efectos prácticos se puede considerar como aleatorio.

Modelos de distribución de probabilidad: variables discretas 11

11.1. INTRODUCCIÓN

La impresión que puede quedar tras la lectura del capítulo anterior podría ser que el trabajo con variables aleatorias es sumamente laborioso, dado que el cálculo del valor esperado, la varianza y la probabilidad acumulada asociada a cada valor exigiría un conocimiento exhaustivo de los valores y sus probabilidades. Afortunadamente, en la mayoría de los casos reales nos encontraremos con variables cuya función de probabilidad o de densidad de probabilidad se ajusta a alguna fórmula concreta. Cuando ocurre así, se dice que la función de probabilidad o de densidad de probabilidad de la variable se ajusta al modelo teórico expresado en esa fórmula. La consecuencia fundamental será que en el estudio de esas variables se podrán aplicar las propiedades de los modelos a los que se ajustan, que para los modelos más habituales son bien conocidas.

En este capítulo y el próximo vamos a exponer los modelos de distribución más utilizados en la estadística aplicada a la psicología. Mientras en este capítulo exponemos los modelos para variables aleatorias discretas, en el próximo expondremos los modelos para variables aleatorias continuas. En estos modelos encontraremos tanto aquellos a los que se ajustan variables que representan variables psicológicas (como los modelos de distribución binomial y normal), como aquellos que son instrumentos para el análisis estadístico y que nos resultarán muy útiles en la inferencia estadística.

Expondremos tres modelos para variables discretas: los modelos de distribución uniforme, binomial y multinomial. El primero no aparece habitualmente en variables de interés por sí mismas, pero creemos que como concepto es conveniente que sea conocido por los estudiantes de psicología, pues con frecuencia su nombre se utiliza para describir algunas situaciones (véase apartado 11.2). Por el contrario, el segundo tiene un alto interés histórico, pedagógico y práctico, que quedará de manifiesto a lo largo del capítulo. El tercero tiene muchas aplicaciones en la psicometría, que es la rama de la psicología que se dedica a desarrollar procedimientos para evaluar y medir las características psicológicas.

11.2. DISTRIBUCIÓN UNIFORME

En algunas ocasiones todos los valores que puede asumir una variable aleatoria discreta son equiprobables. Esto no es fácil de encontrar en variables propiamente psicológicas, pero sí en otros ámbitos, como por ejemplo el de los juegos de azar. Al lanzar un dado imparcial, todas las caras tienen la misma probabilidad de ser observadas; al hacer girar una ruleta todos los números de la mesa tienen la misma probabilidad de que la bola acabe en su posición, etc. En psicología se dan a veces casos en los que se asumen a priori distribuciones de este tipo. Así, se supone que las probabilidades de acierto (1) y error (0) en el primer ensayo (antes de cualquier aprendizaje) que ejecuta una rata en un laberinto en forma de T son iguales; cuando pedimos a un individuo que diga un número al azar suponemos una distribución uniforme de éstos; cuando alguien nos dice que puede transmitir telepáticamente un color, asumiremos que de no tener ese poder las probabilidades de que el receptor elija cualquiera de los colores del experimento es la misma. En todos estos casos se dice que las variables se distribuyen según el modelo uniforme, o simplemente que son uniformes. Otra aplicación que se encuentran los investigadores de la psicología es la asignación de participantes en una investigación a los grupos experimentales. Se utilizan procedimientos de este tipo, por ejemplo, para asignar a cada participante al grupo de control o a cada uno de los grupos experimentales, con un sistema que garantice que las probabilidades de formar parte de cada grupo son iguales.

Una variable aleatoria discreta se distribuye según el modelo *uniforme* si todos los valores con probabilidad no nula son equiprobables. Es decir, si la variable X sólo puede adoptar un número J de valores, entonces la función de probabilidad es la misma para todos esos valores:

$$f(x_i) = \frac{1}{J}$$

Con respecto al valor esperado y la varianza, no hay una fórmula general, puesto que dependerá de las circunstancias. Un caso frecuente es aquel en el que la variable adopta una serie de valores que se corresponden con una sucesión de números enteros, como en el caso del dado (1, 2, 3, 4, 5 y 6), que representamos en la figura 11.1. En estos casos, si la variable se distribuye según el modelo uniforme, su valor esperado es igual al promedio entre los valores máximo y mínimo; es decir:

$$E(X) = \frac{X_{máx} + X_{mín}}{2} \qquad [11.1]$$

Mientras que su varianza es:

$$\sigma^2(X) = \frac{(X_{máx} - X_{mín} + 1)^2 - 1}{12} \qquad [11.2]$$

En el ejemplo del dado, el valor esperado nos da igual a 3,5 y la varianza 2,92; naturalmente, son los mismos resultados que hubiéramos obtenido al aplicar las fórmulas generales [10.3 y 10.4].

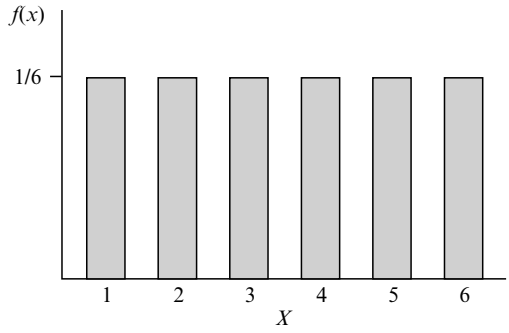


Figura 11.1.—Representación gráfica de la variable «valor observado al lanzar un dado imparcial», distribuida según el modelo uniforme.

Las fórmulas [11.1] y [11.2] no serían válidas si no fueran valores del tipo de los del dado. Por ejemplo, con una distribución ajustada a un modelo uniforme en la que los valores posibles (con probabilidad no nula) fueran 1, 2, 3 y 8, su valor esperado y su varianza no se podrían obtener de esa forma.

11.3. DISTRIBUCIÓN BINOMIAL

La distribución binomial es una de las que se conocen desde más antiguo. Fue extensamente estudiada por Jacob Bernouilli, hasta el punto de que en muchas ocasiones se dice que representa la frecuencia de resultados positivos en una serie de *ensayos de Bernouilli*.

Para que la distribución de probabilidad de una variable aleatoria se ajuste al modelo binomial, tienen que darse algunas condiciones. En primer lugar, debe estar involucrada una variable aleatoria dicotómica (capítulo 10), que es una que sólo admite dos resultados, habitualmente representados por los valores 1 y 0 (véase el cuadro 10.3). Esta variable puede ser una variable dicotómica «natural» o puede ser una variable «dicotomizada» artificialmente. Por ejemplo, la extracción de un individuo de la población y la asignación de un 1 en caso de ser varón y un 0 en caso de ser mujer es una variable genuinamente dicotómica. En cambio, la extracción de un individuo de la población y la asignación de un 1 si supera la puntuación 15 en el test BDI de Beck para la depresión y un 0 en caso contrario es una variable dicotomizada, puesto que, aunque el BDI tiene muchos posibles resultados numéricos diferentes, nosotros los clasificamos en dos: los que cumplen la condición de ser superiores a 15 y los que no la cumplen. En general, las va-

riables en las que se basa una variable binomial se pueden definir como aquellas que asumen la regla de asignar un 1 si se cumple una cierta condición y un 0 si no se cumple.

La segunda condición es que haya una repetición de N ensayos de la variable dicotómica, en los que la probabilidad de que en cada repetición se verifique la condición (y se asigne un 1) sea constante. Dicho de otra forma, la verificación de la condición en cada ensayo debe ser independiente de la verificación en los anteriores. A la probabilidad de verificación de la condición en cada ensayo independiente la representaremos por p.

La tercera y última condición es que se defina una variable, X, como el «número de casos que en la secuencia de N ensayos dicotómicos verifican la condición especificada»; o sea, el número de unos observados.

Podemos resumir los requisitos para la generación de una variable binomial de la siguiente forma:

Si:

a) Se define una variable dicotómica como el cumplimiento o el incumplimiento de una condición.
b) Se realiza una secuencia de N observaciones de esos ensayos dicotómicos en los que la probabilidad de verificación de la condición en cada repetición, π, es constante.
c) Se define una variable aleatoria, X, como el número de casos de esa secuencia en los que se cumple la condición.

entonces, la variable X se ajusta a un modelo binomial con parámetros N y π; se representa por:

$$X \sim B(N; \pi)$$

Pongamos como ejemplo el caso de los tres adolescentes sin experiencia previa que ya hemos descrito en capítulos anteriores, que eligen una de las dos cajas ofrecidas. Definimos la condición «elegir la caja con el logotipo a la izquierda» o, por abreviar, «izquierda». La presentación de las cajas a cada adolescente es un ensayo dicotómico creado por el cumplimiento/incumplimiento de esa condición. Dado que son adolescentes sin experiencia previa, la probabilidad de que cada uno de ellos elija la de la izquierda (cumpla la condición) es la misma. Muy verosímilmente esa probabilidad será 0,50, pero incluso aunque no lo fuera ello no sería un obstáculo para que la distribución sea binomial. Lo importante es que π sea constante en todos los ensayos, sin importar cuál sea ese valor. Por último, definimos la variable X: «Número de adolescentes que eligen la de la izquierda», o utilizamos la definición genérica «número de ensayos en los que se verifica la condición especificada». En estas circunstancias, podemos decir que la

variable X se distribuye según el modelo binomial con parámetros 3 y 0,50; o, expresado de otra forma:

$$X \sim \text{B}(3;\ 0{,}50)$$

De la forma como se genera una variable aleatoria binomial podemos deducir algunas de sus características:

a) Los valores de una variable binomial están comprendidos entre 0 y N, donde N es el número de ensayos dicotómicos realizados. Es decir, el número más pequeño posible de casos en los que se verifica la condición es «ninguno» y el máximo es «todos».

b) Si representamos el resultado de cada ensayo dicotómico con ceros y unos, el valor que adopta la variable X no es más que la suma de esa secuencia de unos y ceros.

c) El valor esperado de una variable binomial se obtiene a partir de las propiedades de la suma de variables aleatorias y de la definición de valor esperado. Dado que una binomial es la suma de una secuencia de N valores, y cada uno de ellos se puede considerar una variable aleatoria dicotómica, su valor esperado será igual a la suma de los valores esperados de cada una de ellas. En el capítulo anterior vimos que el valor esperado de una variable dicotómica es igual a la probabilidad de observar el valor 1, que hemos representado por π; por tanto,

$$E(X) = \mu = \pi + \pi + \ldots (N \text{ veces}) = N \cdot \pi \qquad [11.3]$$

d) Siguiendo la misma lógica, obtenemos la varianza de una variable binomial:

$$\sigma^2(\text{X}) = \pi \cdot (1 - \pi) + \pi \cdot (1 - \pi) + \ldots (N \text{ veces}) = N \cdot \pi \cdot (1 - \pi) \qquad [11.4]$$

e) Se demuestra que la función de probabilidad de una variable binomial viene dada por la expresión (véase el apéndice de este capítulo):

$$f(x_i) = \binom{N}{x_i} \cdot \pi^{x_i} \cdot (1 - \pi)^{N - x_i} \qquad [11.5]$$

Dado que lo importante de los modelos de distribución es que cuando una variable se ajusta a ellos se facilite el trabajo aplicando sus propiedades y características, vamos a ver cómo esa aplicación nos proporciona los mismos resultados que los obtenidos al aplicar las fórmulas generales. Así, la función de probabilidad obtenida mediante la fórmula [11.5] es la misma que la obtenida mediante la aplicación del enfoque clásico o a priori. Por ejemplo, la variable del ejemplo anterior, referida a las cajas elegidas por los adolescentes, tendría la siguiente función de probabilidad (que es la misma que obtuvimos por el procedimiento general en el cuadro 10.1) y que representamos en la figura 11.2:

X_i	$f(x_i)$
3	$\binom{3}{3} \cdot 0{,}50^3 \cdot 0{,}50^0 = 0{,}125$
2	$\binom{3}{2} \cdot 0{,}50^2 \cdot 0{,}50^1 = 0{,}375$
1	$\binom{3}{1} \cdot 0{,}50^1 \cdot 0{,}50^2 = 0{,}375$
0	$\binom{3}{0} \cdot 0{,}50^0 \cdot 0{,}50^3 = 0{,}125$

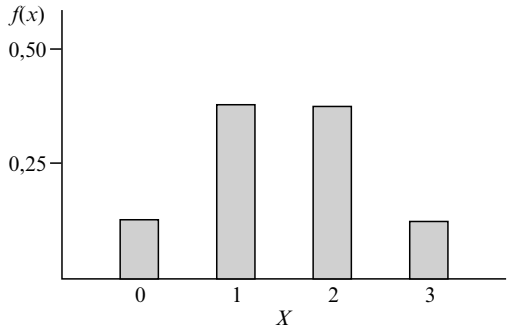


Figura 11.2.—Función de probabilidad de la variable «número de adolescentes que eligen la caja con el logotipo a la izquierda», distribuida según el modelo binomial.

Igualmente, el valor esperado y la varianza son también los mismos que obtuvimos en el cuadro 10.1:

$$E(X) = N \cdot \pi = 3 \cdot 0{,}50 = 1{,}5$$

$$\sigma^2(X) = N \cdot \pi \cdot (1 - \pi) = 3 \cdot 0{,}50 \cdot 0{,}50 = 0{,}75$$

No obstante, para abreviar los cálculos en la obtención de la función de probabilidad se han construido tablas en las que se han recogido las probabilidades asociadas a los valores de variables binomiales (tabla I del apéndice final). En el cuadro 11.1 recogemos algunos ejemplos numéricos del uso de esta tabla.

En la tabla I del apéndice final sólo se recogen valores de N hasta 20 y un número limitado de valores de π. El lector se preguntará cómo puede hallar las probabilidades asociadas en otros casos. Aunque por supuesto siempre se puede aplicar la expresión [11.5], ésta resulta muy laboriosa para obtener probabilidades acumuladas. Existe un procedimiento alternativo, basado en la aplicación de una fórmula de aproximación de este modelo a otro para variables continuas que expondremos en el apéndice del capítulo 12, así como las condiciones en las que se puede aplicar.

CUADRO 11.1

Ejemplos de obtención de probabilidades asociadas a una variable binomial

En la tabla de la binomial (tabla I del apéndice final) se incluye la función de probabilidad para valores seleccionados de N y π. Normalmente, lo que nos interesará es obtener probabilidades acumuladas u otras probabilidades asociadas a esas funciones. Por ello normalmente habrá que sumar las probabilidades de varios valores, tal y como mostramos en los siguientes ejemplos.
Supongamos que la variable X se distribuye según el modelo binomial con parámetros $N = 7$ y $\pi = 0{,}40$. Deseamos obtener: *a*) la probabilidad de observar un valor como máximo igual a 3; *b*) la de observar un valor como mínimo igual a 5, y *c*) la de observar un valor comprendido entre 2 y 4, ambos incluidos.

a) Se trata de la función de distribución (probabilidad acumulada) del valor 3 y, por tanto, la podemos calcular sumando las funciones de probabilidad de todos los valores desde 0 hasta ese valor. Es decir:

$$P(X \leq 3) = F(3) = f(0) + f(1) + f(2) + f(3) = \\ = 0{,}028 + 0{,}131 + 0{,}261 + 0{,}290 = 0{,}710$$

b) Se trata del complementario de la probabilidad acumulada del valor 4. Es decir:

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - F(4) = 1 - [f(0) + f(1) + ... + f(4)] = \\ = 1 - [0{,}028 + 0{,}131 + ... + 0{,}194] = 1 - 0{,}904 = 0{,}096$$

c) Se trata de la suma de las probabilidades acumuladas de los valores entre 2 y 4, ambos incluidos. Es decir:

$$P(2 \leq X \leq 4) = f(2) + f(3) + f(4) = \\ = 0{,}261 + 0{,}290 + 0{,}194 = 0{,}745$$

También se puede obtener como la función de distribución (probabilidad acumulada) del valor 4 menos la función de distribución (probabilidad acumulada) del valor 1:

$$F(4) - F(1) = 0{,}904 - 0{,}159 = 0{,}745$$

Una inspección de la tabla de la binomial nos permite apreciar una característica adicional de esta distribución: cuando $\pi = 0{,}50$ la distribución es simétrica. Eso significa que la probabilidad de los valores máximo y mínimo (N y 0) serán iguales, así como los de los pares de valores que equidisten de los extremos (e.g., los de 1 y N-1, los de 2 y N-2, etc.). Es lógico que sea así, puesto que definir la condición de la binomial con una de las opciones o la otra es arbitrario e indiferente (las probabilidades son ambas 0,50). En estas condiciones, será igual la probabilidad de obtener, por ejemplo, 3 caras y 7 cruces que 3 cruces y 7 caras con una moneda imparcial (véase en la tabla). Dicho de otra forma, si $X \sim B(N; 0{,}50)$, entonces: $f(x_i) = f(N - x_i)$. Con valores diferentes de π la distribución es asimétrica, pero será tanto menos asimétrica cuanto más se acerque π a 0,50.

11.4. DISTRIBUCIÓN MULTINOMIAL

En ocasiones se trabaja con variables aleatorias que, en lugar de adoptar dos posibles valores (dicotomía), pueden adoptar más de dos (politomía). En estos casos se pueden obtener las probabilidades asociadas a cualquier combinación de resultados mediante el modelo multinomial.

Supongamos una secuencia de N ensayos independientes e igualmente distribuidos en los que cada uno puede dar lugar a k resultados distintos con probabilidades $\pi_1, \pi_2, ..., \pi_k$ (siendo $\pi_1 + \pi_2 + ... + \pi_k = 1$). La probabilidad de que el resultado final consista en X_1 casos del primer tipo, X_2 del segundo tipo, ..., X_k del último tipo (siendo $X_1 + X_2 + ... + X_k = N$) viene dada por la fórmula:

$$f(x_1, x_2, ..., x_k) = \frac{N!}{x_1! \cdot x_2! \ldots x_k!} \cdot \pi_1^{x_1} \cdot \pi_2^{x_2} \cdot \pi_{3 \ldots}^{x_3} \pi_k^{x_k} \qquad [11.6]$$

Al comparar esta fórmula con la fórmula [11.5] es fácil advertir que la binomial no es más que un caso particular de la multinomial, en el que $k = 2$.

Veamos un ejemplo de cálculo. Supongamos que la organización de trasplantes ha estudiado las actitudes hacia la donación de órganos (variable X), encontrando que la probabilidad de tener una actitud contraria es 0,15, la de estar a favor 0,45 y la de mostrarse indiferente 0,40. Si se extrae una m.a.s. de 20 sujetos, podemos hallar, por ejemplo, la probabilidad de que esa muestra de 20 esté constituida por 5 contrarios, 5 favorables y 10 indiferentes. Sustituimos en la fórmula:

$$f(5, 5, 10) = \frac{20}{5! \cdot 5! \cdot 10!} \cdot 0{,}15^5 \cdot 0{,}45^5 \cdot 0{,}40^{10} = 0{,}41$$

Una aplicación de uso frecuente se encuentra en aquellos estudios en los que se explora si la distribución (multinomial) de una variable en dos poblaciones distintas (o en dos condiciones distintas) es la misma o es diferente. Por ejemplo, podríamos preguntarnos si la variable que acabamos de describir sobre las actitudes ante el trasplante de órganos es similar en un país mediterráneo como España, frente a un país culturalmente distante, como es Suecia. Representamos las dos distribuciones multinomiales en la siguiente tabla, en la que el primer subíndice de cada probabilidad representa el valor y el segundo representa la población (1, España; 2, Suecia).

	Favorables	Desfavorables	Indiferentes	
España	π_{11}	π_{21}	π_{31}	$\Sigma = 1$
Suecia	π_{12}	π_{22}	π_{32}	$\Sigma = 1$

Las tres probabilidades de cada fila constituyen la distribución multinomial de cada población. En situaciones como ésta nos preguntaremos si las dos distribuciones multinomiales son iguales; es decir, si las actitudes hacia la donación se reparten entre estas tres categorías de igual forma en ambos países ($\pi_{11} = \pi_{12}; \pi_{21} = = \pi_{22}; \pi_{31} = \pi_{32}$).

PROBLEMAS Y EJERCICIOS

1. En un experimento se presentan un par de estímulos auditivos de distinta intensidad a 5 participantes. La tarea consiste en indicar si los pares de estímulos son de igual o de distinta intensidad, con el fin de discriminar entre ellos. Supongamos que los sujetos responden al azar. Obtenga la distribución de probabilidad asociada al número de aciertos por azar.

2. Si de una población de estudiantes de primaria, de los cuales un 40 por 100 estudian en centros bilingües (español-inglés), extraemos una m.a.s. de 15 niños:

a) ¿Cuál es la probabilidad de que 10 de esos niños estudien en centros bilingües?

b) ¿Cuál es la probabilidad de que menos de 7 niños estudien en centros bilingües?

3. Extraemos, con reposición, una m.a.s. de seis estudiantes de primer curso de grado en una facultad de psicología donde hay 4 grupos por curso. Diga si en los casos que se presentan a continuación es posible obtener las probabilidades indicadas con una aplicación directa del modelo binomial:

a) La de que incluya cinco varones.

b) La de que incluya tres estudiantes del grupo 1, dos del grupo 2 y uno del grupo 3.

c) La de que incluya un estudiante de cada grupo.

d) La de que la cuarta parte sean mujeres.

e) La de que dos estudiantes sean mujeres del grupo 1 o del grupo 2.

4. Si un sujeto responde al azar a un examen tipo test en el que cada pregunta tiene una sola respuesta correcta, la variable «número de aciertos» tiene como valor esperado 6 y como varianza 4, calcule el número de preguntas que tenía el examen y el número de alternativas de cada pregunta.

5. Indique cuánto hay que descontar por cada pregunta mal contestada en un examen tipo test de cinco preguntas con tres alternativas de respuesta, donde sólo hay una correcta, para que un estudiante que responda al azar tenga como valor esperado 0 puntos.

6. Si al tirar un dardo tenemos una probabilidad de 0,30 de acertar en la diana, calcule el número mínimo de dardos que tenemos que usar para que al tirarlos tengamos una probabilidad superior a 0,90 de acertar, como mínimo, con un dardo.

7. Se hipotetiza que la probabilidad de acertar un ítem que mide razonamiento lógico es igual a la de fallarlo. Se pasa un mismo ítem de razonamiento lógico a una muestra de 12 participantes y se obtiene que lo aciertan 11 participantes. ¿Qué se podría concluir?

8. Tras un estudio realizado en una población de niños escolarizados, se ha obtenido que un 10 por 100 de niños escolarizados tienen alta capacidad cognitiva. Si se extrae una m.a.s. de 9 niños escolarizados, obtenga las probabilidades de que:

a) En la muestra haya al menos un niño con alta capacidad cognitiva.
b) En la muestra haya al menos un niño que *no* tenga alta capacidad cognitiva.

9. Tras realizar una encuesta exhaustiva sobre la calidad de vida en la población europea, se ha obtenido que el 70 por 100 consideran que su calidad de vida es satisfactoria, mientras que el 30 por 100 considera que no lo es. Considerando, en términos prácticos, que la población es infinita, se extrae una m.a.s. de tres europeos. Obtenga:

a) La probabilidad de que en la muestra haya dos europeos que consideren que su calidad de vida es satisfactoria.
b) La probabilidad de que haya un europeo que considere que su calidad de vida no es satisfactoria.

Comente los resultados de los dos apartados anteriores.

10. En una oposición para un puesto de la administración central, se ha observado que la probabilidad de acertar un ítem que evalúa conocimientos de la Constitución Española es igual a 0,50. Si se extrae una muestra con reposición de seis opositores, obtenga la probabilidad de que:

a) Sólo lo acierte un opositor.
b) Sólo lo acierten cinco opositores.

Comente los resultados de los dos apartados anteriores.

11. Si un estudiante responde al azar un examen tipo test de 14 preguntas de dos opciones (verdadero/falso), obtenga:

a) La probabilidad de que acierte sólo 3 preguntas.
b) La probabilidad de que acierte como mínimo 3 y como máximo 8.
c) El valor esperado y la varianza del número de aciertos.

12. Continuamos con el ejercicio anterior. Si el número de alternativas es igual a 6, de las cuales sólo una es verdadera, y manteniendo todo lo demás igual, responda a las mismas cuestiones del ejercicio anterior.

13. Generalizando los resultados de los dos ejercicios anteriores, la prueba está formada por N preguntas de K alternativas, de las cuales sólo una es correcta. Si el estudiante responde al azar, obtenga:

a) La función de probabilidad de la variable X: número de respuestas correctas.
b) El valor esperado y la varianza de la variable X.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. La probabilidad del primer valor es: $f(0) = \binom{5}{0} \cdot 0{,}50^0 \cdot 0{,}50^5 = 0{,}031$. Aplicando este procedimiento al resto de valores, se obtiene:

X_i	0	1	2	3	4	5
$f(x_i)$	0,031	0,156	0,312	0,312	0,156	0,031

2. Como $X \sim B(15; 0{,}40)$, entonces:

a) $P(X = 10) = 0{,}024$.
b) $P(X < 7) = 0{,}61$.

3. *a*) Sí.
b) No.
c) No.
d) Sí.
e) Sí.

4. El examen tenía 18 preguntas, con tres alternativas de respuesta de cada una.

5. Hay que restar 0,50 puntos por cada respuesta errónea.

6. Consultando la tabla correspondiente, deducimos que el número mínimo de dardos que tenemos que usar es 7.

7. Si la hipótesis fuera cierta, la probabilidad de que once participantes acertaran el ítem sería muy baja (0,003), por lo que debemos sospechar que la hipótesis es falsa.

8. *a*) 0,613.
b) Prácticamente 1.

9. *a*) 0,441.
b) 0,441.

Son iguales, ya que los dos sucesos (*calidad de vida satisfactoria* frente a *calidad de vida no satisfactoria*) son complementarios. Si en la muestra hay dos europeos que consideran satisfactoria su calidad de vida, quiere decir que en la misma muestra hay un europeo que considera que no lo es.

10. *a*) 0,094.
b) 0,094.

Se observa que ambas probabilidades son iguales (véase el comentario del último párrafo del apartado 11.3).

11. *a*) 0,22.
b) 0,78.
c) $E(X) = 7$ y $\sigma^2(X) = 3{,}5$.

12. *a*) 0,227.
b) 0,421.
c) $E(X) = 2{,}338$ y $\sigma^2(X) = 1{,}948$.

13. *a*) $X \sim B(N; \pi = 1/K)$.
b) $E(X) = N \cdot (1/K)$ y $\sigma^2(X) = N \cdot (1/K) \cdot [(K - 1)/K)]$.

APÉNDICE

Combinatoria

En la fórmula [11.5], que permite obtener la función de probabilidad de la distribución binomial, aparece la expresión $\binom{N}{x_i}$, que no es más que un resumen de otra expresión que calcula el número de combinaciones (ordenaciones) de unos y ceros que se pueden dar en N experimentos aleatorios independientes[1].

$$\binom{N}{x_i} = \frac{N!}{x_i! \cdot (N - x_i)!} \qquad [A.1]$$

Donde $N!$, $x_i!$ y $(N - x_i)!$ son los factoriales de N, x_i y $(N - x_i)$, respectivamente. Recuérdese que el factorial de un número natural, N, es igual al producto de los N primeros números naturales. Por tanto:

$$N! = 1 \cdot 2 \ldots (N - 1) \cdot (N)$$

Un caso especial es el del valor 0, para el que se asume que $0! = 1$. Vamos a ilustrar su uso con el ejemplo del apartado 11.3, que se refería a la distribución B(3;0,50):

$$\binom{3}{3} = \frac{3!}{3! \cdot (3 - 3)!} = 1$$

$$\binom{3}{2} = \frac{3!}{2! \cdot (3 - 2)!} = 3$$

$$\binom{3}{1} = \frac{3!}{1! \cdot (3 - 1)!} = 3$$

$$\binom{3}{0} = \frac{3!}{0! \cdot (3 - 0)!} = 1$$

[1] En realidad, la expresión [A.1] tiene una aplicación más general. Se usa en toda aquella situación en la que se hace necesario contar el número de combinaciones de N elementos tomados de x_i en x_i.

Modelos de distribución de probabilidad: variables continuas

12

12.1. INTRODUCCIÓN

Tal y como anticipamos en el capítulo anterior, en éste se exponen los modelos de distribución de probabilidad para variables continuas (véanse sus características en el capítulo 10). Comenzaremos por la distribución rectangular, que no es otra cosa que la análoga a la uniforme que hemos visto en variables discretas, pero para variables continuas. Después nos detendremos en la distribución normal, que no sólo es útil para la inferencia estadística, sino que es el modelo al que se aproximan muchas variables de interés en psicología. Por el contrario, otros modelos se incluyen porque son muy útiles como instrumento para el análisis estadístico, tal y como el lector podrá constatar al abordar la estadística inferencial, como son los modelos χ^2, t y F. La mayor parte de las técnicas inferenciales que se utilizan para la investigación en psicología tienen distribuciones de probabilidad que se ajustan a las de los modelos teóricos para variables continuas que vamos a describir a continuación.

En casi todos los modelos seguiremos el mismo esquema, explicando primero algunas condiciones, cuyo cumplimiento garantiza que la distribución de una variable se ajusta a él; después expondremos algunas de sus propiedades y, por último, abordaremos el trabajo práctico con ellos, que con frecuencia se reducirá exclusivamente al uso adecuado de alguna de las tablas que incluimos en el apéndice final del libro.

Al referirnos a valores concretos de variables aleatorias continuas, con frecuencia aparecerá un subíndice que indicará una probabilidad. En todos los casos esos subíndices reflejarán la función de distribución (área izquierda, probabilidad acumulada, o centil) de ese valor, y que representaremos en general por la letra p.

12.2. DISTRIBUCIÓN RECTANGULAR

Es la equivalente a la distribución uniforme de las variables discretas y, de hecho, a veces se hace referencia a ella con ese mismo nombre. Una variable aleatoria se ajusta a este modelo si todos los valores con probabilidad no nula tienen el mismo valor de la función de densidad de probabilidad. Como consecuen-

cia, su representación gráfica tiene forma de rectángulo; de ahí su nombre (figura 12.1).

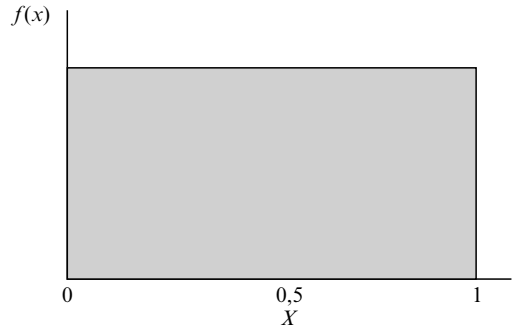


Figura 12.1.—Representación gráfica de la función de densidad de una variable distribuida según el modelo rectangular.

12.3. DISTRIBUCIÓN NORMAL

La curva normal o distribución normal es algo más que una fórmula matemática. Refleja fenómenos naturales, puesto que es frecuente encontrar variables con distribuciones empíricas muy semejantes a ella. Históricamente, los primeros desarrollos de la curva normal están un poco confusos (Walker, 1975), pero lo que sí parece claro es que el primero en expresar su fórmula fue De Moivre (1733), en un intento por dar una solución práctica al cálculo de las probabilidades acumuladas asociadas a la binomial cuando N es un número grande. De ahí procede lo que en el apéndice de este capítulo expondremos como un procedimiento de aproximación de la binomial a la normal. Los desarrollos posteriores realizados por otros autores han hecho que con frecuencia esta distribución se asocie a sus nombres. Así, también recibe los nombres de distribución de Gauss o de Laplace-Gauss. Otros nombres hacen referencia a la forma de su figura, como el de «campana» de Gauss.

La importancia de la curva normal estriba no sólo en su utilidad para el análisis estadístico, sino en que muchas variables de interés para los psicólogos, así como otras variables procedentes de la biología o la física, tienen distribuciones que se asemejan a la normal lo suficiente como para trabajar «como si» fueran normales, obteniendo valores muy aproximados. La estatura, el peso, la agudeza visual, la fuerza, etc., son variables que se aproximan a este modelo. Ya dentro de la psicología, variables que representan diversas capacidades o dimensiones de personalidad son variables con distribuciones empíricas muy parecidas a la normal.

En la representación gráfica de la curva (figura 12.2) se aprecia el porqué de su universalidad. En muchas variables existe un valor central (la media, μ) en torno al cual se concentran la mayor parte de los casos, mientras que a medida

que nos fijamos en valores más alejados de la media observamos que éstos son menos frecuentes. Esta reducción gradual en la frecuencia no es lineal, sino que es mayor al principio y menor después (la curva pasa de convexa a cóncava al alejarse de la media).

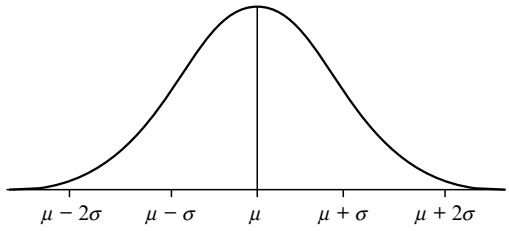


Figura 12.2.—Representación gráfica de la función de densidad de una variable aleatoria distribuida según el modelo normal.

La distribución normal refleja el incuestionable hecho de que la mayor parte de nosotros adoptamos valores intermedios en nuestras características psicológicas y sólo unos pocos sobresalen por adoptar valores especialmente altos o bajos. Es más, cuanto más se alejan los valores de la media más difícil es encontrar casos que adopten esos valores. Esto ocurre con múltiples variables de la naturaleza. Sin embargo, una de ellas tuvo especial importancia en el estudio y desarrollo de la curva normal: los errores. Por ejemplo, los errores perceptivos cometidos por observadores humanos al hacer mediciones se pueden cuantificar calculando la diferencia entre la medición hecha y la cantidad real. Cuando se carecía de instrumentos de precisión, las observaciones de los astrónomos se basaban en grandes cantidades de mediciones recogidas por distintos observadores y en diferentes momentos. El estudio de los valores registrados para una misma magnitud mostró que éstos adoptan una forma parecida a la de la curva normal. Esto impulsó el estudio de una fórmula que describiese la distribución de los errores. Obviamente, son más frecuentes las mediciones cercanas al valor real (errores pequeños) que las lejanas (errores grandes). De hecho, cuanto mayor es el error menor es su frecuencia.

Matemáticamente, una variable aleatoria se distribuye según el modelo normal (abreviaremos diciendo que es una variable normal), con parámetros μ y σ, si su función de densidad de probabilidad para todo valor de X viene dada por la fórmula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2} \qquad [12.1]$$

donde aparecen dos números con nombre propio en matemáticas: π y e. La forma compacta de representar esto es:

$$X \sim N(\mu; \sigma)$$

La fórmula [12.1] es la que alcanzó De Moivre, pero el uso del nombre «curva normal» para designarla es posterior. En tiempos de Galton ya se utilizaba, pero es seguro que el nombre es anterior a él.

Se puede demostrar que al aplicar a este modelo las fórmulas del valor esperado [10.13] y la varianza [10.14] para variables aleatorias continuas se obtienen los dos parámetros que hemos mencionado antes y que intervienen en la fórmula. En concreto, el parámetro que aparece en el numerador del exponente es el valor esperado (μ), y el que aparece en el denominador de la fórmula y en el del exponente es la varianza (σ^2). Ello nos permite reconocer en el exponente la fórmula de tipificación; es decir, que para variables tipificadas (z_i) esta fórmula toma un aspecto más sencillo, dado que la desviación típica es 1 y el valor esperado 0. Si X es una variable normal con valor esperado μ y desviación típica σ y hacemos el cambio de variable, que como recordará el lector es una «tipificación» (véase fórmula [4.8]):

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

entonces la función de densidad de esta nueva variable será:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2}$$

Las variables cuya distribución se ajustan al modelo normal adoptan una representación gráfica como la de la figura 12.2, en la que se pueden apreciar algunas de sus propiedades:

a) La distribución normal es simétrica con respecto a un valor central (μ), y en ese valor central coinciden la media (o valor esperado), la mediana (divide a la curva en dos zonas de igual área a su izquierda y a su derecha) y la moda (es el punto de la curva con máxima ordenada).
b) Es asintótica con respecto al eje de abscisas. Por mucho que se extienda, nunca llega a tocar los ejes; sólo en $\pm\infty$ la altura de la curva llegaría a ser igual a 0.
c) Hay toda una familia de curvas normales, dependiendo de los valores de μ y σ. De entre ellas, la más importante es aquella que tiene media 0 y desviación típica 1, para la que Sheppard (1899) propuso el nombre de *distribución normal unitaria*.
d) Los puntos de inflexión se encuentran en los puntos correspondientes a la media más/menos una desviación típica ($\mu \pm \sigma$).
e) Cualquier combinación lineal de variables aleatorias normales se ajusta también al modelo normal.

Antes de continuar, queremos resaltar una idea para contrarrestar un error frecuente en los aprendices de la estadística. Cuando una variable X con distribución normal se tipifica, esas típicas (z) también se ajustan al modelo normal.

Sin embargo, lo contrario no es cierto. Algunos caen en la tentación de tomar como si fuera normal cualquier variable que esté tipificada, pero no es verdad que toda variable tipificada se ajuste al modelo normal.

La mayor parte del trabajo práctico con variables aleatorias normales consiste en hallar probabilidades asociadas a valores. Tal y como vimos en el capítulo 10, esto significaría integrar la función de densidad entre los valores de interés. Para evitar tener que resolver este tipo de operaciones, se han construido tablas apropiadas con las áreas ya calculadas (la primera de ellas fue publicada por Sheppard en 1902) y cuyo uso se basa en el empleo de una regla de gran interés aplicado que nosotros llamaremos *regla de tipificación*. Según esta regla, la función de distribución asociada a un valor de una variable aleatoria, X, con distribución normal, es la misma que la función de distribución de la tipificada de ese valor en la normal unitaria (figura 12.3). Por eso, las tablas se han construido sólo para la distribución normal unitaria. Para obtener las áreas asociadas a un valor de cualquier otra distribución normal basta con tipificar ese valor (como las típicas son una transformación lineal con media 0 y desviación típica 1, su distribución es la normal unitaria) y acudir con la z obtenida a la tabla II del apéndice final.

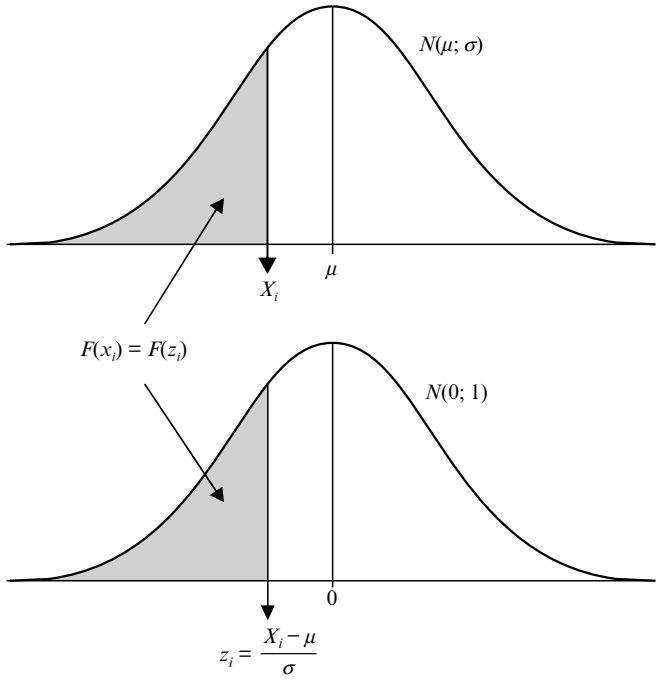


Figura 12.3.—Representación gráfica de la equivalencia entre las funciones de distribución de valores de variables normales (con parámetros μ y σ) y las de sus tipificadas en la distribución normal unitaria.

Dada su importancia, vamos a resaltar esta regla en un recuadro independiente:

> Según la *regla de tipificación* para variables normales, la función de distribución asociada a un valor de una variable aleatoria Normal, X, es igual a la de la tipificada de ese valor en la distribución Normal Unitaria. Es decir,
>
> Si $a)$ $X \sim N(\mu; \sigma)$
> $b)$ Transformamos la variable a $z_i = (X_i - \mu)/\sigma$
>
> entonces, $F(x_i) = F(z_i)$, donde $z \sim N(0; 1)$

Para referirnos a un valor concreto de la distribución normal unitaria utilizaremos la letra z, y a su derecha el subíndice correspondiente a la probabilidad acumulada para ese valor (z_p). Así, por ejemplo:

$$z_{0,67} = 0,44$$

indica que en la distribución normal unitaria el valor tipificado 0,44 tiene una probabilidad acumulada (función de distribución o área izquierda) igual a 0,67. Obviamente, cualquier z con un subíndice menor de 0,50 será un valor negativo, mientras que el valor 0 tendría un subíndice 0,50 ($z_{0,50} = 0$), puesto que el valor 0 es tanto la media como la mediana de la distribución normal unitaria.

Nuestro trabajo práctico con variables aleatorias normales se reduce a la obtención de las probabilidades de obtener un valor menor o igual que uno específico (o área izquierda de ese valor), la de obtener un valor mayor o igual que uno específico (o área derecha de ese valor) o la de obtener un valor comprendido entre dos valores específicos (o área limitada por esos dos valores). El procedimiento para obtener esas probabilidades consistirá en aplicar la regla de tipificación y consultar la tabla, tal y como se describe en los ejemplos del cuadro 12.1. Con frecuencia, lo que interesará será la tarea inversa, es decir, la de identificar la puntuación que deja una cantidad de área concreta a su izquierda o a su derecha. En el cuadro 12.2 hemos incluido algunos ejemplos, en los que lo que se busca es el valor para el que la probabilidad acumulada (o su complementaria) es igual a una cantidad específica.

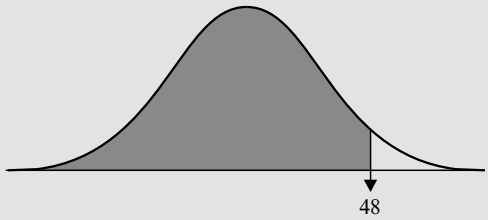
La distribución normal se utiliza también para obtener por aproximación las probabilidades asociadas a otros modelos. Ya hemos mencionado el caso de la distribución binomial, que exponemos en el apéndice de este capítulo, pero hay otras fórmulas de aproximación para otros modelos.

CUADRO 12.1

Ejemplos de obtención de probabilidades asociadas a variables normales

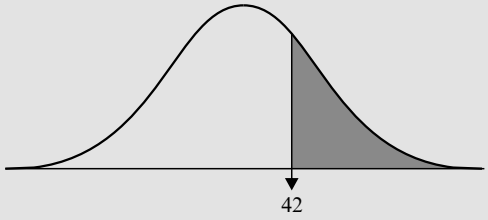
Supongamos que la variable X se distribuye $N(40; 5)$ y queremos obtener las siguientes probabilidades: *a*) la de observar un valor como mucho igual a 48; *b*) la de observar un valor como mínimo igual a 42, y *c*) la de observar un valor comprendido entre 34 y 38,25.

a) En el primer caso se trata de obtener la probabilidad acumulada del valor 48; para ello basta con tipificar y acudir con ese valor tipificado a la tabla de la normal unitaria (tabla II del apéndice final), dado que ésta nos proporciona directamente las áreas izquierdas. Es decir:



$$P(X \leq 48) = P\left(z \leq \frac{48 - 40}{5}\right) = P(z \leq 1{,}60) = 0{,}9452$$

b) En el segundo caso se trata de obtener el complementario de la probabilidad acumulada del valor 42. Hallamos la probabilidad acumulada del valor 42 por el procedimiento empleado en el apartado anterior y la restamos de 1. Es decir:

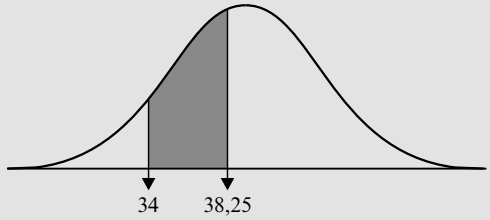


$$P(X \geq 42) = 1 - P(X \leq 42) = 1 - P\left(z \leq \frac{42 - 40}{5}\right) = 1 - P(z \leq 0{,}40) =$$

$$= 1 - 0{,}6554 = 0{,}3446$$

c) En el último caso se trata de obtener el área limitada por los valores 34 y 38,25. Tal y como vimos en el capítulo 10, el procedimiento más apropiado para ello consiste en hallar la diferencia entre la probabilidad acumulada del valor mayor y la del valor menor. Es decir:

CUADRO 12.1 *(continuación)*



$$P(34 \le X \le 38{,}25) = P\left(\frac{34-40}{5} \le z \le \frac{38{,}25-40}{5}\right) = P(-1{,}20 \le z \le -0{,}35) =$$

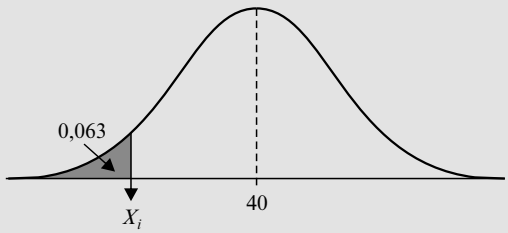
$$= F(-0{,}35) - F(-1{,}20) = 0{,}3632 - 0{,}1151 = 0{,}2481$$

CUADRO 12.2

Ejemplos de obtención de las puntuaciones de una variable normal con probabilidades específicas asociadas

Supongamos de nuevo que la variable X se distribuye $N(40; 5)$ y queremos obtener los valores de esta variable para los cuales se cumplen las siguientes condiciones: *a*) aquel para el que la probabilidad de observar un valor como mucho igual a él es 0,063; *b*) aquel para el que la probabilidad de observar un valor como mínimo igual a él sea 0,33, y *c*) aquellos dos valores que limiten el 50 por 100 central del área.

a) En el primer caso se trata de obtener el valor que deja un área a su izquierda igual a 0,063. Por la regla de tipificación, y acudiendo a la tabla II del apéndice final, comprobamos que se trata del valor cuya típica sea igual a −1,53. Basta con destipificar ese valor con respecto a la media y la desviación típica de la distribución:

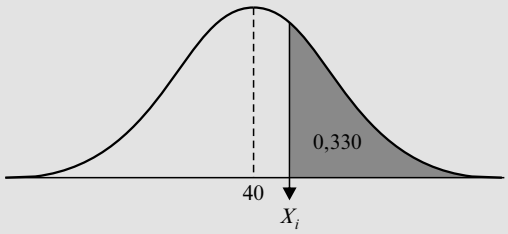


$$z_{0{,}063} = -1{,}53 = \frac{X_i - 40}{5} \qquad \text{despejando,} \quad X_i = -1{,}53 \cdot 5 + 40 = 32{,}35$$

b) En el segundo caso se trata de obtener el valor que deja un área a su derecha igual a 0,33. Como la tabla asocia a cada valor su área izquierda, y dado que la puntuación que deje a su derecha un área igual a 0,33 es la misma que deja a su izquierda un

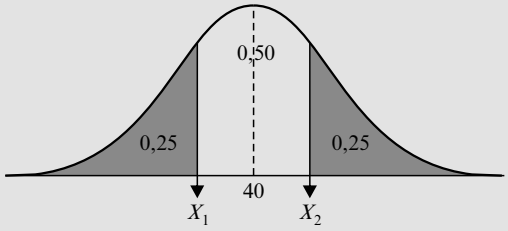
CUADRO 12.2 *(continuación)*

valor igual a 1 − 0,33 = 0,67, buscamos este valor de probabilidad en la tabla y aplicamos el mismo procedimiento que en el apartado anterior:



$$z_{0,67} = 0,44 = \frac{X_i - 40}{5} \qquad \text{despejando,} \quad X_i = 0,44 \cdot 5 + 40 = 42,2$$

c) Se trata de obtener aquellas dos puntuaciones que, tal y como aparece en la figura, dejen a su izquierda y derecha, respectivamente, áreas iguales a 0,25. Según la tabla, esas puntuaciones tendrán como típicas los valores 0,67 y −0,67. Destipificamos esos dos valores y obtenemos lo siguiente:



$$z_{0,25} = -0,67 = \frac{X_1 - 40}{5} \qquad \text{despejando,} \quad X_1 = -0,67 \cdot 5 + 40 = 36,65$$

$$z_{0,75} = 0,67 = \frac{X_2 - 40}{5} \qquad \text{despejando,} \quad X_2 = 0,67 \cdot 5 + 40 = 43,35$$

12.4. DISTRIBUCION χ^2 DE PEARSON

Al igual que en la distribución normal, en el modelo χ^2 de Pearson (que se lee «Ji-cuadrado») hay una fórmula (véase Amón, 1996) que define la función de densidad de probabilidad de una variable aleatoria. Sin embargo, lo que en estadística inferencial nos interesará más es un caso particular, del que se derivó históricamente esta distribución: la suma de variables aleatorias normales unitarias independientes elevadas al cuadrado. El número de valores sumados es el único parámetro de este modelo de distribución, pues las características de las norma-

les unitarias son constantes (valor esperado 0 y desviación típica 1). Es, por tanto, ese número de sumandos, que designaremos con el nombre de *grados de libertad,* el único parámetro que distingue a los miembros de esta familia de distribuciones (el concepto de grados de libertad excede los planteamientos de este libro; remitimos al lector interesado a libros de estadística inferencial: Glass y Stanley, 1970; Hays, 1988; Howell, 2009; Pardo, Ruiz y San Martín, 2009; Solanas, Salafranca, Fauquet y Núñez, 2005). En síntesis, unas condiciones que generan una distribución χ^2 son las siguientes:

Si	*a*)	$z_1, z_2, ..., z_k$ son valores de la distribución normal unitaria, independientes entre sí.
	b)	Formamos la variable $T = z_1^2 + z_2^2 + ... + z_k^2$.
entonces,		la variable aleatroria T se ajusta al modelo χ^2 con k grados de libertad, que representaremos de la siguiente forma: $T \sim \chi_k^2$.

Para referirnos a valores específicos de una distribución χ^2 pondremos en el subíndice derecho los grados de libertad y en el izquierdo la probabilidad acumulada (o función de distribución, o área izquierda) del valor correspondiente; es decir, ${}_p\chi_k^2$. Así, por ejemplo:

$$ {}_{0,30}\chi_{12}^2 = 9{,}034 $$

indica que en una distribución χ^2 con 12 grados de libertad, el valor que deja a su izquierda un área de 0,30 es 9,034. Obviamente, ese mismo valor es el que deja un área a su derecha igual a 0,70 (véase la tabla III del apéndice final).

De las condiciones de generación se deducen algunas de sus propiedades:

a) Dado que se suman valores tipificados elevados al cuadrado, se trata de una suma de sumandos necesariamente positivos y, por tanto, una variable distribuida según χ^2 no puede adoptar valores negativos; es decir:

$$ \text{Si} \quad T \sim \chi_k^2, \qquad \text{entonces} \quad P(T \leq 0) = 0 $$

b) Dado que la distribución normal unitaria es asintótica por ambas colas, la distribución de los cuadrados de sus valores será asintótica por la cola derecha (véase la figura 12.4).

c) Una consecuencia de que la distribución χ^2 sea asintótica por la derecha pero no por la izquierda es que la distribución tiene asimetría positiva.

d) El valor esperado y la varianza de una variable distribuida según χ^2 con k grados de libertad son:

$$ E(\chi_k^2) = k \qquad \sigma^2(\chi_k^2) = 2 \cdot k $$

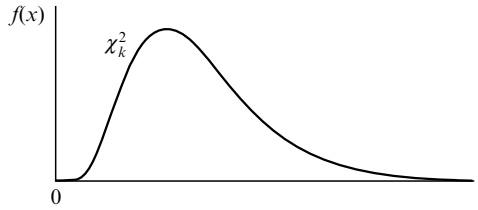


Figura 12.4.—Representación gráfica de la función de densidad de una variable distribuida según χ^2.

e) La asimetría de la distribución va disminuyendo a medida que crecen los grados de libertad; además, se puede demostrar que la distribución χ^2 tiende a la normal a medida que los grados de libertad tienden a infinito. La convergencia con la distribución normal es tan rápida que se ha llegado a establecer una fórmula de aproximación (véase en el apéndice del presente capítulo).

f) Si sumamos dos variables independientes distribuidas según χ^2, como sabemos que cada una representa una suma de típicas al cuadrado, el resultado será una suma de valores de la normal unitaria elevados al cuadrado en la que el número de sumandos será igual a los sumandos de una más los de la otra. Esta propiedad del modelo χ^2 recibe un nombre especial, que resaltamos a continuación:

La *propiedad aditiva* de χ^2 establece que:

Si
- *a*) T_1 y T_2 son variables aleatorias distribuidas según χ^2 con c y k grados de libertad, respectivamente.
- *b*) Formamos la variable aleatoria T_3 mediante la suma de valores de T_1 y T_2 extraídos de forma *independiente*, $T_3 = T_1 + T_2$.

entonces, T_3 se distribuye también según el modelo χ^2 y sus grados de libertad son iguales a la suma de los grados de libertad de T_1 y T_2.

Es decir,

Si
- *a*) $T_1 \sim \chi^2_c$ y $T_2 \sim \chi^2_k$.
- *b*) $T_3 = T_1 + T_2$.

entonces, $T_3 \sim \chi^2_{c+k}$

Como en el caso de la distribución normal, el trabajo práctico con variables distribuidas según χ^2 consiste sobre todo en la obtención de probabilidades asociadas a los valores, o viceversa (cuadro 12.3). Para ello se han construido tablas como la de nuestro apéndice final (tabla III), que incluye los valores de las varia-

bles χ^2 desde 1 hasta 30 grados de libertad y con las probabilidades acumuladas de uso más frecuente. Cuando los grados de libertad son mayores de 30 se puede aplicar la fórmula de aproximación del apéndice de este capítulo.

CUADRO 12.3

Ejemplos de utilización de la tabla de χ^2

Para obtener los valores que cumplen ciertas condiciones de probabilidad, basta con tomarlos directamente de la tabla III del apéndice final. Supongamos que la variable T se distribuye χ^2_{12} y queremos hallar: *a*) el valor tal que la probabilidad de obtener como máximo ese valor sea igual a 0,05; *b*) el valor tal que la probabilidad de obtener como mínimo ese valor sea 0,10, y *c*) los valores que generan una partición de la distribución en cuatro zonas con la misma área.

a) Según la tabla, el primer valor será ${}_{0,05}\chi^2_{12} = 5,226$.

b) Dado que la tabla asocia a cada valor una probabilidad acumulada, el segundo valor se obtiene buscando el que tiene una probabilidad acumulada igual a 0,90, puesto que ese mismo valor será el que deje un área derecha de 0,10; en concreto:

$$ {}_{0,90}\chi^2_{12} = 18,549 $$

c) En el tercer caso, se trata de encontrar los valores con probabilidades acumuladas iguales a 0,25, 0,50 y 0,75 (los tres cuartiles); según la tabla:

$$ {}_{0,25}\chi^2_{12} = 8,438 \qquad {}_{0,50}\chi^2_{12} = 11,340 \qquad {}_{0,75}\chi^2_{12} = 14,845 $$

Para obtener las probabilidades asociadas a ciertos valores basta con utilizar la tabla de χ^2 de forma inversa, es decir, localizando el valor de interés en el interior de la tabla y leyendo la columna en la que aparece. Supongamos ahora que la variable T se distribuye χ^2_{16} y queremos obtener la probabilidad de observar: *a*) un valor como mucho igual a 9,312; *b*) un valor como mínimo igual a 12,624, y *c*) un valor comprendido entre 19,369 y 20,465. Las probabilidades correspondientes son:

a) $P(T \leq 9,312) = F(9,312) = 0,10$.

b) $P(T \geq 12,624) = 1 - P(T \leq 12,624) = 1 - F(12,624) = 1 - 0,30 = 0,70$.

c) $P(19,369 \leq T \leq 20,465) = F(20,465) - F(19,369) = 0,80 - 0,75 = 0,05$.

12.5. DISTRIBUCIÓN *T* DE STUDENT

Para que una variable aleatoria se distribuya según el modelo *t* de Student se debe ajustar a la fórmula desarrollada al efecto por William S. Gosset. Hay un caso de gran interés para nosotros que se ajusta al modelo *t* y que se basa en la observación de valores independientes de una normal unitaria y de una χ^2 y la composición de un estadístico a partir de ellas. Obviamente, la distribución variará en función de los k grados de libertad de la variable χ^2 y, por tanto, tendremos de nuevo una familia

de distribuciones t, con un parámetro igual a los grados de libertad de la variable χ^2 que interviene en su generación. Diremos también que esa variable compuesta se distribuye según la t de Student con esos mismos grados de libertad. En concreto:

Si *a*) Las variables X e Y se distribuyen, respectivamente, $N(0; 1)$ y χ^2_k.

b) Formamos la variable T extrayendo valores independientes de X e Y y sustituyendo en la expresión:

$$T = \frac{X}{\sqrt{Y/k}}$$

entonces, la variable aleatoria T se ajusta al modelo t de Student con k grados de libertad, que representaremos de la siguiente forma:

$$T \sim t_k$$

Para referirnos a valores concretos de una distribución t pondremos de nuevo en el subíndice derecho los grados de libertad, mientras que a la izquierda pondremos la probabilidad acumulada (área izquierda o función de distribución) del valor correspondiente; es decir, ${}_pt_k$. Así, por ejemplo:

$${}_{0,10}t_{22} = -1{,}321$$

indica que en una distribución t con 22 grados de libertad el valor que deja un área izquierda de 0,10 es −1,321. Dado que esta distribución, al igual que la normal, es simétrica con respecto al valor 0, cualquier valor con un área izquierda menor de 0,50 será negativo, mientras que el valor 0 de una distribución t dejará a su izquierda y a su derecha áreas iguales a 0,50, con independencia de los grados de libertad que tenga.

De las condiciones de generación de la variable T se deducen algunas de sus propiedades (véase la figura 12.5):

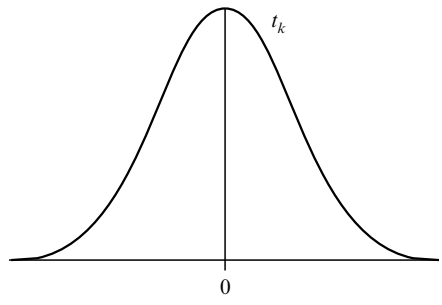


Figura 12.5.—Representación gráfica de la función de densidad de una variable distribuida según el modelo t de Student.

a) La distribución *t* de Student es simétrica con respecto al valor 0; es decir, ${}_pt_k = -{}_{1-p}t_k$.
b) En el valor 0 coinciden el valor esperado, la mediana y la moda.
c) El valor esperado y la varianza de una variable distribuida según t_k son:

$$E(t_k) = 0 \qquad \sigma^2(t_k) = k/(k-2)$$

para $k > 2$.
d) A medida que los grados de libertad se van incrementando, la distribución *t* de Student se va pareciendo más a la normal; de hecho, tiende a la normal cuando los grados de libertad tienden a infinito.

El trabajo práctico con la distribución *t* de Student se reduce de nuevo a la obtención de los valores que tienen asociadas ciertas probabilidades y las probabilidades asociadas a ciertos valores. Hemos incluido algunos ejemplos numéricos de ello en el cuadro 12.4.

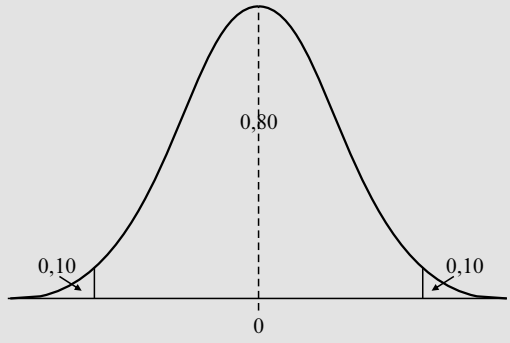
CUADRO 12.4

Ejemplos de utilización de la tabla t de Student

Para mostrar la aplicación de la tabla IV del apéndice final en la obtención de los valores que cumplen ciertas condiciones probabilísticas supondremos que la variable *T* se distribuye según el modelo t_{15} y obtendremos lo siguiente: *a*) el valor para el que la probabilidad acumulada es 0,10; *b*) aquel para el que la probabilidad de observar valores mayores es igual a 0,25, y *c*) aquellos dos valores que generan una partición en tres zonas tales que el área de la central sea cuatro veces mayor que la suma de las otras dos, y estas dos sean de igual área.
Según la tabla, los dos primeros valores serán:

a) ${}_{0,10}t_{15} = -1,341$ *b*) ${}_{0,75}t_{15} = 0,691$

c) En el tercer caso se trata de obtener los dos valores que generan la partición que aparece en la siguiente gráfica y que obtenemos también de la tabla:



${}_{0,10}t_{15} = -1,341$ y ${}_{0,90}t_{15} = 1,341$

CUADRO 12.4 *(continuación)*

Para obtener las probabilidades asociadas a valores concretos de esta distribución basta con utilizar la tabla de forma inversa. Supongamos que la variable T se distribuye según el modelo t de Student con 30 grados de libertad y queremos obtener las probabilidades de que esta variable adopte: *a*) valores como mucho iguales a 1,697; *b*) como mínimo iguales a −0,256, y *c*) comprendidos entre −1,310 y 0. Acudiendo a la tabla, obtenemos estas probabilidades:

a) La primera probabilidad es el área izquierda de ese valor, es decir:

$$P(T \leq 1{,}697) = F(1{,}697) = 0{,}95$$

b) La segunda es el complementario de la probabilidad acumulada, o área derecha:

$$P(T \geq -0{,}256) = 1 - P(T \leq -0{,}256) = 1 - F(-0{,}256) = 1 - 0{,}40 = 0{,}60$$

c) La tercera se refiere a un área central limitada por esos dos valores y, por tanto, es igual al área izquierda del mayor menos la del menor, es decir:

$$P(-1{,}310 \leq T \leq 0) = F(0) - F(-1{,}310) = 0{,}50 - 0{,}10 = 0{,}40$$

12.6. DISTRIBUCIÓN *F* DE SNEDECOR

La distribución F de Snedecor también se puede generar a partir de una peculiar composición de otras variables aleatorias con distribución conocida. En concreto, a partir de dos variables aleatorias independientes distribuidas según el modelo χ^2, no necesariamente con los mismos grados de libertad. Si se obtiene un estadístico a partir del cociente entre un valor de cada una de esas variables dividido por sus grados de libertad, tendremos una variable distribuida según la F de Snedecor. De nuevo tendremos tantas distribuciones F como valores pueden adoptar los grados de libertad de las dos variables que intervienen. Es importante hacer notar que los grados de libertad no serán intercambiables; siempre identificaremos cuáles son los grados de libertad de la expresión del numerador y cuáles son los de la expresión del denominador. Por tanto:

Si
- *a*) Las variables X e Y se distribuyen según el modelo χ^2 con m y n grados de libertad, respectivamente.
- *b*) Extraemos valores independientes de X e Y.
- *c*) Formamos la variable T extrayendo valores independientes de X e Y y sustituimos en la expresión:

$$T = \frac{X/m}{Y/n}$$

(continuación)

entonces,	la variable aleatoria T se distribuye según el modelo F de Snedecor con m y n grados de libertad; lo expresaremos de la siguiente forma, donde el primer subíndice siempre representará los grados de libertad del numerador y el segundo los del denominador: $$T \sim F_{m,n}$$

De nuevo, para referirnos a valores concretos de una distribución F pondremos en el subíndice derecho los grados de libertad del numerador y del denominador, por ese orden, mientras que en el de la izquierda reflejamos la probabilidad acumulada (o área izquierda) del valor correspondiente ($_pF_{m,n}$). Así, por ejemplo:

$$_{0,90}F_{2,4} = 4{,}32$$

indica que en una distribución F con 2 grados de libertad en el numerador y 4 en el denominador, el valor que deja un área izquierda de 0,90 es 4,32.

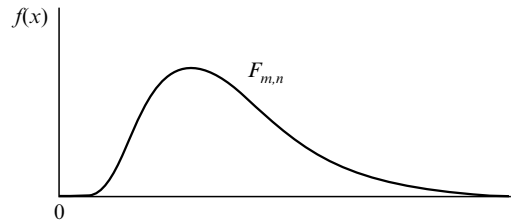


Figura 12.6.—Representación gráfica de la función de densidad de variables distribuidas según el modelo F de Snedecor.

La distribución F de Snedecor tiene también algunas propiedades importantes, de las que resaltaremos las siguientes (véase la figura 12.6):

a) Dado que se basa en valores positivos (los valores de las distribuciones χ^2 lo son necesariamente y los grados de libertad son enteros positivos), una variable distribuida según este modelo sólo puede adoptar valores positivos.
b) Las variables distribuidas según F también tienen como valor mínimo el cero, mientras que son asintóticas por la cola derecha; dado que las variables χ^2 son asintóticas por la derecha, la F también lo es.
c) Se trata de una distribución asimétrica, aunque a medida que los grados de libertad se van incrementando la distribución se va haciendo menos asimétrica; de hecho, la distribución F tiende a la normal cuando ambos grados de libertad tienden a infinito.

d) El valor esperado y la varianza de una variable distribuida según $F_{m,n}$ son:

$$E(F_{m,n}) = n/(n-2) \qquad \sigma^2(F_{m,n}) = \frac{2 \cdot n^2 \cdot (m+n-2)}{m \cdot (n-4) \cdot (n-2)^2}$$

e) Si una variable se distribuye según t_k y elevamos los valores al cuadrado, esos cuadrados se distribuyen según $F_{1,k}$. Elevemos al cuadrado la expresión con la que generábamos una distribución t_k:

$$T = \frac{z^2}{Y/k}$$

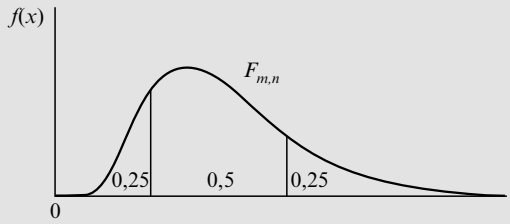
En esta expresión reconocemos en el numerador una variable distribuida según χ^2 con un grado de libertad, dividida por sus grados de libertad (dividido por 1), mientras que en el denominador aparece una variable distribuida según χ^2_k dividida por sus grados de libertad (k); por tanto, si elevamos al cuadrado una variable que se distribuye t_k, entonces se distribuye según $F_{1,k}$.

Como en casos anteriores, el trabajo práctico con esta distribución se suele reducir a la obtención de los valores que limitan ciertas áreas y de las probabilidades asociadas a valores concretos de la misma. En el cuadro 12.5 presentamos algunos ejemplos numéricos de cómo se utiliza la tabla V del apéndice final.

CUADRO 12.5

Ejemplos de utilización de la tabla F de Snedecor y de una de sus propiedades

Para mostrar la aplicación de la tabla de F en la obtención de los valores que cumplen ciertas condiciones probabilísticas supondremos que la variable T se distribuye según el modelo $F_{12,15}$. Obtenemos mediante la tabla V del apéndice final lo siguiente: *a*) el valor para el que la probabilidad de observar como mucho ese valor es 0,25; *b*) aquel para el que la probabilidad de observar como mínimo ese valor es 0,90, y *c*) aquellos valores que generan una partición del área como la que aparece en la figura siguiente:



a) En el primer caso se trata del valor con área izquierda igual a 0,25:

$$_{0,25}F_{12,15} = 0{,}676$$

CUADRO 12.5 *(continuación)*

b) En el segundo caso se trata del valor con un área derecha igual a 0,90 y, por tanto, su probabilidad acumulada es 0,10:

$$_{0,10}F_{12,15} = 0,475$$

c) En el tercero se trata de los siguientes valores:

$$_{0,25}F_{12,15} = 0,676 \qquad _{0,75}F_{12,15} = 1,44$$

Para obtener los valores que cumplen ciertas condiciones probabilísticas utilizamos la tabla de forma inversa. Así, si la variable T se distribuye según el modelo F con 10 y 8 grados de libertad, podemos obtener: *a*) la probabilidad de observar un valor como mucho igual a 1,02; *b*) como mínimo igual a 3,35, y *c*) comprendido entre 0,164 y 0,259.

a) La primera probabilidad no es más que la probabilidad acumulada del valor, es decir:

$$P(T \leq 1,02) = F(1,02) = 0,50$$

b) La segunda es el complementario de la probabilidad acumulada, es decir:

$$P(T \geq 3,35) = 1 - P(T \leq 3,35) = 1 - F(3,35) = 1 - 0,95 = 0,05$$

c) La tercera es igual al área izquierda del valor mayor menos el área izquierda del valor menor, es decir:

$$P(0,164 \leq T \leq 0,259) = F(0,259) - F(0,164) = 0,025 - 0,005 = 0,020$$

PROBLEMAS Y EJERCICIOS

1. Sabiendo que la variable X se distribuye según una normal con media 20 y varianza 64, determine la probabilidad de extraer al azar una observación cuya puntuación sea:

a) Menor de 26.
b) Menor de 18.
c) Mayor de 30.
d) Mayor de 13,2.
e) Entre 16 y 28.
f) Entre 24 y 36.

2. Considerando la distribución del ejercicio anterior, calcule cuál es la probabilidad de que una observación se separe de la media en, como mínimo, media desviación típica.

3. Considerando la distribución de los dos ejercicios anteriores, calcular la probabilidad de extraer al azar una observación que tenga una puntuación diferencial como mínimo igual a 18.

4. Si la variable V sigue una distribución $N(10; 4)$, calcular el valor o valores que cumplen que:

a) $F(v_i) = 0{,}6064$.
b) La probabilidad de obtener ese valor o cualquier otro inferior es igual a 0,2033.
c) La probabilidad de obtener ese valor o cualquier otro superior es igual a 0,0136.
d) Entre ellos se encuentre el 75 por 100 central de la distribución.

5. En una población de estudiantes de secundaria se ha observado que las puntuaciones, X, en una prueba de matemáticas se aproximan a una distribución $N(5; 2)$, mientras que en una prueba de historia las puntuaciones, Y, se aproximan a una distribución $N(6; 3)$. Asumiendo que ambas puntuaciones son independientes en dicha población, obtenga:

a) La probabilidad de seleccionar un estudiante al azar que tenga en la prueba de matemáticas una puntuación igual o superior a 6 y al mismo tiempo una puntuación en la prueba de historia igual o superior a 7.
b) Si se obtiene una m.a.s. de 8 estudiantes de dicha población, ¿cuál es la probabilidad de que la mitad de la muestra tenga una puntuación en historia igual o superior a 7,56?

6. Se ha observado que el nivel de rendimiento en una tarea de memoria, variable X, de una determinada población se distribuye según una normal. Sabiendo que el 80 por 100 de la población supera un valor igual a 14,64 y que el tercer cuartil es igual a 20,68, obtenga la media y la varianza de dicha variable.

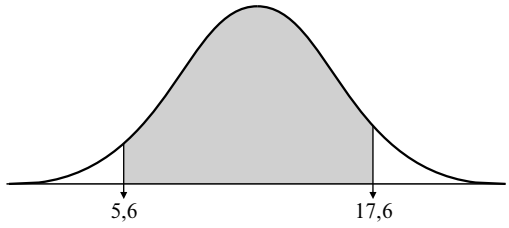
7. Se asume que el nivel de ansiedad-estado, variable X, sigue una distribución $N(8; 2)$ en una población determinada. Si se toma una m.a.s de 1.000 personas de dicha población:

a) ¿Cuántas personas cabría esperar que se encuentren entre los niveles de ansiedad 7 y 9?

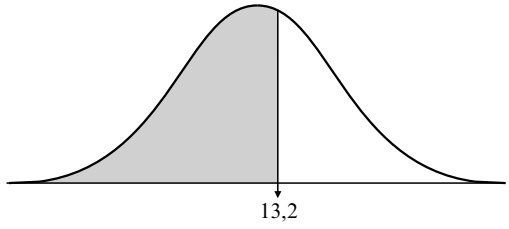
b) ¿Cuántas cabría esperar que tengan un nivel de ansiedad como mínimo igual a 12?

8. Sabiendo que $X \sim N(12; 4)$, obtenga las probabilidades asociadas a las áreas sombreadas.

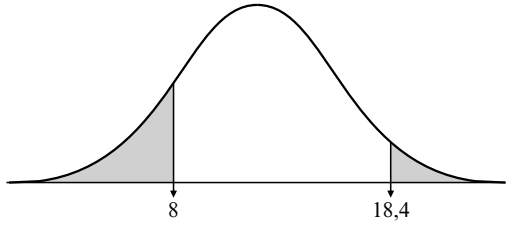
a)



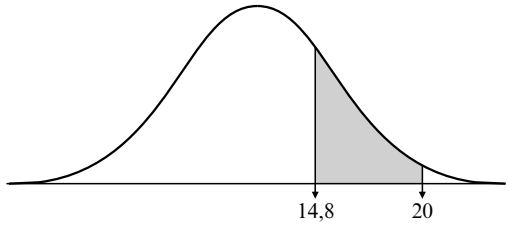
b)



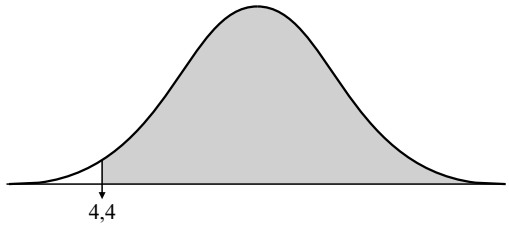
c)



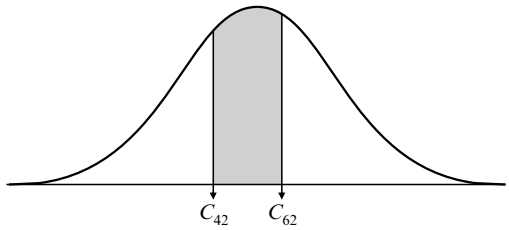
d)



e)

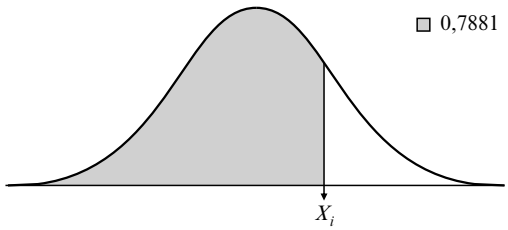


f)

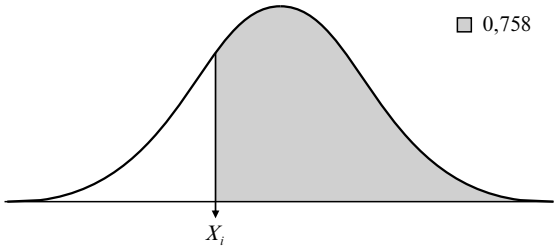


9. Sabiendo que $X \sim N(12; 4)$, obtenga la puntuación o puntuaciones asociadas a las áreas indicadas en la leyenda de cada gráfico.

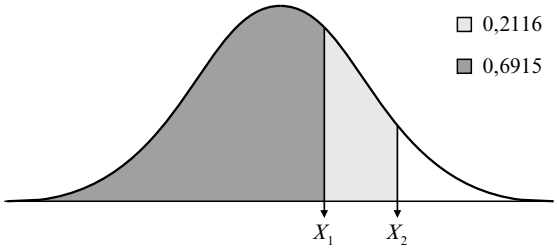
a)



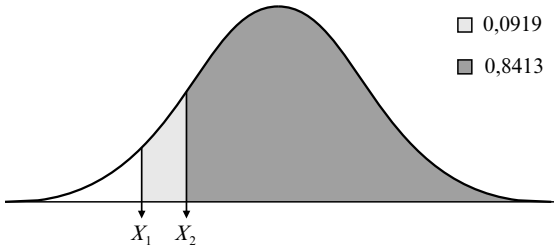
b)



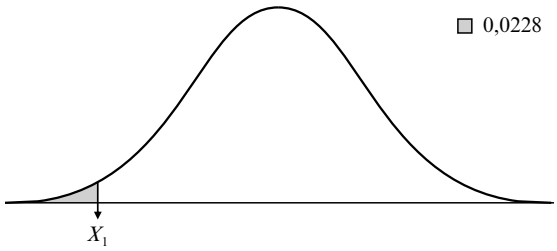
c)



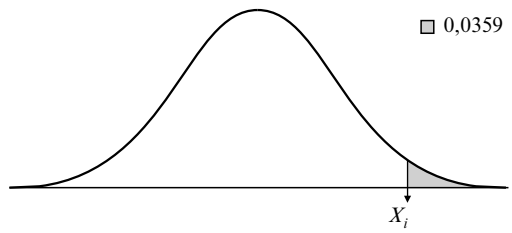
d)



e)



f)



10. Tras administrar una prueba diagnóstica sobre estrés, X, a un grupo de 300 personas, se observa que las puntuaciones en dicha prueba se aproximan a una distribución normal con media 30 y varianza 36. Responder a las siguientes cuestiones:

a) ¿Cuántas personas no superan la puntuación 39?
b) Si se establece que toda aquella persona que tenga una puntuación igual o superior a 42 tiene un riesgo alto de sufrir estrés, ¿cuántas personas cabría esperar que satisfagan este criterio?
c) Si en el 10 por 100 inferior de la distribución se sitúan las personas que cabe esperar que su riesgo de sufrir estrés sea bajo, obtener la puntuación mínima que se ha de tener en la prueba para pertenecer a este grupo.

11. En un estudio sobre discriminación de colores se ha observado que las puntuaciones que obtienen los profesionales que trabajan en artes plásticas siguen una distribución $N(80; 10)$, mientras que las puntuaciones de los profesionales de otras áreas siguen una distribución $N(60; 20)$. Extrayendo de manera aleatoria e independiente un profesional de cada población, obtenga las probabilidades de que:

a) Ambos tengan una puntuación como mínimo igual a 67,5.
b) Ambos tengan una puntuación como mínimo igual a 70.

12. Se ha aplicado un test a 1.000 personas cuyas puntuaciones, X, se aproximan a una normal. Del total de las puntuaciones, 330 son iguales o inferiores a 23,24, mientras que 123 son iguales o superiores a la puntuación 29,64. Según lo anterior, obtenga la media y la desviación típica de la distribución.

13. Obtenga las puntuaciones que permiten dividir a la distribución $N(22; 6)$, que sigue la variable W, en cuatro partes cada una de ellas, conteniendo el 25 por 100 del área de dicha distribución.

14. Se ha administrado un cuestionario en la población de titulados universitarios europeos acerca del nivel de satisfacción sobre su trabajo. Se ha obtenido que en la subpoblación de hombres las puntuaciones se aproximan a una distribución $N(6; 1)$ y en la subpoblación de mujeres se aproximan a una distribución $N(4; 2)$. Responder a las siguientes cuestiones:

a) Si se extrae al azar a un hombre, ¿cuál es la probabilidad de que supere la mediana de la distribución de mujeres?

b) Si se extraen de manera aleatoria e independiente dos hombres, ¿cuál es la probabilidad de que ambos superen la mediana de la distribución de mujeres?

15. Se ha observado que en la población infantil el nivel de activación se distribuye aproximadamente normal con media 12 y desviación típica 3, mientras que en la población de adultos también se aproxima a una normal pero con media 10 y desviación típica 2. Responder a las siguientes cuestiones:

a) ¿Qué porcentaje de adultos no superan la moda de la distribución de los niños?

b) ¿Qué porcentaje de los niños superan el tercer cuartil de la distribución de los adultos?

c) ¿Qué puntuación en la distribución de adultos no es superada por el mismo porcentaje que no supera la puntuación 13 en la población infantil?

16. En una investigación se establece que la recta de regresión que permite pronosticar la orientación espacial, O, a partir del cociente intelectual, CI, es $O' = 10 + 0{,}5 \cdot CI$. Además, se asume que la variable CI sigue una distribución $N(100; 15)$. Si se extrae una persona al azar, ¿cuál es la probabilidad de que el pronóstico que se hiciera a esa persona en O estuviera entre 60 y 67,5?

17. ¿Qué proporción de las puntuaciones de una distribución normal se separan de la media en, como mínimo, 1/4 parte de la desviación típica?

18. Si el rendimiento en una tarea perceptiva, X, en estudiantes que sólo cursan el grado universitario de Física se distribuye según $N(20; 3)$, mientras que el rendimiento en la misma tarea en estudiantes que sólo cursan el grado de Pedagogía, Y, se distribuye $N(25; 4)$, y se extrae un estudiante de cada población, calcule la probabilidad de que ambos no superen la puntuación 24.

19. En un centro escolar en el que se imparten actividades extraescolares, hay un grupo que realiza patinaje artístico. Sabiendo que el rendimiento en patinaje se mide mediante un test cuyas puntuaciones se distribuyen, aproximadamente, según la distribución normal con media 50 y desviación típica 9, y que se ha extraído una m.a.s. de 200 estudiantes, indique cuántos estudiantes esperaremos tener en las categorías que aparecen a continuación: nivel alto a partir de 65 puntos; nivel medio entre 45 y 65 puntos, y nivel bajo menos de 45 puntos.

20. Indique qué puntuaciones tendríamos que haber utilizado para definir las categorías del ejercicio anterior de manera que hubiésemos obtenido las siguientes frecuencias en cada categoría: nivel alto 25 estudiantes; nivel medio 150 estudiantes, y nivel bajo 25 estudiantes.

21. Tres de cada diez españoles son favorables al retraso de la edad de jubilación a los 67 años. Si extrajésemos una m.a.s. de 100 españoles:

a) ¿Cuál sería la probabilidad de que fueran favorables 35 o más?
b) ¿Y la de que fueran favorables entre 25 y 40 (ambos inclusive)?

22. Dada la variable X, que se distribuye $N(\mu; \sigma)$, extraemos una m.a.s. de 12 observaciones y definimos la variable Y que consiste en la tipificación de las X, elevadas al cuadrado y sumadas. Según esto, conteste a las siguientes preguntas:

a) Diga cuál es el modelo de distribución de Y.
b) Indique cuánto valen las probabilidades de que la puntuación en Y sea, respectivamente, menor de 14,845, mayor de 3,571, y una cantidad comprendida entre 9,034 y 23,337.

23. Continuando con el ejercicio anterior, diga para qué valor o valores de Y se verifican las siguientes condiciones:

a) La probabilidad de obtener valores mayores es 0,20.
b) La probabilidad de obtener valores menores es 0,30.
c) Acotan un área central de 0,50.
d) Corresponde al C_{90}.
e) Corresponde al D_2.

24. En un sondeo electoral, ¿cuál es la probabilidad de que en una m.a.s. de 80 encuestados haya mayoría de partidarios de la «abstención» cuando en realidad en la población éstos son sólo el 40 por 100?

25. En la tarea utilizada por Botella, Villar y Ponsoda (1988) se presentaba a los sujetos una letra tomada del grupo T, E, F, I y L, equiprobablemente, y los sujetos debían identificar la letra presentada, sabiendo que necesariamente era una de esas cinco. Cada observador pasó 70 ensayos. Responda a las siguientes cuestiones, referidas a un sujeto que respondiese al azar en la tarea:

a) Indique el número esperado de aciertos.
b) Obtenga la probabilidad de que acierte 20 ensayos o más.
c) Halle la probabilidad de que acierte más de lo esperado por azar.
d) ¿A cuánto habría que elevar el número de letras para que la proporción esperada de aciertos fuese el 10 por 100 de los ensayos?

26. En un estudio de observación sobre el juego infantil, ponemos un espejo unidireccional en una escuela infantil y evaluamos a un niño cada cinco minutos durante cuatro horas (48 evaluaciones). Anotamos el número de veces en las que el niño está jugando solo. Si un niño está, en realidad, el 40 por 100 del tiempo jugando solo, calcule la probabilidad de que en ese muestreo el niño esté jugando solo:

a) Menos de la mitad de las veces.
b) Más de la cuarta parte de las veces.

27. En una ciudad, el 40 por 100 de los ciudadanos están a favor del partido A, el 25 por 100 a favor del partido B, el 20 por 100 a favor del partido C y el 15 por 100 restante son indecisos. Si extraemos una m.a.s. de 40 ciudadanos:

a) ¿Cuál es la probabilidad de que en esa muestra sean mayoría los que están a favor del partido A?
b) ¿Cuál es la de que haya al menos siete ciudadanos a favor del partido C?

28. Si la variable X se distribuye según el modelo binomial con parámetros $N = 20$ y $\pi = 0{,}40$, y calculamos la probabilidad de observar un valor igual o mayor de 12, diga en cuánto nos equivocaríamos si la obtenemos por la aproximación a la normal, en lugar de obtener la probabilidad exacta por medio de la propia binomial.

29. Una empresa multinacional tiene vacantes varios puestos de trabajo con un perfil de ventas. En un primer proceso de selección se ha administrado un test de aptitud X, cuyas puntuaciones se distribuyen en la población aproximadamente $N(30; 6)$. Si se establece como criterio que sólo pasan este proceso los candidatos con una puntuación superior a 33 en X:

a) ¿Cuál es la probabilidad de que un candidato, extraído al azar, se quede fuera del proceso de selección?
b) ¿Qué límite deberíamos haber puesto en el test para que sólo el 34 por 100 de los candidatos pasen el criterio de selección?
c) ¿Hasta qué puntuación habría que elevar el criterio de selección para reducir a la mitad el porcentaje de candidatos que podrían ser admitidos con el criterio original?
d) Supongamos ahora que se admite a los candidatos que cumplan, indistintamente, o el criterio original del test X o que tengan una puntuación en inteligencia (I) superior a 120, sabiendo que, en la población, $I \sim N(100; 15)$, y que esa variable es independiente de las puntuaciones en el test de aptitud, ¿cuál es la probabilidad de que un candidato extraído al azar sea admitido en el proceso de selección?

30. Utilizando los datos del problema anterior, si extraemos una m.a.s. de diez sujetos, ¿cuál es la probabilidad de que más de la mitad cumpla el criterio del test de aptitud?

31. La variable cociente intelectual, CI, se distribuye $N(100; 15)$. Si llevamos a cabo una investigación sobre inteligencia y extraemos una m.a.s. de 50 participantes, ¿cuál es la probabilidad de que al menos diez de los participantes superen la puntuación correspondiente al tercer cuartil de la población?

32. Sabiendo que la variable U se distribuye χ^2_{16}, obtener:

a) Un valor que es superado por el 2 por 100 de la distribución.
b) La probabilidad de extraer al azar una puntuación como mínimo igual a 12,624.
c) Las puntuaciones que determinan una zona central de la distribución con área igual al 50 por 100.
d) La puntuación correspondiente al D_2.
e) La probabilidad de extraer al azar una puntuación igual o inferior a 26,296.

33. Si la variable V sigue una distribución χ^2_8, obtener las probabilidades de encontrar valores:

a) Menores o iguales que 17,535.
b) Mayores o iguales que 3,490.
c) Entre 9,524 y 11,030.

34. Si la variable Y sigue una distribución t de Student con 10 grados de libertad, obtenga las probabilidades de encontrar valores:

a) Menores que 1,372.
b) Mayores que 0,542.
c) Menores que −1,812.
d) Comprendidos entre −2,228 y 0.

35. Si la variable X sigue la distribución t_{17}, obtenga el valor o valores de la variable:

a) Que delimitan un área central del 60 por 100.
b) Que es superado por el 1 por 100 de la distribución.
c) Que acumula una probabilidad igual a 0,25.
d) Que corresponde a Q_2.
e) Que corresponde al C_{60}.

36. Si las variables X, V e Y se distribuyen según χ^2 con 3, 6 y 9 grados de libertad, respectivamente, y sabiendo que son independientes, responda a las siguientes cuestiones:

a) La probabilidad de obtener valores en V iguales o superiores al D_2 de la distribución de la variable Y.
b) Si se define la variable $W = X + V + Y$, obtenga el C_{99} de dicha variable.

37. La variable Y sigue una distribución χ^2 con 25 grados de libertad. Si se extrae un valor al azar, ¿cuál es la probabilidad de que sea como máximo igual a 29,339? Obtenga el error que se cometería si se utiliza la aproximación a la normal.

38. Sabiendo que la variable X se distribuye χ^2 con k grados de libertad, determine el valor de los grados de libertad de:

a) ${}_{0,02}\chi^2_k = 9{,}915$
b) ${}_{0,995}\chi^2_k = 20{,}278$
c) ${}_{0,50}\chi^2_k = 3{,}357$
d) ${}_{0,10}\chi^2_k = 18{,}114$
e) ${}_{0,90}\chi^2_k = 21{,}064$
f) ${}_{0,995}\chi^2_k = 14{,}860$

39. Si la variable X sigue una distribución t de Student con k grados de libertad, ¿es posible que se cumpla que ${}_{0,10}t_k > 0$?

40. Partiendo de la variable X, distribuida $N(0; 1)$ y la variable Y distribuida χ^2_{14}, y siendo X e Y variables independientes, se define la variable $U = \dfrac{X}{\sqrt{Y/14}}$. Obtenga:

a) $P(U \leq 0)$
b) El C_{20} de U
c) $P(U \geq 2{,}624)$

41. Considerando que la variable X se distribuye t_6, obtenga:

a) $F(0{,}906)$
b) $P(-1{,}943 \leq X \leq 0{,}553)$

42. Si la variable del ejercicio anterior se transforma mediante la expresión $W = X^2$. Obtenga:

a) $P(W \leq 13{,}7)$
b) $P(W \geq 0{,}111)$
c) $F(0{,}017)$

43. Sea la variable Y que se distribuye según χ^2_{19}. Calcule:

a) ${}_{0,10}\chi^2_{19}$
b) ${}_{0,999}\chi^2_{19}$
c) ${}_{0,05}\chi^2_{19}$
d) ${}_{0,01}\chi^2_{19}$
e) ${}_{0,99}\chi^2_{19}$
f) ${}_{0,95}\chi^2_{19}$

44. La variable aleatoria X se distribuye según el modelo χ^2 con 15 grados de libertad y la variable aleatoria Y se distribuye según el modelo χ^2 con 5 grados de libertad. Asumiendo que X e Y son independientes, si se define la variable aleatoria $V = X + Y$:

a) ¿Cuál es el modelo de distribución para la variable V?
b) ¿Cuál es el valor de $E(V)$ y $\sigma(V)$?
c) Calcule: $P(V \leq 34{,}17)$, $P(8{,}26 \leq V \leq 16{,}226)$ y $P(V \geq 35{,}02)$.

45. Sabemos que la variable tiempo de reacción *TR* ante una tarea auditiva (medida en milisegundos) se distribuye según χ^2 con veinte grados de libertad en la población. Si extraemos de forma aleatoria e independientemente una muestra de cinco sujetos de dicha población, ¿cuál es la probabilidad de que al menos tres superen el valor 22,775?

46. Sea X una variable aleatoria distribuida según F de Snedecor con $m = 5$ y $n = 10$ grados de libertad. Calcule:

a) $P(X \leq 2{,}52)$
b) $P(X \leq 3{,}33)$
c) $P(3{,}33 \leq X \leq 5{,}64)$
d) ${}_{0{,}05}F_{5,10}$
e) ${}_{0{,}01}F_{5,10}$
f) ${}_{0{,}10}F_{5,10}$
g) $P(X \geq 1{,}59)$
h) C_{25}
i) $P(X \geq 4{,}24)$
j) $P(4{,}24 \leq X \leq 5{,}64)$
k) $P(X \geq 10{,}5)$
l) $P(X \leq 12{,}4)$

47. Sea X una variable aleatoria con distribución normal. Se define la variable aleatoria $Y = 2 \cdot X$ con distribución $N(4; 2)$. Según lo anterior, calcule:

a) $E(X)$
b) $\sigma^2(X)$
c) $P(0{,}5 \leq X \leq 2{,}5)$

48. La variable estatura (medida en centímetros) sigue el modelo $N(165; 9)$ en la población de mujeres nacidas en el año 1995, mientras que en la de varones nacidos el mismo año tal distribución es $N(175; 11)$. Según lo anterior, calcule:

a) La probabilidad de que una mujer tenga una estatura superior a 167 cm.
b) La misma probabilidad en caso de ser varón.

49. Sabiendo que la variable $T \sim \chi^2_k$, donde k son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con las probabilidades acumuladas correspondientes a los valores que se indican en la misma.

k	Valor χ^2	Probabilidad acumulada
10	3,059	( )
5	3,000	( )
20	22,775	( )
26	31,795	( )
18	13,675	( )
2	1,386	( )

50. Sabiendo que la variable $T \sim t_k$, donde k son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con las probabilidades acumuladas correspondientes a los valores que se indican en la misma.

k	Valor *t*	Probabilidad acumulada
10	1,812	( )
5	−0,559	( )
21	−1,323	( )
25	−0,684	( )
80	1,292	( )
15	0	( )

51. Sabiendo que la variable $T \sim F_{m,n}$, donde m y n son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con las probabilidades acumuladas correspondientes a los valores que se indican en la misma.

m	*n*	Valor *F*	Probabilidad acumulada
2	5	1,85	( )
4	6	0,942	( )
10	9	2,42	( )
2	12	0,106	( )
10	10	0,336	( )
60	24	1,02	( )

52. Sabiendo que la variable $T \sim \chi^2_k$, donde k son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con los valores correspondientes a las probabilidades acumuladas que se indican en la misma.

k	Probabilidad acumulada	Valor χ^2
9	0,70	( )
5	0,25	( )
2	0,90	( )
16	0,50	( )
18	0,10	( )
20	0,01	( )

53. Sabiendo que la variable $T \sim t_k$, donde k son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con los valores correspondientes a las probabilidades acumuladas que se indican en la misma.

k	Probabilidad acumulada	Valor *t*
15	0,950	( )
18	0,200	( )
2	0,300	( )
70	0,400	( )
5	0,975	( )
100	0,500	( )

54. Sabiendo que la variable $T \sim F_{m,n}$, donde m y n son los grados de libertad que se especifican para cada caso, rellene la tabla inferior con los valores correspondientes a las probabilidades acumuladas que se indican en la misma.

m	*n*	Probabilidad acumulada	Valor *F*
12	20	0,75	( )
6	30	0,50	( )
60	120	0,90	( )
1	10	0,10	( )
20	20	0,05	( )
10	120	0,25	( )

55. Si la variable aleatoria T_1 se distribuye según t_{15} y la variable aleatoria T_2 se distribuye según t_{80}, calcule la probabilidad de que T_2 supere la puntuación correspondiente al centil 95 de T_1.

56. La variable aleatoria T_1 se distribuye según t_7 y la variable aleatoria T_2 se distribuye según $F_{24,15}$. Obtenga la probabilidad de que una observación de T_2 extraída al azar sea superior a la puntuación correspondiente al centil 95 de T_1.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*) 0,7734.
b) 0,4013.
c) 0,1056.
d) 0,8023.
e) 0,5328.
f) 0,2857.

2. 0,617.

3. 0,0122.

4. *a*) $V_i = 11{,}08$.
b) $V_i = 6{,}68$.
c) $V_i = 18{,}84$.
d) $V_1 = 5{,}4$ y $V_2 = 14{,}6$.

5. *a*) Suceso A: «$X \geq 6$»; Suceso B: «$Y \geq 7$». $P(A \cap B) = 0{,}1144$.
b) 0,136.

6. $\mu_X = 18$ y $\sigma_X^2 = 16$.

7. *a*) Aproximadamente 383 personas.
b) Aproximadamente 159 personas.

8. *a*) 0,8644.
b) 0,6179.
c) 0,2135.
d) 0,2192.
e) 0,9713.
f) 0,20.

9. *a*) 15,2
b) 9,2.
c) $X_1 = 14$ y $X_2 = 17{,}2$.
d) $X_1 = 6$ y $X_2 = 8$.
e) $X_i = 4$
f) $X_i = 19{,}2$.

10. *a*) Aproximadamente 280 personas.
b) Aproximadamente 7.
c) $X_i = 22{,}32$.

11. *a*) 0,3148.
b) 0,2595.

12. $\mu_X = 25$ y $\sigma_X = 4$.

13. $Q_1 = 17{,}98$; $Q_2 = 22$; $Q_3 = 26{,}02$.

14. *a*) 0,9772.
b) 0,9549.

15. *a*) Aproximadamente el 84,13 por 100 de la población de adultos no supera la moda de la distribución de la población infantil.
b) El 58,71 por 100 de la población infantil supera el Q_3 de la distribución de la población de adultos.
c) El 62,93 por 100 de la población infantil no supera la puntuación 13; la puntuación pedida para la población adulta es 10,66. Obsérvese que, al ser la misma proporción en las dos poblaciones, se obtienen la misma puntuación típica, pero diferente puntuación directa.

16. La probabilidad de que a esa persona se le pronosticara un valor en O comprendido entre 60 y 67,5 es igual a 0,3413.

17. 0,8026.

18. 0,3645.

19. Nivel alto, aproximadamente 10 estudiantes.
Nivel medio, aproximadamente 133 estudiantes.
Nivel bajo, aproximadamente 58 estudiantes.

20. Nivel alto: puntuaciones superiores a 60,35.
Nivel medio: puntuaciones entre 39,65 y 60,35 puntos.
Nivel bajo: puntuaciones inferiores a 39,65.

21. *a*) 0,1635.
b) 0,8739.

22. *a*) $Y \sim \chi^2_{12}$.
b) 0,75; 0,99; 0,675.

23. *a*) ${}_{0,80}\chi^2_{12} = 15,812$.
b) ${}_{0,30}\chi^2_{12} = 9,034$.
c) ${}_{0,25}\chi^2_{12} = 8,438$ y ${}_{0,75}\chi^2_{12} = 14,845$.
d) ${}_{0,90}\chi^2_{12} = 18,549$.
e) ${}_{0,20}\chi^2_{12} = 7,807$.

24. 0,0262.

25. *a*) $E(X) = N \cdot \pi = 14$.
b) 0,0505.
c) 0,4404.
d) Deberían ser diez letras.

26. *a*) 0,8980.
b) 0,1660.

27. *a*) 0,0735.
b) 0,7224.

28. Con la binomial $P(X \geq 12) = 0{,}057$, y con la aproximación a la normal: $P(X \geq 12) = 0{,}0548$. Por tanto, la diferencia sería 0,0022.

29. *a*) 0,6915.
b) 32,46.
c) 36,12.
d) 0,3720.

30. 0,048 (redondeando a $\pi = 0{,}30$).

31. 0,8365.

32. *a*) ${}_{0,98}\chi^2_{16} = 29{,}633$.
b) 0,70.
c) 11,912 y 19,369.
d) ${}_{0,20}\chi^2_{16} = 11{,}152$.
e) 26,296.

33. *a*) 0,975.
b) 0,90.
c) 0,10.

34. *a*) 0,90.
b) 0,30.
c) 0,05.
d) 0,475.

35. *a*) −0,863 y 0,863.
b) 2,567.
c) −0,689.
d) 0.
e) 0,257.

36. *a*) 0,50.
b) 34,805.

37. El error de la aproximación es: 29,339 − 29,414 = −0,075.

38. *a*) ${}_{0,02}\chi^2_{21} = 9,915$.
b) ${}_{0,995}\chi^2_{7} = 20,278$.
c) ${}_{0,50}\chi^2_{4} = 3,357$.
d) ${}_{0,10}\chi^2_{27} = 18,114$.
e) ${}_{0,90}\chi^2_{14} = 21,064$.
f) ${}_{0,995}\chi^2_{4} = 14,860$.

39. No, ya que cualquier valor de una variable que siga una distribución *t* de Student, sean cuales sean sus grados de libertad, que acumule una probabilidad por la izquierda menor que 0,5, siempre tiene que ser negativo.

40. *a*) 0,50.
b) −0,868.
c) 0,01.

41. *a*) 0,80.
b) 0,65.

42. *a*) 0,99.
b) 0,75.
c) 0,10.

43. *a*) ${}_{0,10}\chi^2_{19} = 11,651$.
b) ${}_{0,999}\chi^2_{19} = 43,820$.
c) ${}_{0,05}\chi^2_{19} = 10,117$.
d) ${}_{0,01}\chi^2_{19} = 7,633$.
e) ${}_{0,99}\chi^2_{19} = 36,191$.
f) ${}_{0,95}\chi^2_{19} = 30,144$.

44. *a*) $V \sim \chi^2_{20}$.
b) $E(V) = 20$ y $\sigma(V) = 6,325$.
c) 0,02.

45. 0,162.

46. *a*) $P(X \leq 2{,}52) = 0{,}90$.
b) $P(X \leq 3{,}33) = 0{,}95$.
c) $P(3{,}33 \leq X \leq 5{,}64) = 0{,}04$.
d) ${}_{0{,}05}F_{5,10} = 0{,}211$.
e) ${}_{0{,}01}F_{5,10} = 0{,}100$.
f) ${}_{0{,}10}F_{5,10} = 0{,}303$.
g) $P(X \geq 1{,}59) = 0{,}25$.
h) $C_{25} = 0{,}529$.
i) $P(X \geq 4{,}24) = 0{,}025$.
j) $P(4{,}24 \leq X \leq 5{,}64) = 0{,}015$.
k) $P(X \geq 10{,}5) = 0{,}001$.
l) $P(X \leq 12{,}4) = 0{,}9995$.

47. *a*) $E(X) = 2$.
b) $\sigma^2(X) = 1$.
c) 0,6247.

48. *a*) En mujeres 0,4129.
b) En hombres 0,7673.

49.

k	Valor χ^2	Probabilidad acumulada
10	3,059	0,02
5	3,000	0,30
20	22,775	0,70
26	31,795	0,80
18	13,675	0,25
2	1,386	0,50

50.

k	Valor *t*	Probabilidad acumulada
10	1,812	0,95
5	−0,559	0,30
21	−1,323	0,10
25	−0,684	0,25
80	1,292	0,90
15	0	0,50

51.

m	*n*	Valor *F*	Probabilidad acumulada
2	5	1,85	0,75
4	6	0,942	0,50
10	9	2,42	0,90
2	12	0,106	0,10
10	10	0,336	0,05
60	24	1,02	0,50

52.

k	Probabilidad acumulada	Valor χ^2
9	0,70	10,656
5	0,25	2,675
2	0,90	4,605
16	0,50	15,338
18	0,10	10,865
20	0,01	8,260

53.

k	Probabilidad acumulada	Valor *t*
15	0,950	1,753
18	0,200	−0,862
2	0,300	−0,617
70	0,400	−0,254
5	0,975	2,571
100	0,500	0

54.

m	*n*	Probabilidad acumulada	Valor *F*
12	20	0,75	1,390
6	30	0,50	0,912
60	120	0,90	1,320
1	10	0,10	0,017
20	20	0,05	0,471
10	120	0,25	0,670

55. Aproximadamente 0,05.

56. 0,10.

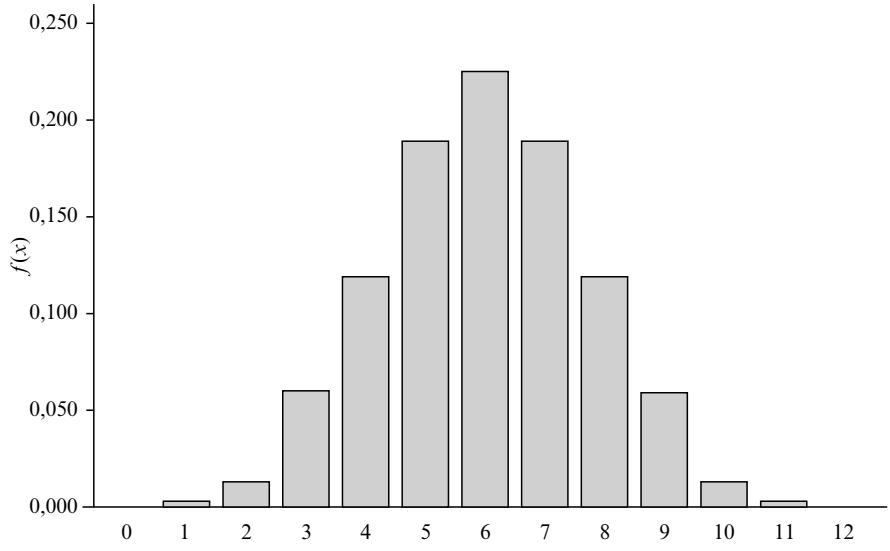
APÉNDICE

Procedimientos de aproximación

A veces ocurre que la obtención de un valor o de una probabilidad implican un cálculo laborioso y no se dispone de una tabla suficientemente amplia que abarque el valor buscado. Por ejemplo, no se pueden hacer tablas para cualquier valor de grados de libertad, dado que son infinitos. Para estos casos, con frecuencia se dispone de procedimientos que proporcionan aproximaciones a los valores o probabilidades buscados. En este apéndice incluimos dos de los más utilizados. En ambos casos se trata de aproximaciones a la distribución normal.

a) *Aproximación de la distribución binomial a la distribución normal*

Las probabilidades asociadas a valores de variables binomiales se pueden obtener mediante una aproximación a la normal. En la figura siguiente se aprecia que la curva que aparecería al unir los extremos superiores de las barras que representan las funciones de probabilidad de un modelo $B(12; 0{,}50)$ tienen un cierto parecido a esa distribución. Ese parecido es tanto mayor cuanto más simétrica sea la distribución y cuanto mayor sea el número de barras en las que se basa. Esto se traduce en que el valor de π no sea demasiado extremo ni el valor esperado de la distribución sea demasiado pequeño.



El procedimiento consiste en que, para calcular la probabilidad acumulada de un valor de una variable distribuida según el modelo binomial, se calcula la de ese mismo valor en una hipotética distribución normal en la que el valor espera-

do y la varianza fueran los mismos que los de la binomial original, pero haciendo una corrección por continuidad. Esta corrección permite mejorar la aproximación, asumiendo el hecho de que la binomial es discreta y la normal continua. La corrección por continuidad consiste en que no se tipifica el valor correspondiente, sino que se suma o resta media unidad para que el valor tipificado incluya las barras de los valores de interés de forma completa. Veámoslo con algunos ejemplos y comprobemos el grado de aproximación que se consigue.

Ejemplo. Supongamos una variable, X, distribuida según el modelo $B(12; 0{,}50)$. Queremos obtener las probabilidades de observar: *a*) valores como máximo iguales a 4; *b*) valores como mínimo iguales a 9, y *c*) valores comprendidos entre 4 y 7, ambos incluidos.

a) Se trata de calcular la probabilidad acumulada del valor 4, que según la tabla de la binomial es 0,194. El valor que debemos tipificar en la aproximación es el que deja por debajo todos los valores de interés, y por encima todos los demás, es decir, 4,5. Como el valor esperado y la varianza de la binomial son $N \cdot \pi$ y $N \cdot \pi \cdot (1 - \pi)$, sustituimos en la fórmula de tipificación:

$$P(X \leq 4) \approx P\left(z \leq \frac{4{,}5 - 12 \cdot 0{,}50}{\sqrt{12 \cdot 0{,}50 \cdot 0{,}50}}\right) = P(z \leq -0{,}87) = 0{,}1922$$

b) Se trata de la probabilidad combinada del valor 9 y los superiores. Al tipificar el valor 8,5 obtenemos la probabilidad asociada al conjunto de valores 0-8, cuyo complementario será la probabilidad buscada y que según la tabla de la binomial es $1 - 0{,}927 = 0{,}073$. Por tanto:

$$1 - P(X \leq 8) \approx 1 - P\left(z \leq \frac{8{,}5 - 12 \cdot 0{,}50}{\sqrt{12 \cdot 0{,}50 \cdot 0{,}50}}\right) = 1 - P(z \leq 1{,}44) =$$

$$= 1 - 0{,}9251 = 0{,}0749$$

c) Se trata de obtener por aproximación la probabilidad que resulta de la suma de las funciones de probabilidad de los valores 4, 5, 6 y 7, que según la tabla de la binomial es igual a 0,733. Para ello tipificamos los valores 3,5 y 7,5. Es decir:

$$P(4 \leq X \leq 7) \approx P\left(\frac{3{,}5 - 12 \cdot 0{,}50}{\sqrt{12 \cdot 0{,}50 \cdot 0{,}50}} \leq z \leq \frac{7{,}5 - 12 \cdot 0{,}50}{\sqrt{12 \cdot 0{,}50 \cdot 0{,}50}}\right) =$$

$$= P(-1{,}44 \leq z \leq 0{,}87) = F(0{,}87) - F(-1{,}44) = 0{,}8078 - 0{,}0749 = 0{,}7329$$

Como se puede apreciar, en los tres casos se obtienen probabilidades lo suficientemente cercanas a las de las tablas de la binomial. Esta aproximación es tanto mejor cuanto mayor es N y cuanto más cercano a 0,50 es π. Muchos autores ponen límites a los valores paramétricos para los que considera razonable la

aproximación. Así, es frecuente imponer las dos condiciones siguientes: *a*) que el valor esperado sea como mínimo igual a 5, y *b*) que π sea un valor comprendido entre 0,20 y 0,80.

b) *Aproximación de χ^2 a la distribución normal*

Esta aproximación permite obtener las áreas asociadas a la distribución χ^2 a partir de las de la distribución normal con bastante precisión, cuando el número de grados de libertad es moderadamente grande. En concreto, se demuestra que si T es una variable que se distribuye χ^2_k, entonces la expresión $\sqrt{2 \cdot T}$ se distribuye aproximadamente $N(\sqrt{2 \cdot k - 1}; 1)$, siendo esta aproximación tanto mejor cuanto mayor es el número de grados de libertad. Esta fórmula se suele expresar con la variable despejada:

$$_p\chi^2_k \approx (1/2) \cdot (z_p + \sqrt{2 \cdot k - 1})^2$$

Según esta fórmula, el valor de una distribución χ^2 con k grados de libertad que tiene una probabilidad acumulada igual a p se puede obtener por aproximación a partir del miembro derecho de la fórmula, en el que interviene el valor de la normal unitaria con esa misma función de distribución y los k grados de libertad de la distribución χ^2. Esta aproximación se suele considerar como suficientemente buena cuando los grados de libertad son mayores de 30.

Veamos algunos ejemplos de cómo se aplica. Supongamos que la variable T se distribuye según χ^2_{95} y queremos obtener el valor con probabilidad acumulada 0,10. Dado que los grados de libertad son mayores de 30, utilizamos la fórmula de aproximación; para ello tomamos de la tabla de la distribución normal unitaria el valor con esa misma probabilidad acumulada (−1,28) y sustituimos en la fórmula:

$$_{0,10}\chi^2_{95} \approx (1/2) \cdot (z_{0,10} + \sqrt{2 \cdot 95 - 1})^2 = (1/2) \cdot (-1,28 + \sqrt{189})^2 = 77,72$$

En ocasiones, lo que nos interesará será obtener las áreas asociadas a valores de distribuciones con unos grados de libertad superiores a 30. Supongamos que en este mismo ejemplo queremos obtener la probabilidad de observar un valor como mucho igual a 103,50 (o el área que queda a la izquierda de 103,50). Para ello, sustituimos en la fórmula y despejamos z_p:

$$_p\chi^2_{95} = 103,50 \approx (1/2) \cdot (z_p + \sqrt{2 \cdot 95 - 1})^2$$

despejando:

$$z_p \approx \sqrt{2 \cdot 103,5} - \sqrt{2 \cdot 95 - 1} = 14,3875 - 13,7477 = 0,64$$

De aquí deducimos que la probabilidad acumulada del valor 103,5 en la distribución χ^2_{95} es aproximadamente igual a la del valor 0,64 en la normal unitaria. Podemos obtener esa probabilidad de la tabla de ésta; en concreto:

$$z_p \approx 0,64 = z_{0,7389}$$

por tanto:

$$P(T \leq 103{,}50) \approx 0{,}7389$$

Podemos comprobar el grado de aproximación que se consigue con esta fórmula aplicándola a un caso límite. En la tabla podemos observar que el valor de la distribución χ^2_{30} que tiene un área izquierda igual a 0,90 es 40,256. Veamos ahora qué valor hubiéramos obtenido con la fórmula de aproximación:

$$_{0{,}90}\chi^2_{30} \approx (1/2) \cdot (1{,}28 + \sqrt{2 \cdot 30 - 1})^2 = 40{,}151$$

Como se puede apreciar, la aproximación es ya bastante buena con 30 grados de libertad, por lo que para mayores grados de libertad podemos aceptar con confianza la aproximación que nos proporciona la fórmula.

PARTE CUARTA

Introducción a la inferencia estadística

Distribución muestral de un estadístico 13

13.1. INTRODUCCIÓN

La estadística inferencial trata, como su propio nombre indica, sobre las inferencias que se realizan con respecto a poblaciones o a variables aleatorias, a partir de la información contenida en las muestras. Uno de sus objetivos principales es el de establecer conclusiones relativas a los parámetros poblacionales a partir de los estadísticos muestrales. Para ello es imprescindible conocer la relación entre ambos, es decir, determinar si se puede establecer una relación entre las características poblacionales y el comportamiento de los estadísticos muestrales correspondientes a esas características. En este capítulo abordaremos el concepto esencial que permite establecer esa relación, conocido como «distribución muestral de un estadístico». Nos centraremos sobre todo en la media aritmética, en la que aprovecharemos para exponer las propiedades y características de este concepto. Después lo extenderemos a otros dos estadísticos, la correlación de Pearson y la proporción. Pero, antes de nada, vamos a retomar el concepto de muestra aleatoria simple, imprescindible para abordar la inferencia estadística.

13.2. MUESTREO ALEATORIO SIMPLE

En el apartado 10.6 habíamos dedicado unas líneas al muestreo aleatorio y habíamos definido lo que se conoce como *muestra aleatoria simple* (m.a.s.). Este concepto es de gran importancia para la estadística inferencial, dado que la aplicación correcta de prácticamente todos sus procedimientos exige que los datos sean obtenidos en estas condiciones. Por ello, vamos a repasarlo aquí y a insistir en su definición.

La idea es considerar que, cuando nos disponemos a extraer una muestra aleatoria de N observaciones, cada una de las observaciones que van a componer esa muestra es una variable aleatoria (valores X_1, X_2,…, X_N). Como tales variables aleatorias, tienen su valor esperado, su varianza y su distribución de probabilidad (o de densidad de probabilidad). Pues bien, ese conjunto de N valores será una muestra aleatoria simple si esas N variables aleatorias son independientes y tienen

la misma distribución. Esto significa que las expectativas respecto a los valores de cada observación no cambian en función de los valores observados en las demás. Repetiremos aquí el cuadro que presentábamos en el capítulo 10 para resaltar su definición.

> Una *muestra aleatoria simple* (m.a.s.) compuesta por N elementos es una secuencia de N variables aleatorias, independientes e igualmente distribuidas. Es decir, si representamos por X_k al elemento extraído en k-ésimo lugar y por X_{k+1} al elemento extraído en el lugar siguiente a aquél, entonces:
>
> $$P(X_{k+1} = x_i) = P(X_k = x_i) \qquad \text{para todo } k \text{ y } x_i$$

Los procedimientos de extracción al azar (como por ejemplo las urnas o los procedimientos informáticos de generación de números aleatorios) son una garantía razonable de que las N variables que componen la muestra sean independientes. En cambio, para que tengan la misma distribución es necesario que la población sea infinita o que las extracciones se realicen con reposición de los elementos ya extraídos. En la práctica, como en psicología estudiamos sobre todo procesos y los parámetros representan propensiones (véase el apartado 14.5.4), no tenemos problemas con los tamaños de las poblaciones.

13.3. LA DISTRIBUCIÓN MUESTRAL DE UN ESTADÍSTICO

Supongamos la extracción de N observaciones de una variable aleatoria y el cálculo con ellas de un determinado estadístico (por ejemplo, la media aritmética). Nuestro objetivo principal en este capítulo es determinar las probabilidades asociadas a los posibles valores que puede adoptar ese estadístico en la muestra. Si extraemos infinitas muestras del mismo tamaño, N, en las mismas condiciones, y calculamos en cada una de ellas el estadístico al que nos referimos, entonces podemos considerar al propio estadístico como una variable aleatoria, con su propia distribución de probabilidad. Si somos capaces de establecer cuál es esa distribución, entonces podremos calcular probabilidades asociadas a los posibles valores de ese estadístico. La distribución que empareja posibles valores de un estadístico con probabilidades de verificación recibe el nombre de *distribución muestral* de un estadístico. Cada uno de los estadísticos que hemos estudiado a lo largo de este libro tiene su propia distribución muestral; muchos de ellos se ajustarán a uno de los modelos de distribución de probabilidad que hemos estudiado en el capítulo anterior. Para cualquier estadístico, T, su distribución muestral es la función de probabilidad (o función de densidad de probabilidad) que relaciona cada posible valor de T con su probabilidad, $f(t)$. La desviación típica de la distribución muestral de un estadístico recibe el nom-

bre de *error típico*. Podemos, por tanto, definir este concepto de la siguiente forma:

> Se llama distribución muestral de un estadístico, para muestras de tamaño N, a una «distribución de probabilidad teórica» de los valores de ese estadístico cuando éstos se calculan sobre muestras aleatorias simples de tamaño N, extraídas de la población de referencia.

Aunque de esta definición se desprende que existe una distribución muestral para cada posible tamaño muestral, éstas suelen expresarse de manera compacta, para cualquier valor de N.

13.4. DISTRIBUCIÓN MUESTRAL DE LA MEDIA

La media es, sin duda, el parámetro de mayor interés, como indicador de tendencia central. Tal y como veremos en capítulos posteriores, uno de los usos más frecuentes de la estadística aplicada en psicología consiste en contrastar, mediante unos datos muestrales, hipótesis relativas a los valores de las medias poblacionales. Para poder hacerlo es imprescindible conocer la *distribución muestral de la media*. Para mostrar cómo se comporta este estadístico y exponer las características de su distribución muestral, recordaremos que la media aritmética se puede expresar como una suma ponderada de los valores de la muestra; es decir, como:

$$\bar{X} = \frac{1}{N} \cdot X_1 + \frac{1}{N} \cdot X_2 + ... + \frac{1}{N} \cdot X_N \qquad [13.1]$$

De esta forma será más fácil deducir las características de su distribución muestral. Vamos a exponer dos casos: cuando la variable aleatoria se ajusta al modelo normal y cuando no es así.

13.4.1. La variable se distribuye según el modelo normal

Supongamos que trabajamos con una variable aleatoria que se distribuye $N(\mu; \sigma)$ y que extraemos muestras aleatorias simples de N observaciones. Para mostrar cuál es en este caso la distribución muestral de la media, vamos a detenernos en tres de sus propiedades: el valor esperado, la varianza (y el error típico) y la forma de la distribución, aunque sólo demostraremos las dos primeras. Lo que vamos a explicar sobre el valor esperado y la varianza no exige que se cumpla la normalidad (es válido también para el apartado 13.4.2, en el que ésta no es una condición necesaria). Las normalidad sólo es relevante para la tercera de estas propiedades: la forma de la distribución.

a) *Valor esperado*. Si extraemos una m.a.s. de tamaño N, como las observaciones son por definición independientes, cada una de ellas tiene el mismo valor esperado, μ. Por otro lado, sabemos que el valor esperado de una combinación lineal de variables es igual a la misma combinación lineal de sus valores esperados (apartado 10.3.3). Aplicando esto a la expresión [13.1]:

$$E(\bar{X}) = \frac{1}{N} \cdot E(X_1) + \frac{1}{N} \cdot E(X_2) + ... + \frac{1}{N} E(X_N) =$$

$$= \frac{1}{N} \cdot E(X_i) + \frac{1}{N} \cdot E(X_i) + ... + \frac{1}{N} \cdot E(X_i)$$

Como se trata de una suma de N expresiones idénticas:

$$E(\bar{X}) = N \cdot \frac{1}{N} \cdot E(X_i) = E(X_i) = \mu \qquad [13.2]$$

b) *Varianza.* Recordemos que la varianza de una suma de variables no es igual a la simple suma de sus varianzas, sino que hay que sumarle las covarianzas multiplicadas por 2 (véase en el apartado 10.3.3). Sin embargo, en este caso hemos partido de que las observaciones son independientes (es una m.a.s.), por lo que las covarianzas son nulas. En resumen, aplicando a [13.1] las propiedades que conocemos de la varianza de variables aleatorias (véase también en 10.3.3), la varianza de cada sumando será igual a $(1/N)^2 \cdot \sigma^2$. Por tanto, la varianza de la expresión [13.1] será igual a la suma de N veces esa expresión:

$$\sigma^2(\bar{X}) = \frac{1}{N^2} \cdot \sigma^2 + \frac{1}{N^2} \cdot \sigma^2 + ... + \frac{1}{N^2} \cdot \sigma^2 = \frac{N}{N^2} \cdot \sigma^2 = \frac{\sigma^2}{N} \qquad [13.3]$$

Naturalmente, su raíz cuadrada (la desviación típica) es lo que hemos llamado *error típico de la media:*

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{N}}$$

c) *Distribución.* Una variable aleatoria que se define como la combinación lineal de variables aleatorias independientes que se ajusten al modelo normal, también se distribuye según el modelo normal. La demostración de esta afirmación excede el alcance de este libro (véase en Amón, 1996).

Estos tres elementos nos permiten establecer la distribución muestral de la media en el caso particular de las condiciones especificadas, que resumimos en el cuadro 13.1.

Si se cumplen las condiciones especificadas en el cuadro 13.1, la distribución muestral de la media es $N(\mu; \sigma/\sqrt{N})$. Esto nos va a permitir hacer cálculos relativos a las probabilidades asociadas a los valores de la media, en las condiciones conocidas, mediante la aplicación de la regla de tipificación expuesta en el capítulo 12. Todo ello se ilustra en el cuadro 13.2.

CUADRO 13.1

Distribución muestral de la media

Si *a*) la variable X se distribuye $N(\mu; \sigma)$,
b) extraemos muestras aleatorias simples de tamaño N, y
c) calculamos en cada muestra la media aritmética de los N valores,

entonces, los valores de esas medias aritméticas constituyen una variable aleatoria que se distribuye $N(\mu; \sigma/\sqrt{N})$.

CUADRO 13.2

Ejemplos de la distribución muestral de la media

Supongamos que la estatura de los varones españoles se distribuye $N(175; 8)$. Si extraemos una muestra aleatoria simple de 16 varones españoles, calcule las probabilidades de que su media aritmética sea: *a*) menor de 170; *b*) mayor de 174, y *c*) comprendida entre 177 y 181.
Dadas las condiciones, la media aritmética de una m.a.s. de tamaño 16 se distribuye $N(175; 8/\sqrt{16})$. En consecuencia, la expresión:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{\bar{X} - 175}{8/\sqrt{16}} \qquad \text{se distribuye } N(0; 1)$$

Por tanto, las probabilidades pedidas son:

a)

$$P(\bar{X} \leq 170) = P\left(z \leq \frac{170 - 175}{8/\sqrt{16}}\right) = P(z \leq -2{,}50) = 0{,}0062$$

b)

$$P(\bar{X} \geq 174) = P\left(z \geq \frac{174 - 175}{8/\sqrt{16}}\right) = 1 - P(z \leq -0{,}50) = 1 - 0{,}3085 = 0{,}6915$$

c)

$$P(177 \leq \bar{X} \leq 181) = P\left(\frac{177 - 175}{8/\sqrt{16}} \leq z \leq \frac{181 - 175}{8/\sqrt{16}}\right) = P(1 \leq z \leq 3) = F(3) - F(1) =$$
$$= 0{,}9987 - 0{,}8413 = 0{,}1574$$

Obtenga también las probabilidades de obtener una media menor de 173 en los casos de que la muestra fuera de 25, 49, 100 y 144 sujetos.
Lo que cambia en cada uno de estos casos es el valor del error típico de la media:

$$P(\bar{X} \leq 173) = P\left(z \leq \frac{173 - 175}{8/\sqrt{25}}\right) = P(z \leq -1{,}25) = 0{,}1056$$

CUADRO 13.2 *(continuación)*

$$P(\bar{X} \leq 173) = P\left(z \leq \frac{173 - 175}{8/\sqrt{49}}\right) = P(z \leq -1{,}75) = 0{,}0401$$

$$P(\bar{X} \leq 173) = P\left(z \leq \frac{173 - 175}{8/\sqrt{100}}\right) = P(z \leq -2{,}50) = 0{,}0062$$

$$P(\bar{X} \leq 173) = P\left(z \leq \frac{173 - 175}{8/\sqrt{144}}\right) = P(z \leq -3{,}00) = 0{,}0013$$

Tal y como se muestra en el segundo ejemplo del cuadro 13.2, una de las características más importantes de la distribución muestral de la media es que su error típico depende del tamaño muestral, N. Si observamos la fórmula [13.3] advertimos que lo único que cambia al aumentar N es la desviación típica de la distribución muestral (el error típico de la media); es decir, a medida que aumenta N, la distribución se va haciendo más y más homogénea. Veamos lo que ocurrirá en los casos extremos, cuando N alcanza sus valores mínimo y máximo.

El mínimo posible valor de N es 1 (la muestra está integrada por un único valor). En este caso, la media muestral no es otra cosa que ese mismo único valor y, por tanto, la expresión $N(\mu; \sigma/\sqrt{N})$ queda reducida a $N(\mu; \sigma)$, que es la distribución de la variable original. Por tanto, se puede decir que la distribución de una variable aleatoria no es otra cosa que la distribución muestral de las medias que se obtendrían si esas muestras fueran de tamaño 1.

Por otro lado, el máximo valor posible de N se produce cuando la muestra es tan grande como la población (o infinita). En este caso, la media de la muestra sería la misma que la de la población, puesto que muestra y población coinciden. Pero, además, en cada extracción obtendríamos la misma muestra (toda la población), con la misma media; por tanto, no habría variación en los valores de media obtenidos. El error típico de la media sería cero.

En resumen, la variabilidad de la distribución muestral de la media depende del tamaño de las muestras a las que se refiera, pues el error típico de las medias es una función inversa de N. Sus valores máximo y mínimo son σ y 0, respectivamente.

13.4.2. La variable no se distribuye según el modelo normal

La situación expuesta en el apartado anterior es la más sencilla, pero, desafortunadamente, no es nada frecuente. Es mucho más habitual que la variable no se ajuste al modelo normal o, sencillamente, se desconozca la distribución de la variable. En estos casos se puede recurrir a un importante teorema de la estadística, conocido como teorema central del límite (o teorema del límite central), que establece lo siguiente:

Independientemente de cómo sea la forma de la distribución de la variable, la distribución muestral de la media se parece más y más al modelo normal a medida que aumenta el tamaño de las muestras (N) sobre las que se calcula, y tiende a ella cuando el tamaño de las muestras tiende a infinito.

Gracias a este teorema podemos calcular con bastante aproximación las probabilidades asociadas a los valores de las medias sin conocer la forma de la distribución de la variable, siempre y cuando las muestras sean lo suficientemente grandes como para poder asumir que el parecido de la distribución muestral de la media con la distribución normal es suficientemente grande. En muchos manuales se considera que esta aproximación es suficientemente buena cuando N es igual o mayor que 30 (cuadro 13.3).

CUADRO 13.3

Aplicación del teorema central del límite

Supongamos que los pesos de los recién nacidos tienen una media igual a 3,600 y una desviación típica igual a 0,250. Calcule las probabilidades de que una m.a.s. de 36 recién nacidos tenga un peso medio: *a*) superior a 3,500; *b*) inferior a 3,650, y *c*) comprendido entre 3,510 y 3,580.

Aunque no sabemos cómo se distribuye la variable original, como la muestra es suficientemente grande el teorema central del límite nos permite deducir con bastante aproximación la distribución muestral de la media. En concreto:

Si *a*) $E(X) = 3{,}600$ y $\sigma(X) = 0{,}250$

b) nos referimos a muestras aleatorias simples de tamaño $N = 36$,

entonces, $\bar{X}$ se aproxima a $N(3{,}600;\ 0{,}250/\sqrt{36})$

Por tanto, las probabilidades pedidas son:

a)

$$P(\bar{X} > 3{,}500) = 1 - P\left(z \leq \frac{3{,}500 - 3{,}600}{0{,}250/\sqrt{36}}\right) = 1 - P(z \leq -2{,}40) = 1 - 0{,}0082 = 0{,}9918$$

b)

$$P(\bar{X} \leq 3{,}650) = P\left(z \leq \frac{3{,}650 - 3{,}600}{0{,}250/\sqrt{36}}\right) = P(z \leq 1{,}20) = 0{,}8849$$

c)

$$P(3{,}510 \leq \bar{X} \leq 3{,}580) = P\left(\frac{3{,}510 - 3{,}600}{0{,}250/\sqrt{36}} \leq z \leq \frac{3{,}580 - 3{,}600}{0{,}250/\sqrt{36}}\right) = P(-2{,}16 \leq z \leq -0{,}48) =$$
$$= 0{,}3121 - 0{,}0154 = 0{,}2967$$

13.5. DISTRIBUCIÓN MUESTRAL DE LA CORRELACIÓN

Supongamos dos variables cuantitativas linealmente independientes ($\rho = 0$; recordemos que, igual que los parámetros media y varianza se representan por μ y σ, la correlación poblacional se representa por la letra griega «rho»), como por ejemplo la estatura y la inteligencia. Que estas variables sean linealmente independientes significa que el parámetro (ρ) es igual a 0, pero no significa que en una muestra limitada de individuos la correlación muestral (r_{xy}) tenga que ser aritméticamente igual a 0. De hecho, la distribución de valores de r_{xy} que se obtendrían con un determinado tamaño muestral, N, si se extrajeran aleatoriamente, constituye la *distribución muestral de la correlación*. Siendo la correlación poblacional igual a 0, hay que esperar que las correlaciones muestrales se sitúen en torno a ese valor, pero no tienen por qué ser exactamente iguales al parámetro.

Aquí vamos a exponer sólo el caso en que la correlación paramétrica es nula ($\rho = 0$) y ambas variables se ajustan al modelo normal. En el cuadro 13.4 se resumen esas condiciones y el estadístico (distribuido según t de Student con $N - 2$ grados de libertad) que permite calcular las probabilidades asociadas a la correlación muestral.

CUADRO 13.4

Distribución muestral de la correlación

Si
a) las variables aleatorias X e Y se ajustan al modelo normal,
b) X e Y son linealmente independientes ($\rho = 0$), y
c) se extraen muestras aleatorias simples de tamaño N,

entonces, el estadístico $T = \dfrac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r_{xy}^2}}$ se distribuye t_{N-2} [13.4]

La fórmula [13.4] permite obtener las probabilidades asociadas a valores concretos de la correlación muestral, r_{xy}, tal y como se ilustra en el cuadro 13.5.

CUADRO 13.5

Ejemplos numéricos de la distribución muestral de la correlación

Supongamos que en la población española la estatura y la inteligencia son linealmente independientes. Si extraemos una m.a.s. de 50 individuos y medimos en cada uno su estatura y su inteligencia, calcule las probabilidades de que la correlación entre esas variables en la muestra sea: *a*) menor de 0,229; *b*) mayor de 0,037, y *c*) comprendida entre −0,269 y −0,040.

Dadas las condiciones, el estadístico [13.4] se distribuye t_{48}. Por tanto, las probabilidades pedidas son:

CUADRO 13.5 *(continuación)*

a)

$$P(r_{xy} \leq 0{,}229) = P\left(T \leq \frac{0{,}229 \cdot \sqrt{48}}{\sqrt{1 - 0{,}229^2}}\right) = P(T \leq 1{,}630) = 0{,}95$$

b)

$$P(r_{xy} \geq 0{,}037) = P\left(T \geq \frac{0{,}037 \cdot \sqrt{48}}{\sqrt{1 - 0{,}037^2}}\right) = 1 - P(T \leq 0{,}257) = 1 - 0{,}60 = 0{,}40$$

c)

$$P(-0{,}269 \leq r_{xy} \leq -0{,}040) = P\,\frac{-0{,}269 \cdot \sqrt{48}}{\sqrt{1 - (-0{,}269)^2}} \leq T \leq \frac{-0{,}040 \cdot \sqrt{48}}{\sqrt{1 - (-0{,}040)^2}}\Bigg) =$$

$$= P(-1{,}935 \leq T \leq -0{,}277) = F(-0{,}277) - F(-1{,}935) = 0{,}400 - 0{,}025 = 0{,}375$$

13.6. DISTRIBUCIÓN MUESTRAL DE LA PROPORCIÓN

También resulta de gran utilidad para la inferencia estadística estudiar cómo se distribuye la proporción de N observaciones independientes que cumplen una condición especificada, o *distribución muestral de la proporción.*

El número de casos, dentro de una muestra de N observaciones, que cumplen una determinada condición se distribuye según el modelo binomial (capítulo 11). Sin embargo, en la mayoría de los casos los tamaños muestrales permitirán emplear el procedimiento de aproximación al modelo normal que hemos descrito en el capítulo 12. Vamos a describir primero el procedimiento de la binomial, y luego abordaremos el procedimiento de aproximación que se aplica con muestras grandes.

La proporción, P, de casos que cumplen una condición no es más que una transformación simple del número de casos, dividiendo la frecuencia, X, por el número de observaciones ($P_i = X_i/N$), por lo que sus probabilidades son las mismas que las de las frecuencias absolutas. Veamos un ejemplo. El número de caras que se pueden obtener en ocho lanzamientos de una moneda imparcial es una variable aleatoria que se distribuye B(8; 0,50). Su distribución se resume en la primera de las dos tablas siguientes:

X_i	0	1	2	3	4	5	6	7	8
$f(x_i)$	0,004	0,031	0,109	0,219	0,273	0,219	0,109	0,031	0,004

P_i	0/8	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
$f(P_i)$	0,004	0,031	0,109	0,219	0,273	0,219	0,109	0,031	0,004

Las probabilidades asociadas a los valores no cambian si esos valores se transforman en valores de proporción, P, o frecuencias relativas (0/8, 1/8, ...,8/8), que

es lo que hemos hecho en la segunda de las tablas anteriores. Por ejemplo, la probabilidad de obtener una frecuencia de 3 es la misma que la de obtener una proporción de 0,375 (3/8). Por tanto, podemos decir que las probabilidades asociadas a las proporciones se ajustan a un modelo binomial (cuadro 13.6).

CUADRO 13.6

Distribución muestral de la proporción

Si	*a*)	la probabilidad de que al hacer una observación se verifique una determinada condición es igual a π,
	b)	se realizan N de esas observaciones, de forma independiente, y
	c)	se calcula la proporción (P) de esas N observaciones que verifican la condición,
entonces,		la variable aleatoria P_i se distribuye $B(N; \pi)$, con las probabilidades correspondientes a X_i.

En el cuadro 13.7 se muestran ejemplos de la obtención de probabilidades asociadas a la proporción.

CUADRO 13.7

Ejemplos de la distribución muestral de la proporción

Supongamos que en la población española la proporción de las personas que tienen grupo sanguíneo 0 es igual a 0,40. Si extraemos una m.a.s. de 15 individuos, calcule las probabilidades de que la proporción de los que en esa muestra tengan el grupo sanguíneo 0 sea: *a*) menor o igual que 0,50; *b*) mayor de 0,25, y *c*) comprendida entre 0,60 y 0,80.
Dadas las condiciones, las probabilidades asociadas a la proporción son las que proporciona el modelo $B(15; 0{,}40)$. Por tanto, las probabilidades pedidas (según la tabla del modelo binomial) son:

a) Una proporción de 0,50 implica en este caso $0{,}50 \cdot 15 = 7{,}5$ casos; como el número de casos no puede adoptar valores decimales, la probabilidad es la misma que la de obtener 7 casos o menos.

$$P(P \leq 0{,}50) = P(X \leq 7) = 0{,}787$$

b) La proporción 0,25 implica una frecuencia de $0{,}025 \cdot 15 = 3{,}75$. De nuevo se trata de una frecuencia con decimales. Por tanto, una proporción superior a 0,25 implica una frecuencia igual o superior a 4.

$$P(P > 0{,}25) = P(X \geq 4) = 0{,}910$$

c) Las frecuencias correspondientes a esas proporciones son $0{,}60 \cdot 15 = 9$ y $0{,}80 \cdot 15 = 12$; en este caso se trata de valores enteros.

$$P(0{,}60 \leq P \leq 0{,}80) = P(9 \leq X \leq 12) = 0{,}061 + 0{,}024 + 0{,}007 + 0{,}002 = 0{,}094$$

13.6.1. Distribución muestral de la proporción con muestras grandes

En la situación expuesta hasta aquí se emplea la distribución binomial, pero, tal y como ya hemos señalado en el apéndice del capítulo anterior, cuando N es grande los cálculos con esa distribución resultan muy laboriosos. Por ello se suele emplear la aproximación de la binomial al modelo normal.

Sabemos que el valor esperado y la varianza de una variable binomial, X, son iguales a:

$$E(X) = N \cdot \pi \qquad \text{y} \qquad \sigma^2(X) = N \cdot \pi \cdot (1 - \pi)$$

Como la proporción, P, no es más que la frecuencia dividida por el tamaño total ($P_i = X_i/N$), el valor esperado y la varianza de P son iguales a:

$$E(P) = \pi \qquad \text{y} \qquad \sigma^2(P) = \frac{\pi \cdot (1 - \pi)}{N}$$

El teorema central del límite nos asegura que, a medida que aumenta N, las probabilidades que proporciona el modelo binomial se parecen más a las que proporcionaría un modelo normal con esos mismos valores paramétricos. Es decir:

$$z = \frac{P - E(P)}{\sigma(P)}$$

se distribuye aproximadamente $N(0; 1)$. Aplicar esta expresión implica conocer π, pero como suele ser desconocida, se sustituye por su estimación puntual (P). Por tanto, la fórmula que permite hacer los cálculos correspondientes a la distribución muestral de la proporción es:

$$z = \frac{P - \pi}{\sqrt{\pi \cdot (1 - \pi)/N}} \qquad [13.5]$$

Una limitación adicional es que mientras la distribución binomial describe las probabilidades de variables discretas, la distribución normal describe las de variables continuas. Los cálculos de las probabilidades por este procedimiento son más precisos si se hace la llamada *corrección por continuidad* (véase el apéndice del capítulo 12). Para ello hay que corregir la fórmula [13.5] de la forma como aparece en la fórmula [13.6]. En el cuadro 13.8 presentamos ejemplos de cálculos con este procedimiento.

$$z = \frac{\left(P \pm \frac{0{,}5}{N}\right) - \pi}{\sqrt{\pi \cdot (1 - \pi)/N}} \qquad [13.6]$$

CUADRO 13.8

Ejemplos de la distribución muestral de la proporción con muestras grandes

Supongamos que en la población española el 40 por 100 de los hogares disponen de conexión a Internet. Si accedemos a una m.a.s. de 120 hogares, calcule las probabilidades de que el porcentaje de los que tienen conexión a Internet en esa muestra sea: *a*) menor o igual que el 50 por 100; *b*) al menos el 45 por 100, y *c*) comprendido entre el 40 y el 48 por 100. Dadas las condiciones, las probabilidades asociadas a la proporción son las que proporcionaría el modelo $B(120; 0{,}40)$. Dado que N es grande y π no es extremo, aplicamos el procedimiento de aproximación a la distribución normal.
Los porcentajes solicitados se expresan en proporciones dividiéndolos por 100.

a)

$$P(P \leq 0{,}50) = P\left(z \leq \frac{\left(0{,}50 + \frac{0{,}5}{120}\right) - 0{,}40}{\sqrt{0{,}40 \cdot 0{,}60/120}}\right) = P(z \leq 2{,}33) = 0{,}9901$$

b)

$$P(P \geq 0{,}45) = 1 - P\left(z \leq \frac{\left(0{,}45 - \frac{0{,}5}{120}\right) - 0{,}40}{\sqrt{0{,}40 \cdot 0{,}60/120}}\right) = 1 - P(z \leq 1{,}02) =$$

$$= 1 - 0{,}8461 = 0{,}1539$$

c)

$$P(0{,}40 \leq P \leq 0{,}48) = P\left(\frac{\left(0{,}40 - \frac{0{,}5}{120}\right) - 0{,}40}{\sqrt{0{,}40 \cdot 0{,}60/120}} \leq z \leq \frac{\left(0{,}48 + \frac{0{,}5}{120}\right) - 0{,}40}{\sqrt{0{,}40 \cdot 0{,}60/120}}\right) =$$

$$= P(-0{,}09 \leq z \leq 1{,}88) = 0{,}9699 - 0{,}4641 = 0{,}5058$$

PROBLEMAS Y EJERCICIOS

1. Si sabemos que las puntuaciones obtenidas en *CI* a partir de la *escala Wechsler de inteligencia* para adultos se distribuyen según el modelo $N(100; 15)$, obtenga las probabilidades de que en una m.a.s. la media aritmética en *CI* sea menor o igual a 103 si la muestra es de cada uno de los siguientes tamaños: 4, 9, 25, 49 y 81.

2. Continuando con el ejercicio anterior, indique qué tamaño de la muestra habría que extraer para que la probabilidad de que su media aritmética sea mayor de 105 sea igual a 0,0475.

3. Si extraemos una m.a.s. de 25 estudiantes de la población de estudiantes matriculados en el primer curso de grado en 2011 de una facultad de psicología donde la variable *edad* (X) se distribuye $N(18; 1{,}25)$, indique cuál es el valor de la media en edad, tal que la probabilidad de obtener en la muestra un valor al menos tan alejado por encima de la media poblacional como él sea de 0,0516.

4. Los *pesos* de una población de jugadoras de baloncesto femenino se distribuyen $N(67; 6)$. Si las instrucciones del ascensor de los vestuarios prohíben subir a más de cuatro personas, calcule la probabilidad de que una m.a.s. de 4 jugadoras de esa población supere los 300 kilos máximos que el ascensor tiene de tolerancia.

5. Si sabemos que la probabilidad de que una m.a.s. de 35 personas tenga una media menor o igual que 23 es 0,1020 y que la varianza poblacional es 25, ¿cuánto valdrá la media poblacional?

6. Si en una población la variable X se distribuye $N(23; 8)$ y se extrae una m.a.s. de 4 personas, obtenga la probabilidad de que la media aritmética se separe en 3 o más puntos de la media poblacional.

7. En una población de estudiantes de secundaria, de los cuales un 40 por 100 estudian en centros bilingües (español-inglés), se extrae una m.a.s. de 15 estudiantes. Según lo anterior:

a) ¿Cuál es la probabilidad de que un 80 por 100 de la muestra estudie en centros bilingües?

b) ¿Cuál es la probabilidad de que al menos el 40 por 100 de la muestra estudie en centros bilingües?

8. Disponemos de los datos de un organismo internacional sobre la pobreza durante el año 2010 en un país de África, la cual se sitúa en un 40 por 100. Si tomamos una m.a.s. de 200 ciudadanos de ese país, ¿cuál es la probabilidad de que al menos el 50 por 100 de la muestra sean pobres?

9. Sabiendo que en la población humana las variables *tamaño del cerebro* (X) e *inteligencia* (Y) siguen el modelo normal y correlacionan 0, indique cuál es la probabilidad de que al extraer una m.a.s. de 10 personas la correlación de Pearson sea menor de 0,70.

10. Continuando con el ejercicio anterior, indique cuál es el valor de la correlación de Pearson al que le corresponde una probabilidad acumulada de 0,90.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. Como $X \sim N(100; 15)$, entonces $\bar{X} \sim N(100; 15/\sqrt{N})$. Por tanto, para $N = 4$: 0,6554; para $N = 9$: 0,7257; para $N = 25$: 0,8413; para $N = 49$: 0,9192; para $N = 81$: 0,9641.

2. $N = 25$.

3. $\bar{X}_i = 18{,}41$.

4. 0,0038.

5. $\mu = 24{,}07$.

6. 0,4532.

7. *a*) 0,002.
b) 0,597.

8. 0,0025.

9. Algo menor de 0,99.

10. $r_{xy} = 0{,}44$.

La lógica del contraste de hipótesis 14

14.1. INTRODUCCIÓN

Uno de los usos más extendidos de la estadística es el que se refiere a las técnicas de análisis de datos que se emplean en la investigación científica. En este capítulo vamos a describir el procedimiento más empleado, que se conoce como *contraste de hipótesis* (CH), aunque no está exento de críticas (Borges, San Luis, Sánchez-Bruno y Cañadas, 2001; Gigerenzer, 1998; Harlow, Mulaik y Steiger, 1997; Howard, Maxwell y Fleming, 2000; Nickerson, 2000). El CH es un procedimiento para adoptar decisiones categoriales respecto a la plausibilidad de una afirmación; estas afirmaciones son las hipótesis que se quieren estudiar.

14.2. VALORANDO LA EVIDENCIA

Para comprender lo que subyace en un CH conviene señalar antes unos elementos de lógica, cuya comprensión facilitará lo que sigue. Entre los enfoques que se emplean para valorar la evidencia empírica se puede distinguir entre los esquemas confirmatorios y los esquemas falsacionistas. Veamos un ejemplo. Supongamos que tenemos dos monedas que, por comodidad, designaremos como *A* y *B*. Estamos interesados en estudiar si son monedas realmente imparciales, para su uso en juegos de cara o cruz. Lanzamos estas monedas 100 veces y obtenemos unas frecuencias de caras/cruces de 50/50 y 90/10, respectivamente. Estas observaciones constituyen la evidencia empírica con la que vamos a razonar respecto a su imparcialidad. Hay dos opciones respecto a la imparcialidad que, por su propia naturaleza, son mutuamente excluyentes y exhaustivas: que la moneda sea imparcial y que no sea imparcial. Estas dos opciones se pueden traducir fácilmente a hipótesis estadísticas. La hipótesis de imparcialidad se traduce en que la probabilidad de cara (y de cruz) es 0,50, y la de no imparcialidad en que la probabilidad de cara (y de cruz) es distinta de 0,50.

Un enfoque confirmatorio intentaría afianzar la credibilidad de una hipótesis buscando evidencia empírica compatible con ella, mientras que un enfoque falsacionista trataría de reducir la credibilidad de una hipótesis buscando evidencia empírica incompatible con ella. Por ejemplo, desde un enfoque confirmatorio una

persona podría decir que los resultados de la moneda *A* avalan la idea de que se trata de una moneda imparcial, dado que las proporciones de caras y cruces observadas (0,50/0,50) son iguales a las esperadas si realmente fuera una moneda imparcial. Desde un enfoque falsacionista, una persona podría decir que los resultados de la moneda *B* le llevan a creer que se trata de una moneda que no es imparcial, dado que las proporciones de caras y cruces observadas (0,90/0,10) se desvían mucho de lo que esperaríamos si lo fuera.

En pocas palabras, un enfoque confirmatorio concluye que una hipótesis es creíble porque lo observado es compatible con ella, mientras que un enfoque falsacionista concluye que una hipótesis no es creíble (probablemente es falsa) porque lo observado es incompatible (en realidad «poco compatible») con ella.

Esta primera exposición necesita muchas aclaraciones que en breve vamos a abordar, pero no sin antes adelantar que el uso del CH, tal y como se emplea mayoritariamente en la ciencia, se basa en un enfoque falsacionista. Hay un gran acuerdo en que este enfoque es mucho más potente, aunque tenga también algunas limitaciones. Por muchas ocasiones en que observemos evidencia compatible con una hipótesis, nunca podremos estar razonablemente seguros de que es correcta. En cambio, como la evidencia incompatible con una hipótesis es en teoría imposible de ser observada, el hecho de que se haya observado es un argumento muy convincente de su falsedad (León y Montero, 2003).

Por ejemplo, la verosimilitud de la hipótesis «los burros no son capaces de volar» aumenta cada vez que nos acercamos con ademán amenazante a un burro y éste escapa corriendo, en lugar de hacerlo volando. Pero esto no es suficiente. Quizá los burros sí tengan esa capacidad, pero no la emplean porque no perciben que yo sea lo suficientemente amenazador como para emplear este recurso. Si lo intentásemos con ahínco y con un número grande de burros sin que ninguno saliera volando podríamos considerar que la credibilidad de esa hipótesis es muy grande. Aun así, siempre quedaría la posibilidad de que en un caso de amenaza extrema algún nuevo participante de nuestro experimento saliera volando. La estrategia confirmatoria siempre dejaría dudas.

Por el contrario, nos bastaría con encontrar un solo burro volador para poder afirmar que la hipótesis es falsa. Desde un enfoque falsacionista, un solo ejemplar nos serviría para valorar negativamente la hipótesis propuesta, concluyendo que los burros sí son capaces de volar, al menos en casos de pánico extremo.

Una diferencia fundamental entre la lógica clásica y la estadística inferencial es que, mientras la argumentación en la primera es categorial, en la estadística inferencial la argumentación es probabilística (Rivadulla, 1991). Cuando decimos que las proporciones 0,90/0,10 son incompatibles con la hipótesis de que la moneda es imparcial, no queremos decir que realmente lo sean de un modo absoluto. Ciertamente, una moneda imparcial lanzada 100 veces puede dar lugar a 90 o más caras (de hecho, ocurriría aproximadamente en uno de cada trillón de series de 100 lanzamientos). Lo que se quiere decir es que es un resultado tan improbable que, de su observación, es muy razonable deducir que hay una práctica incompatibilidad entre lo observado y lo esperado bajo esa hipótesis.

Entonces, ¿cuál debería ser nuestra conclusión respecto a una moneda que ofrece, por ejemplo, un resultado de 55/45? Siguiendo con la misma lógica, debe-

ríamos preguntarnos cómo de probable es un resultado como el observado (0,55/0,45) si la moneda fuera imparcial. Si esa probabilidad es muy pequeña, concluiremos que la evidencia indica que la moneda no es imparcial, pero si es grande deberemos concluir que el resultado es compatible con la hipótesis de imparcialidad. En el CH se establece un valor arbitrariamente pequeño para considerar de forma operativa a algo «improbable». Se suelen considerar como evidencias improbables aquellas cuyas probabilidades son menores de 0,05 o 0,01. Cuando se produce un resultado como éste se dice que el resultado es *estadísticamente significativo* (volveremos a este concepto en el apartado 14.5.1).

14.3. ELEMENTOS DE UN CONTRASTE DE HIPÓTESIS

En la aplicación de un CH, como el del ejemplo de la imparcialidad de una moneda, intervienen varios elementos que vamos a identificar y comentar. Vamos a señalar *cinco* elementos, empleando como ejemplo la forma de contraste más sencilla, que se refiere al CH sobre la media poblacional (μ). Ilustraremos cada uno de ellos con el siguiente ejemplo.

Supongamos que en los años 50 se hizo un estudio en Australia que determinó que la edad media a la que los niños australianos adquieren una determinada destreza era de 10 años y su desviación típica era 2,5 (adviértase que se entiende que son valores de parámetros poblacionales). Un colega nos plantea que, debido a una serie de factores, esa edad media no es correcta para los niños actuales. Vamos a hacer un estudio que nos permita alcanzar una conclusión al respecto. La estrategia va a ser recoger información relevante, establecer como hipótesis que la media poblacional es 10 y calcular la probabilidad de haber obtenido la evidencia, que, de hecho, hemos recogido si esa hipótesis fuese verdadera. Si esa probabilidad es pequeña, concluiremos que la hipótesis es falsa (la evidencia es incompatible con la hipótesis y el resultado es estadísticamente significativo). Si no es pequeña, mantendremos la hipótesis como una opción plausible. La información relevante la obtenemos seleccionando una m.a.s. de 81 niños y anotando la edad a la que han adquirido esa destreza. Esas edades tienen una media de 10,8 (media muestral o $\bar{X}$). Veamos los elementos involucrados en este CH.

a) ***Hipótesis.*** El primer paso es siempre el establecimiento de dos hipótesis relativas a los parámetros, conocidas como *hipótesis nula* e *hipótesis alternativa*. La hipótesis nula, que se representa como H_0, es aquella bajo la cual hallamos la probabilidad asociada a la evidencia. Es la que se pretende falsar, mediante la observación de evidencia incompatible (poco probable) con ella. Cuando la probabilidad es muy pequeña, concluiremos que es falsa, mientras que si no lo es la mantendremos. La hipótesis alternativa, que se representa como H_1, es la contraria (o negación) de la nula.

En nuestro ejemplo, la hipótesis nula establece que la media poblacional (μ) es igual a 10. Al valor de la media establecido en la hipótesis nula se le representa por μ_0. La hipótesis alternativa, como negación de la anterior, simplemente establece que esa media es diferente de 10:

$$H_0: \mu = 10$$
$$H_1: \mu \neq 10$$

A menudo las hipótesis científicas se plantean de una forma algo diferente a como lo hemos hecho en este ejemplo. En este planteamiento se rechaza H_0 cuando la media muestral observada se aleje del valor planteado en la hipótesis nula, sin importar que esa media se aleje por ser demasiado grande o demasiado pequeña. Este planteamiento es adecuado cuando no tenemos ninguna expectativa respecto a la dirección de la eventual discrepancia entre el estadístico y el parámetro.

En la investigación científica es habitual que el CH se aplique habiendo una expectativa direccional de la discrepancia. Así, en este ejemplo nuestro colega podría habernos planteado que cree que la media poblacional no es 10, sino que debido a una serie de factores cree que la media actual es mayor (que esa destreza se adquiere más tarde, por término medio). Si queremos realizar un contraste que responda a este debate, no tendría sentido realizar el contraste de la forma como lo hemos hecho. No tendría sentido, porque sería absurdo concluir que la media poblacional actual es *mayor* de 10, empleando como argumento que la media muestral de nuestro estudio es *menor* que ese valor.

Estas dos maneras alternativas de plantear los contrastes reciben los nombres de *contrastes bilaterales* y *contrastes unilaterales*, respectivamente. En los primeros, al rechazar la H_0 nos quedaremos con una alternativa no direccional (por ejemplo, que la media poblacional es distinta de 10), mientras que en los segundos nos quedaremos con una alternativa direccional (por ejemplo, que la media poblacional es mayor de 10).

En los contrastes unilaterales, la H_0 especificará no sólo el valor puntual del parámetro, sino toda la zona que incluye los valores de dirección opuesta a la esperada. En nuestro ejemplo, sería:

$$H_0: \mu \leq 10$$
$$H_1: \mu > 10$$

Si se rechaza H_0, la alternativa con la que nos quedamos es una que especifica la dirección buscada en la discrepancia. Dependiendo de la dirección, se dice que es un contraste *unilateral izquierdo* o *unilateral derecho*.

Por tanto, la formulación de las hipótesis en contrastes unilaterales sería:

Unilateral izquierdo: $H_0: \mu \geq \mu_0$
$H_1: \mu < \mu_0$

Unilateral derecho: $H_0: \mu \leq \mu_0$
$H_1: \mu > \mu_0$

b) ***Supuestos.*** Para poder calcular la probabilidad en la que se basa la decisión hay que asumir que se cumplen algunas condiciones, llamadas *supuestos*. El procedimiento de obtención de los datos (el diseño) será la garantía de algunos supuestos, otros se podrán comprobar, mientras que algunos otros se asumirán con cierto riesgo. Cuanto más arriesgados sean los supuestos que no se pueden comprobar, menos confiable será el resultado de un CH. Para hacer referencia a esta característica de los contrastes se emplea el término *robustez* (véase el apéndice de este capítulo).

Para nuestro ejemplo, vamos a asumir que las N observaciones son independientes, algo que no parece arriesgado, dado que en la información disponible se señala que se ha trabajado con una m.a.s. Habitualmente, nuestros CH se ajustarán al esquema de algunos de los procedimientos más frecuentes. En nuestro ejemplo, podemos emplear lo que sabemos acerca de la distribución muestral de la media cuando se conoce σ. No sabemos si la distribución de la variable es normal, pero sabemos que gracias al teorema central del límite, con muestras moderadamente grandes la distribución muestral de la media se aproxima a la normalidad (véase el apartado 13.3.2). En realidad, en este esquema nuestros supuestos son los mismos que establecíamos como condiciones al estudiar la distribución muestral de la media cuando se conoce σ y la muestra es grande:

— Las observaciones son independientes (m.a.s.).
— Conocemos la varianza poblacional ($\sigma^2 = 2{,}5^2$).
— La muestra es grande ($N = 81$, mayor de 30); permite asumir la normalidad.

c) ***Estadístico de contraste y distribución muestral.*** El estadístico de contraste es una transformación de un resultado muestral que nos resulta un instrumento útil porque su distribución de probabilidad es conocida cuando H_0 es verdadera y se cumplen los supuestos. De esta forma, podemos calcular la probabilidad asociada a la evidencia.

En nuestro ejemplo, el estadístico no es otro que el que define la distribución muestral de la media, conocida σ, para muestras grandes; su distribución se aproxima a la normal unitaria. Su fórmula es la siguiente, en la que el valor paramétrico es sustituido por el que se indica en H_0, dado que vamos a calcular una probabilidad bajo H_0 verdadera:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} \qquad [14.1]$$

Su distribución es aproximadamente $N(0; 1)$. Sustituyendo:

$$Z = \frac{10{,}8 - 10}{2{,}5/\sqrt{81}} = 2{,}88$$

d) ***Regla de decisión.*** Establece el criterio probabilístico que va a conducir a la decisión sobre H_0. Dicha regla debe indicar qué conjunto de posibles valores del estadístico de contraste (Z) son los menos compatibles con H_0 y cuya proba-

bilidad conjunta es igual a ese valor de probabilidad que hemos calificado como de «práctica incompatibilidad». A esa pequeña probabilidad en la que se basa la regla de decisión se le llama *nivel de significación* y se representa por α, mientras que a su complementario ($1 - \alpha$) se le llama *nivel de confianza*. Al conjunto de valores que agregan esa probabilidad se le llama *región crítica*, mientras que el resto de valores conforman la *región de confianza* (véase la figura 14.1).

Siguiendo los procedimientos expuestos en relación con la distribución muestral de la media, podemos formular la regla de decisión del ejemplo de nuestro contraste bilateral de la siguiente forma. Primero establecemos un valor para el nivel de significación (por ejemplo, $\alpha = 0{,}01$). A continuación nos preguntamos qué valores del estadístico de contraste constituyen la región crítica. A los valores que delimitan la región crítica se les denomina *puntos críticos*. Si la H_0 es verdadera, la región de confianza estará constituida por los que abarcan un área de 0,99 y son más compatibles con H_0.

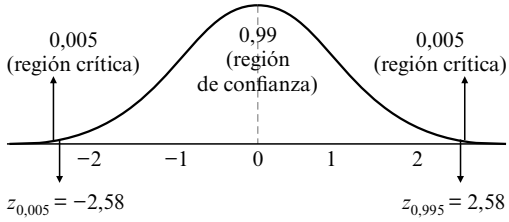


Figura 14.1.—Representación de la región de confianza y la región crítica en un CH (véase explicación en el texto).

En el ejemplo de la figura 14.1, la región de confianza es la que abarca un área central igual a 0,99. En una distribución normal unitaria sería la limitada por los puntos críticos ±2,58 (según la tabla II del apéndice final). De esta forma, se consideran valores incompatibles (poco probables) con H_0 a aquellos que se alejan mucho del esperado, tanto por ser demasiado grandes (superiores a 2,58) como por ser demasiado pequeños (inferiores a −2,58). Es decir, la regla de decisión sería:

«Rechazar H_0 si $Z \geq 2{,}58$ o $Z \leq -2{,}58$; no rechazarla en caso contrario»

Veamos ahora cómo se establece la regla de decisión si el contraste es unilateral. El cambio principal es que, en lugar de establecer una región crítica dividida en dos zonas, la región crítica se concentra en la cola de la distribución que recoge la dirección esperada de la discrepancia. Esta dirección esperada es siempre la que se señala en la hipótesis alternativa (H_1).

Para un mismo nivel de significación, los puntos críticos empleados en contrastes unilaterales y bilaterales serán diferentes. Si en el contraste bilateral empleábamos los valores ±2,58, en uno unilateral de la cola derecha emplearemos el valor que deja a su derecha el área completa de α (en el ejemplo 0,01); según

la tabla II del apéndice final, este valor es 2,33. Por tanto, la regla de decisión en este contraste unilateral sería:

«Rechazar H_0 si $Z \geq 2{,}33$; mantenerla en caso contrario»

La representación gráfica de este contraste es la que aparece en la figura 14.2 (compárese con la de la figura 14.1). La región crítica se concentra en la cola derecha, porque la hipótesis de nuestro colega (aquella con la que queremos quedarnos en caso de rechazar H_0) es que la media actual es mayor de 10. Si su hipótesis hubiera sido que la media actual es menor, entonces hubiéramos concentrado la región crítica en la cola izquierda y el valor de la regla de decisión hubiera sido −2,33.

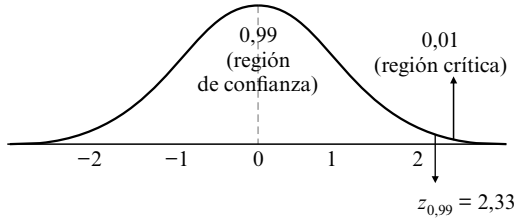


Figura 14.2.—Representación gráfica de un contraste unilateral derecho.

e) ***Decisión y conclusión.*** La decisión se adopta aplicando la regla de decisión establecida en el apartado anterior al valor obtenido como estadístico de contraste, lo que nos lleva en nuestro ejemplo de contraste bilateral a lo siguiente:

«Rechazamos H_0 porque el valor de Z (2,88) es mayor de 2,58 (cae en la región crítica)»

Esta decisión se suele expresar como una conclusión, en los mismos términos en que se expresó la hipótesis de partida: «Hay evidencia suficiente para rechazar la hipótesis de que la media poblacional es igual a 10». En otras palabras, si la media poblacional fuera realmente 10, la probabilidad de obtener en una m.a.s. una media tan alejada del parámetro como la que hemos obtenido (10,8) es menor del nivel de significación establecido ($\alpha = 0{,}01$).

14.4. UNA FORMA ALTERNATIVA DE DECIDIR

Si bien la forma que hemos detallado de desarrollar un CH establece los pasos en orden adecuado, lo cierto es que desde que los CH no se hacen a mano, con una calculadora y unas tablas, sino que se hacen con los ordenadores, por lo que la manera de realizarlos es algo distinta. No se establece de forma explícita la regla de decisión, sino que se obtiene el valor del estadístico de contraste y la probabilidad asociada al mismo. Esta probabilidad, conocida como *nivel crítico*

(o valor p), se compara con el valor establecido para α. Si $p \leq \alpha$ se rechaza H_0, mientras que si $p > \alpha$ se mantiene. El resultado es el mismo, pero resulta más cómodo porque los programas informáticos siempre ofrecen el valor del estadístico y su valor de probabilidad asociado. En lugar de comparar el valor empírico del estadístico de contraste con los puntos críticos establecidos en la regla de decisión, se compara el nivel crítico (p) con el nivel de significación (α). El valor de α se establece antes de obtener los datos, mientras que p se calcula después de obtener los datos. Podemos ofrecer la siguiente definición:

> El *nivel crítico* (p) de un contraste es el menor valor de α con el que se hubiera rechazado H_0.

En nuestro ejemplo unilateral procederíamos de la siguiente forma. Tras obtener el valor 2,88 para el estadístico de contraste, comprobamos que el área que ese valor deja a su derecha (p) es 0,002; es decir, $P(z \geq 2{,}88) = 0{,}002$. La decisión sigue siendo la de rechazar H_0, ya que $p < \alpha$ (0,002 < 0,01).

Naturalmente, el nivel crítico será igual al área asociada si el contraste es unilateral, pero habrá que multiplicarlo por 2 si se trata de un contraste bilateral. En el contraste bilateral de nuestro ejemplo, tendríamos que comparar $2 \cdot 0{,}002 = 0{,}004$ con $\alpha = 0{,}01$. La decisión es la misma que si seguimos los pasos indicados en el apartado 14.3.

Como decíamos, esta forma de realizar el contraste es la más habitual hoy en día, ya que los ordenadores nos proporcionan directamente el valor p. El investigador evalúa este valor comparándolo con el valor de α que ha asumido. Para hacerlo de la otra forma con un ordenador, habría que proporcionar a la máquina cuál es el valor asumido como nivel de significación (α) y acabaríamos indicando exclusivamente si es menor que α o no. Sin embargo, para la publicación de informes de investigación se suele exigir algo más que esa decisión categorial; se exige que se informe del valor de p, ya que al hacerlo se hace público para otros investigadores y les resulta útil cuando aplican otros procedimientos más avanzados (American Psychological Association, 2010).

14.5. OTRAS CUESTIONES RELACIONADAS CON EL CH

14.5.1. Sobre la expresión «estadísticamente significativo»

Las conclusiones de los CH son de dos grandes tipos (Abelson, 1995): con el criterio de probabilidad representado por α, los datos son indistinguibles de los que generaría el azar o son distinguibles de los que generaría el azar. Veamos lo que significan estas dos conclusiones:

a) *Los datos son indistinguibles de los que generaría el azar*. No estamos diciendo, como muchas veces se concluye, que estén producidos realmente

por un proceso aleatorio. Sólo se dice que la diferencia entre los datos (el estadístico) y el valor propuesto para el parámetro en H_0, al estar dentro de los valores asumibles con una probabilidad igual a $1-\alpha$, es indistinguible de los que podría producir el muestreo aleatorio.

b) *Los datos son distinguibles de los que generaría el azar.* Esta conclusión implica una negación de la indistinguibilidad con respecto al azar. Lo que se dice es que la diferencia entre el estadístico y el valor hipotetizado para el parámetro es mayor de lo que esperaríamos con una probabilidad igual a $1-\alpha$, y se distinguen de lo que podría ser el fruto del muestreo aleatorio.

Para describir esta segunda situación se dice que el resultado (la discrepancia entre el estadístico y el valor propuesto para el parámetro) es *estadísticamente significativo*. No se debe entender esta expresión como que el resultado es relevante, importante, ni ningún otro sinónimo. Sólo se quiere decir que si H_0 fuera verdadera, la probabilidad de obtener un resultado como el observado sería menor de α.

14.5.2. ¿Es lo mismo no rechazar H_0 que aceptarla?

A lo largo de este capítulo hemos empleado las expresiones «rechazar» y «mantener» (o «no rechazar») para referirnos a las decisiones respecto a H_0. Quizá el lector se esté preguntando por qué en el segundo caso no se dice simplemente «aceptar» H_0; de hecho, así es como se suelen expresar los estudiantes que dan sus primeros pasos en la investigación y en la aplicación de técnicas de CH. Aunque pueda parecer una diferencia superficial, puramente terminológica, en los manuales de estadística aplicada se suele insistir en la importancia de no emplear el término «aceptar». La razón está de nuevo en la diferente solidez argumentativa de los enfoques basados en la confirmación y la falsación. En realidad, cuando se mantiene H_0 no se está aportando evidencia directa de que ésta sea verdadera; sólo se concluye que la evidencia obtenida es «compatible» con ella (probabilísticamente hablando), pero seguramente también lo es con muchos otros valores que pudiéramos hipotetizar para el parámetro. Por el contrario, cuando se obtiene evidencia «incompatible» con H_0 (probabilísticamente hablando) ésta se considera una evidencia mucho más sólida de su falsedad. Las expresiones que hemos propuesto pretenden resaltar (y servir como recordatorio) del diferente estatus argumentativo de las dos decisiones. En un caso se alcanza la «sólida» conclusión de que H_0 es falsa, mientras que en el otro la conclusión es menos sólida y se prefiere decir, más prudentemente, que H_0 se mantiene como «plausible» o que «no se rechaza», en lugar de concluir que se acepta como verdadera.

14.5.3. Tipos de error en un CH

Como en todo escenario en el que se adoptan decisiones, en un CH se pueden cometer errores. Ya que en los CH las alternativas de la decisión son dos, rechazar o no rechazar, hay dos formas de equivocarse: rechazar H_0 siendo verdadera

y no rechazarla siendo falsa. El esquema pretende establecer un virtuoso equilibrio entre las probabilidades de los dos tipos de error, de manera que se incremente la probabilidad de las dos formas de acierto (rechazar H_0 cuando es falsa y no rechazarla cuando es verdadera). Lo que es imposible es que las probabilidades de cometer los dos tipos de error sean nulas.

Al error que consiste en rechazar una H_0 que es verdadera se le llama *error tipo I*. Su probabilidad (condicionada a que H_0 fuera verdadera) es α, mientras que la probabilidad de no rechazarla es $1-\alpha$. Por el contrario, si H_0 es falsa tenemos también dos probabilidades condicionales, la de rechazar y la de no rechazar, que se suelen representar por $1-\beta$ y β, respectivamente. A la probabilidad de rechazar H_0 siendo falsa ($1-\beta$) se le llama *potencia* del contraste. Al error que consiste en no rechazar una H_0 falsa se le llama *error tipo II* y su probabilidad es β.

Los términos y probabilidades se resumen en la siguiente tabla:

		Decisión sobre H_0	
		No Rechazar	*Rechazar*
H_0	*Verdadera*	$1-\alpha$	α Error Tipo I
	Falsa	β Error Tipo II	$1-\beta$

La probabilidad de cometer un error tipo I la establece el investigador *a priori* al fijar α, por lo que el lector podría preguntarse por qué no emplear un valor menor (por ejemplo 0,00001 o incluso más pequeño) en lugar de los valores 0,05 y 0,01 habituales. La razón es que, aunque al hacerlo se reduce la probabilidad del error tipo I, a cambio se incrementa la probabilidad del error tipo II. El margen de maniobra que tiene el investigador, que se refiere sobre todo al establecimiento del valor de α y del tamaño de la muestra empleada, se emplea para formular reglas de decisión con un equilibrio adecuado entre las probabilidades de los dos tipos de error.

Veámoslo con el ejemplo del contraste unilateral derecho del apartado 14.3, en el que vamos a suponer que H_0 es falsa porque la media poblacional no es en realidad $\mu = 10$, sino $\mu = 11$. En la figura 14.3 se representa la situación, donde la curva izquierda representa la distribución muestral de la media bajo H_0 verdadera ($\mu = 10$) y la curva derecha representa la distribución muestral de la media si $\mu = 11$ (H_0 falsa). En el eje aparecen los valores del estadístico de contraste bajo H_0 verdadera (que ya están tipificados), mientras que en el eje paralelo inferior aparece su equivalencia en términos de medias muestrales. La distribución de la izquierda está centrada sobre el valor 10, que es el valor del parámetro recogido en H_0. El punto crítico (2,33) equivale a un valor de media muestral aproximado de 10,65. Este valor se puede obtener sustituyendo 2,33 como valor de Z en la fórmula del estadístico de contraste y despejando el valor de la media correspondiente. La regla de decisión «rechazar H_0 si $Z \geq 2{,}33$» es idéntica a la regla de decisión «rechazar H_0 si $\overline{X} \geq 10{,}65$».

Veamos ahora cuáles son las probabilidades de rechazar y no rechazar si la media poblacional fuera realmente el valor 11. La curva de la derecha representa esta posibilidad, y las probabilidades que buscamos son las áreas que el valor del punto de corte deja a su izquierda y derecha en esta curva. Tipificando ahora respecto a su media poblacional (11) podemos calcular que la media asociada al punto de corte (10,65) deja a su izquierda un área de 0,1038 (β) y a su derecha 0,8962 ($1-\beta$) (véase el ejercicio 5 de este capítulo).

En la figura 14.3 se aprecia el balance entre las dos probabilidades de error. Al tratar de reducir la probabilidad de cometer un error tipo I (α) desplazando hacia la derecha el punto de corte, se incrementa la probabilidad de cometer un error tipo II (β), ya que ésta es el área izquierda de la otra curva. Por ello se suele establecer un valor de α convencional y tratar de reducir β con otros procedimientos (como por ejemplo incrementando el tamaño de la muestra; no desarrollaremos aquí esta cuestión, que excede el ámbito de este libro, pero se puede acceder a una discusión más detallada en Pardo y San Martín, 2010).

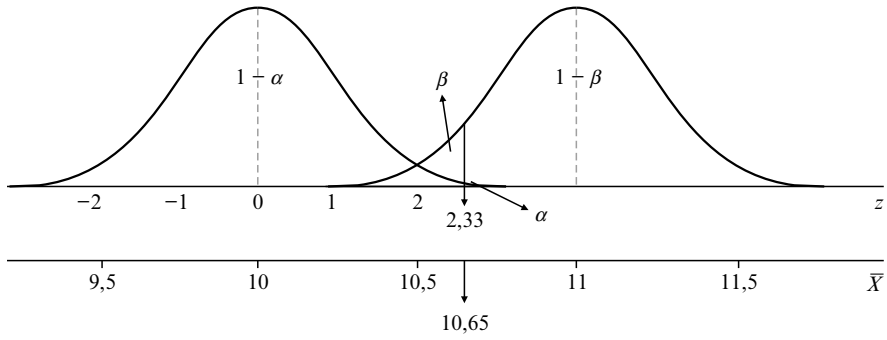


Figura 14.3.—Representación gráfica de la relación entre las probabilidades de los dos tipos de error. Al desplazar el punto de corte hacia la derecha se reduce la probabilidad de cometer un error tipo I (α), pero se incrementa la de cometer un error tipo II (β).

Es importante advertir que la potencia de un contraste se refiere siempre a un valor concreto de H_1 (en nuestro ejemplo, $\mu = 11$). Cuando no se cuenta con un valor alternativo, se puede obtener la llamada función de potencia (función que asocia un valor de la potencia a los posibles valores de μ_1). En la mayoría de los casos en los que se realiza un CH en psicología no se realizará respecto a un valor concreto alternativo, sino respecto a la negación de la hipótesis nula, tal y como hemos planteado en nuestro esquema.

14.5.4. Parámetros poblacionales y propensiones

Al exponer la lógica de los CH estamos haciendo referencia constante al hecho de que la evidencia (o información muestral) nos permite hacer inferencias acerca de los parámetros. Puede parecer que sólo se aplican en situaciones donde

existen parámetros poblacionales en sentido estricto (e.g., socio-demográficos), pero lo cierto es que en psicología son más frecuentes las hipótesis que se refieren a tendencias de comportamiento. La diferencia entre uno y otro reside en un factor extra-estadístico, que es el procedimiento operativo de generación de la variación. Este procedimiento puede referirse a un *muestreo* o a un *proceso.* En el primer caso nos referimos a la extracción de unidades a partir de poblaciones; en el segundo, a la producción de un determinado comportamiento. Esto no es relevante para el procedimiento inferencial del CH, que es un proceso mecánico y ciego respecto a la naturaleza de lo que se estudia, pero sí lo es para la interpretación de los datos. De ello se deduce que la interpretación de los datos tiene que hacer referencia a elementos adicionales al mero tratamiento de los números.

Veamos un ejemplo de la distinción entre población y proceso. Supongamos que queremos estudiar el desarrollo intelectual de los niños sordo-ciegos de nacimiento en España, con objeto de planificar su proceso educativo. Tras una búsqueda exhaustiva, concluimos que en España sólo hay 123 casos y, gracias a la colaboración de múltiples entidades, conseguimos que todos ellos participen en la investigación. ¿Tiene sentido hacer CH sobre los valores obtenidos? Si consideramos que estamos trabajando con la población completa, no tendría sentido hacer inferencias sobre sus parámetros, dado que éstos son directamente comprobables. Por el contrario, sí tiene sentido hacerlo si analizamos la situación en términos de *propensión* (Frick, 1998). La propensión sería la tendencia de un proceso a producir un cierto resultado. En estas condiciones, el proceso de desarrollo de un niño en determinadas condiciones no tiene tamaño (poblacional). Se estudian los productos de ese proceso en términos de tendencias a producir ciertos tipos de resultados observables. La población estaría constituida por todos los casos similares que existen o que potencialmente pudieran existir. La población completa de niños sordo-ciegos españoles es, en realidad, una muestra de los resultados que produciría ese supuesto proceso de desarrollo cuya propensión a generar determinados resultados vamos a estudiar.

En cambio, en los sondeos electorales se pretende estimar el estado de opinión y la intención de voto de una población real mediante la observación de una muestra aleatoria. En este caso sí que se hacen inferencias sobre características (parámetros) poblacionales en un sentido literal. Aquí la población es el conjunto finito y concreto de ciudadanos que el día de las elecciones tienen derecho a votar.

Lo importante es que el procedimiento de CH se aplica a un conjunto limitado de observaciones de una variable aleatoria. Que la variable aleatoria represente la selección de unidades de una población o el registro de un conjunto limitado de producciones de un proceso es indiferente para el CH. En ambos casos se contrastan hipótesis respecto a los parámetros de la variable aleatoria, aunque en el primer caso éstos representen una característica de una población *(parámetro)* y en el segundo representen una tendencia de un proceso *(propensión)*.

PROBLEMAS Y EJERCICIOS

1. A continuación se presentan una serie de situaciones que implican el planteamiento de diferentes hipótesis sobre el parámetro μ. Indique, en cada uno de los casos, cómo se plantea cada una de las hipótesis y el tipo de contraste a que hacen referencia (unilateral o bilateral).

a) Asumiendo que la variable *CI* se distribuye $N(100; 15)$ en la población general, se desea investigar si el valor de la media es diferente en la población de sordomudos.

b) Sabiendo que los varones españoles nacidos entre 1990 y 1993 se distribuyen en la variable estatura $N(175; 4{,}5)$, se desea someter a contraste la hipótesis de que las mujeres españolas nacidas en la misma época tienen una estatura media menor a la de los varones.

c) En una población de estudiantes de secundaria se ha observado que las puntuaciones obtenidas en una prueba de matemáticas en el curso 2007-2008 siguen el modelo $N(5; 2)$. Se desea analizar si el promedio en dicha prueba se ha incrementado en el curso 2010-2011.

2. Asumiendo que la variable X sigue el modelo $N(10; 3)$, se ha sometido a contraste la siguiente hipótesis: H_0: $\mu \geq 10$; H_1: $\mu < 10$. Seleccionada una m.a.s. de 16 sujetos, se obtiene que $\overline{X} = 8{,}2$. Según estos datos, ¿cuál tiene que ser el nivel de significación mínimo para rechazar H_0?

3. Se ha llevado a cabo un contraste de hipótesis sobre μ para la variable X que sigue el modelo $N(20; 2{,}5)$. Asumiendo que se cumplen los supuestos, se establece la siguiente regla de decisión: rechazar H_0 si $Z \leq z_{0,05}$. Sabiendo también que el estadístico de contraste toma el valor $Z = -2{,}54$, conteste a las siguientes cuestiones:

a) ¿Cuáles son las hipótesis del contraste?

b) ¿Qué nivel de significación se ha empleado?

c) ¿Cuál es el nivel crítico?

d) ¿Qué decisión se tomará?

4. Las puntuaciones en un test de apertura emocional se distribuyen normalmente en la población general con varianza 8. Se toma una m.a.s. de 15 personas con baja autoestima para estudiar si estas personas tienen un nivel de apertura emocional similar, y se realiza un contraste sobre μ, siendo las hipótesis: H_0: $\mu = 50$; H_1: $\mu \neq 50$. Según el enunciado, calcule el nivel crítico e indique la decisión estadística a tomar en cada uno de los siguientes casos:

a) Si $\overline{X} = 52$.

b) Si $\overline{X} = 49$.

5. Se ha llevado a cabo un contraste de hipótesis unilateral derecho sobre H_0: $\mu = 6$ con $\alpha = 0{,}05$ y sabiendo que $\sigma = 5$ y $N = 25$. Si la media poblacional fuese en realidad $\mu = 9$, indique la probabilidad asociada al tipo de error que se podría producir en este contraste.

6. Tras haber realizado diferentes CH, se han establecido sus respectivas conclusiones. Indique, razonadamente, si cada una de ellas es correcta.

a) Si H_0: $\mu = 20$; H_1: $\mu \neq 20$ y con $\alpha = 0{,}05$ se ha obtenido que el estadístico de contraste es igual a $Z = -2{,}14$ y $P(z \leq -2{,}14) = 0{,}0404$. Conclusión: como $p < \alpha$, se rechaza H_0.

b) Si H_0: $\mu \geq 34$; H_1: $\mu < 34$ y con $\alpha = 0{,}05$ se ha obtenido que $p = 0{,}12$. Conclusión: como $p > \alpha$, entonces H_0 es verdadera.

c) Si H_0: $\rho = 0$; H_1: $\rho \neq 0$ y con $\alpha = 0{,}01$ se ha obtenido que $p = 0{,}0005$. Conclusión: como $p < \alpha$, entonces se rechaza H_0; por tanto, entre las variables X e Y existe una relación lineal directa.

d) Si H_0: $\mu = 18$; H_1: $\mu \neq 18$, donde σ es conocida, y con $\alpha = 0{,}01$ se ha obtenido que el valor del estadístico de contraste es igual a 3. Conclusión: se rechaza H_0.

e) Si H_0: $\mu = 37$; H_1: $\mu \neq 37$, donde σ es conocida, y con $\alpha = 0{,}05$ se ha obtenido que el valor del estadístico de contraste es igual a 1,86. Como $1{,}86 > (z_{0{,}95} = 1{,}64)$, entonces se rechaza H_0.

f) Si H_0: $\mu \geq 230$; H_1: $\mu < 230$, donde σ es conocida, y con $\alpha = 0{,}05$ se ha obtenido que el valor del estadístico de contraste es igual a $-1{,}83$. Como $-1{,}83 < (z_{0{,}95} = 1{,}64)$, entonces se mantiene H_0.

g) Si H_0: $\mu \leq 6$; H_1: $\mu > 6$, donde σ es conocida, y con $\alpha = 0{,}01$ se ha obtenido que el valor del estadístico de contraste es igual a 2,80. Como $2{,}80 > (z_{0{,}01} = -2{,}33)$, entonces se mantiene H_0.

7. Se asume que la variable X sigue una distribución normal en la que σ es igual a 30. Se quiere poner a prueba la hipótesis H_0: $\mu \leq 120$; H_1: $\mu > 120$ con $\alpha = 0{,}05$. Responda a las siguientes cuestiones:

a) Tomando una m.a.s. de tamaño igual a 25 de las puntuaciones de X, se obtiene que $\overline{X} = 126$. ¿A qué conclusión se llegaría?

b) Si se toma una m.a.s de tamaño igual a 100 y se obtiene que la media de X en la muestra es igual a la del apartado anterior, ¿a qué conclusión se llegaría?

c) Compare los resultados de los dos apartados anteriores y coméntelos brevemente.

8. Se asume que la variable X sigue una distribución normal en la que σ es igual a 40. Se quiere poner a prueba la hipótesis H_0: $\mu \leq 80$; H_1: $\mu > 80$. Para ello se toma una m.a.s. de tamaño 100, obteniéndose que $\overline{X} = 88$. Responda a las siguientes cuestiones:

a) Si se establece un nivel de significación igual a 0,05, ¿a qué conclusión se llegaría?

b) Si se establece un nivel de significación igual a 0,01, ¿a que conclusión se llegaría?

c) Compare los resultados de los dos apartados anteriores y coméntelos brevemente.

9. Se asume que la variable X sigue una distribución normal en la que σ es igual a 52. Se toma una m.a.s. de tamaño 169, obteniéndose que $\bar{X} = 83$. Responda a las siguientes cuestiones:

a) Se pone a prueba la hipótesis H_0: $\mu \geq 90$; H_1: $\mu < 90$, con $\alpha = 0{,}05$. ¿A qué conclusión se llega?

b) Se pone a prueba la hipótesis H_0: $\mu = 90$; H_1: $\mu \neq 90$, con $\alpha = 0{,}05$. ¿A qué conclusión se llega?

c) Compare los resultados de los dos apartados anteriores y coméntelos brevemente.

10. Se asume que la variable X sigue una distribución normal en la que σ es 22. Se toma una m.a.s. de tamaño 121, obteniéndose que $\bar{X} = 71$. Responda a las siguientes cuestiones:

a) Si se pone a prueba H_0: $\mu \leq 70$; H_1: $\mu > 70$, con $\alpha = 0{,}01$, ¿a qué conclusión se llegaría?

b) Si se pone a prueba H_0: $\mu \leq 67$; H_1: $\mu > 67$, con $\alpha = 0{,}01$, ¿a qué conclusión se llegaría?

c) Compare los resultados de los dos apartados anteriores y coméntelos brevemente.

11. Retomando los datos del ejercicio 8, se sabe que H_0 es falsa y que el verdadero valor de μ es 90. Obtenga la probabilidad asociada al error tipo II, β, y la potencia, $1 - \beta$, para cada uno de los valores de α; compare los resultados.

SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. *a*) H_0: $\mu = 100$; H_1: $\mu \neq 100$ (bilateral).
b) H_0: $\mu \geq 175$; H_1: $\mu < 175$ (unilateral izquierdo).
c) H_0: $\mu \leq 5$; H_1: $\mu > 5$ (unilateral derecho).

2. Con cualquier $\alpha > 0{,}0082$ (por ejemplo, los valores habituales 0,01 y 0,05) se rechazaría H_0, pero con cualquier $\alpha \leq 0{,}0082$, se mantendría.

3. *a*) $H_0: \mu \geq 20$; $H_1: \mu < 20$.
b) $\alpha = 0{,}05$.
c) $p = 0{,}0055$.
d) Se rechaza H_0, pues $(Z = -2{,}54) < (z_{0{,}05} = -1{,}64)$; o también porque $p < \alpha$ (es decir, $0{,}0055 < 0{,}05$).

4. *a*) El nivel crítico es: $p = 0{,}0062$; como $p < \alpha$, la decisión es rechazar la H_0.
b) El nivel crítico es: $p = 0{,}1706$; como $p > \alpha$, la decisión es no rechazar H_0.

Como se puede ver, la decisión sobre H_0, pese a referirse a la misma h pótesis, puede ser diferente según sea el valor del estadístico de contraste.

5. El tipo de error es el error tipo II (mantener la H_0 sabiendo que es falsa) y la probabilidad asociada a dicho error es $\beta = 0{,}0869$.

6. *a*) La conclusión es incorrecta. Obsérvese que se está haciendo un contraste bilateral, por lo que el nivel crítico no es el área asociada al valor, sino esa probabilidad multiplicada por 2. En este caso $p = 2 \cdot \cdot 0{,}0404 = 0{,}0808$; como $0{,}0808 \geq 0{,}05$, entonces se mantiene H_0.
b) La decisión es correcta, pero la conclusión es incorrecta. No se puede afirmar que H_0 es verdadera; lo correcto es afirmar que se mantiene H_0 o no se rechaza H_0.
c) La conclusión es incorrecta. Se ha realizado un CH bilateral que sólo permite concluir que la correlación es no nula, que existe algún tipo de relación lineal. Si se quisiera poner a prueba también el signo de la correlación habría que realizar un CH unilateral.
d) La conclusión es correcta. Se trata de un CH bilateral sobre una media conocida σ. Con $\alpha = 0{,}01$, entonces $z_{0{,}005} = -2{,}58$ y $z_{0{,}995} = 2{,}58$; de aquí que el valor del estadístico de contraste, 3, es mayor que $z_{0{,}995} = 2{,}58$ y se rechaza H_0.
e) La conclusión es incorrecta. Se ha planteado un CH bilateral; por tanto, hay que comparar el valor del estadístico de contraste con $z_{\alpha/2}$ y $z_{1-\alpha/2}$; en este caso, $z_{0{,}025} = -1{,}96$ y $z_{0{,}975} = 1{,}96$. Como el valor del estadístico de contraste, 1,86, está entre $z_{0{,}025}$ y $z_{0{,}975}$, se mantiene H_0.
f) La conclusión es incorrecta. Se está planteando un CH unilateral izquierdo. Por tanto, hay que comparar el valor del estadístico de contraste con el valor correspondiente a z_α; en este caso, $z_{0{,}05} = -1{,}64$. Como $-1{,}83 < (z_{0{,}05} = -1{,}64)$, entonces se rechaza H_0.
g) La conclusión es incorrecta. Se está planteando un CH unilateral derecho. Por tanto, hay que comparar el valor del estadístico de contraste con el valor correspondiente a $z_{1-\alpha}$; en este caso, $z_{0{,}99} = 2{,}33$. Como $2{,}80 > (z_{0{,}99} = 2{,}33)$, entonces se rechaza H_0.

7. *a*) Se mantiene H_0.

b) Se rechaza H_0.

c) Comparando los dos apartados anteriores, la única diferencia existente es el tamaño de las muestras. Se observa que el tamaño muestral, N, repercute en el CH, de tal forma que, a medida que se incrementa el tamaño de la muestra, es más probable rechazar H_0. Como idea general, esto no sólo ocurre en el contraste de una media conocida σ, sino en contrastes sobre otros parámetros: el tamaño de la muestra influye en el CH.

8. *a*) Se rechaza H_0.

b) El estadístico de contraste toma el mismo valor, pero en este caso se mantiene H_0.

c) Para un mismo conjunto de datos y un mismo CH, el nivel de significación (α) puede modificar las conclusiones. La decisión sobre qué valor de α tomar depende del *nivel de riesgo* que se quiera asumir. Ese nivel de riesgo dependerá de las consecuencias prácticas (efecto de un tratamiento, aplicar o no un tratamiento) que se obtengan del CH.

9. *a*) Se rechaza H_0.

b) Se mantiene H_0.

c) Para un mismo conjunto de datos y un mismo nivel de significación (α), el tipo de CH (unilateral o bilateral) puede modificar la decisión y las conclusiones.

10. *a*) Se mantiene H_0.

b) Se mantiene H_0.

c) En ambos casos, todo se mantiene igual excepto el valor del parámetro establecido en las hipótesis. En ambos casos se mantiene la H_0. Obsérvese que por este motivo no es correcto concluir que la H_0 es verdadera, ya que se puede llegar a la misma conclusión aunque el valor del parámetro de la H_0 sea diferente. La media muestral es (probabilísticamente) compatible con ambos valores del parámetro.

11. *a*) $\beta = 0{,}1949$ y la potencia es $1 - \beta = 0{,}8051$.

b) $\beta = 0{,}4325$ y $1 - \beta = 0{,}5675$. Comparando los resultados de los dos apartados, se observa que tanto la probabilidad de un error tipo II como la de la potencia dependen de α. Si se reduce la probabilidad de un error tipo I (α) aumenta la del error tipo II (β), disminuyendo la potencia ($1 - \beta$) (si se mantiene constante el tamaño de la muestra).

APÉNDICE

Dos enfoques en el CH: Neyman-Pearson *versus* Fisher

La manera como se realizan hoy en día los CH es un híbrido entre dos enfoques que se desarrollaron en paralelo en la década de 1920 (Gigerenzer, Swijtnik, Porter, Daston, Beatty y Kruger, 1989). El primer enfoque es el de *Neyman-Pearson*, en el que se establecen dos hipótesis puntuales (valores concretos del parámetro) y se establece una regla de decisión que tiene en cuenta las probabilidades de las dos formas de error (α y β). El segundo es el de *Fisher*, en el que se establece una única hipótesis (H_0) y se calcula la probabilidad de obtener un resultado tan extremo como el observado. En este enfoque no se plantea ninguna H_1 específica.

El procedimiento que hemos descrito es el que aparece en la mayoría de los manuales de estadística aplicada y, aunque es una mezcla de ambos, se parece más al enfoque de Fisher, pues la regla de decisión se establece teniendo sólo en consideración la probabilidad de cometer un error tipo I (α), mientras que se intenta incrementar ($1 - \beta$) con otros procedimientos (utilizando muestras grandes), pero no se calcula porque no se establece un valor específico como alternativo a H_0. Sin embargo, se parece al de Neyman-Pearson en que tiene dos hipótesis, aunque la alternativa no sea un valor puntual, sino la negación de la nula.

Hay una diferencia de mayor calado, que tiene un carácter más epistemológico. El enfoque de Neyman-Pearson es en el fondo un enfoque confirmatorio limitado a dos valores puntuales. Dado que se asume que sólo hay dos valores posibles del parámetro, decir que la evidencia es contraria a uno de ellos es lo mismo que decir que se trata de evidencia a favor de la otra. Por el contrario, el enfoque de Fisher se encuadra más directamente en la perspectiva falsacionista. El rechazo de H_0 no apoya nada concreto, sino que se limita a negar el valor implicado en ella.

Sobre la robustez de los contrastes

Un contraste es tanto más robusto cuanto menos sensible es a la violación de los supuestos en los que se basa. Por ejemplo, el contraste sobre la media es robusto con respecto a la normalidad si las muestras son moderadamente grandes. Esto quiere decir que asumir la creencia (falsa) de que la distribución poblacional es normal no tiene consecuencias graves si la muestra es grande ($N \geq 30$), ya que el teorema central del límite establece que la distribución muestral se aproxima a la normal incluso cuando la variable original no lo es. Estamos relativamente protegidos ante situaciones en las que asumimos equivocadamente la normalidad.

Una parte de la estadística aplicada se ha dedicado al desarrollo de procedimientos inferenciales que exijan pocos supuestos, o que dichos supuestos sean poco arriesgados o poco sensibles a la violación de sus supuestos. Estos procedimientos son con frecuencia preferidos por su robustez. En asignaturas posteriores se expondrán algunos de ellos, como por ejemplo los que constituyen la estadística no paramétrica (Pardo y San Martín, 2010).

Contraste de hipótesis sobre algunos parámetros 15

15.1. INTRODUCCIÓN

Una vez expuesta la lógica del CH, y tras haber definido los términos y conceptos involucrados, hay que decir que esa lógica general se concreta en una gran cantidad de técnicas particulares. Cada técnica ha sido desarrollada para ser empleada en un escenario específico, es decir, para las hipótesis referidas a un determinado parámetro, con unos determinados supuestos distribucionales y en unas circunstancias concretas. En asignaturas posteriores se expondrá una amplia variedad de estas técnicas, elegidas por ser algunas de las más empleadas en Psicología. En este capítulo, último del libro, vamos a exponer algunas de las más sencillas, para que el estudiante se vaya familiarizando con ellas y para que las pueda ir aplicando en los contextos en los que sean pertinentes.

Los cinco esquemas concretos de CH que vamos a explicar ilustran la aplicación del esquema general descrito en el capítulo anterior. Los dos primeros se refieren al CH sobre la media poblacional (μ), distinguiendo entre los casos en los que se conoce la varianza (σ^2) y los casos en los que se desconoce. El tercero y el cuarto se refieren, respectivamente, a los contrastes sobre la correlación y la proporción, cuyas distribuciones muestrales hemos discutido ya en el capítulo 13. Por último, expondremos el contraste de independencia entre variables categóricas, a partir de tablas de contingencia. Recordemos los pasos que componen el esquema general:

a) Hipótesis.
b) Supuestos.
c) Estadístico de contraste y distribución muestral.
d) Regla de decisión.
e) Decisión y conclusión.

También indicaremos cómo aplicarlo de la forma alternativa que hemos señalado en el apartado 14.5, basada en la comparación entre el nivel crítico y el nivel de significación (entre p y α).

15.2. CONTRASTE DE HIPÓTESIS SOBRE LA MEDIA (μ)

Para contrastar hipótesis sobre el valor de una media vamos a distinguir dos casos: aquellos en los que se conoce la varianza poblacional y aquellos en los que no se conoce. Aunque el primer caso es muy infrecuente en la práctica, por razones didácticas se suele exponer en primer lugar (ya lo hemos avanzado en el capítulo anterior). Como el lector avezado estará ya suponiendo, para cada caso emplearemos la distribución muestral correspondiente.

15.2.1. Conocida σ

Los pasos a seguir son los siguientes:

a) *Hipótesis.* Si se trata de un contraste bilateral, éstas serán de la forma:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

Si se trata de contrastes unilaterales, H_0 se establece como $\mu \leq \mu_0$ si es unilateral derecho o como $\mu \geq \mu_0$ si es izquierdo; naturalmente, H_1 será $\mu > \mu_0$ en el primer caso y $\mu < \mu_0$ en el segundo.

b) *Supuestos.* Este esquema se aplica cuando se conoce la varianza poblacional (σ^2) y la distribución muestral se ajusta al modelo normal. Esto último se asumirá de forma automática cuando la variable de partida sea normal, pero también se asumirá por aproximación cuando la muestra sea moderadamente grande ($N \geq 30$, por el teorema central del límite; capítulo 13). Por supuesto, también se asumirá que las observaciones constituyen una m.a.s., algo que dependerá del diseño de muestreo empleado. En resumen, al aplicar este esquema se indicarán los siguientes supuestos:

— La población se distribuye $N(\mu; \sigma)$ o, recurriendo al teorema central del límite, la muestra es suficientemente grande como para asumir la normalidad de la distribución muestral.
— La media muestral se ha obtenido sobre una m.a.s.
— Conocemos σ.

c) *Estadístico de contraste y distribución muestral.* El estadístico y su distribución bajo H_0 verdadera son:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}; \quad Z \sim N(0; 1) \qquad [15.1]$$

Esta distribución es exactamente normal si la población de partida es normal, mientras que sólo es aproximadamente normal si invocamos el teorema central del límite.

d) *Regla de decisión*. Basada en el nivel de significación adoptado (α), puede ser bilateral o unilateral. Dependiendo de ello, se formulará de la siguiente forma:

Bilateral	Rechazar si $Z \geq z_{1-\alpha/2}$ o $Z \leq z_{\alpha/2}$ No rechazar si $z_{\alpha/2} < Z < z_{1-\alpha/2}$
Unilateral derecho	Rechazar si $Z \geq z_{1-\alpha}$ No rechazar si $Z < z_{1-\alpha}$
Unilateral izquierdo	Rechazar si $Z \leq z_{\alpha}$ No rechazar si $Z > z_{\alpha}$

e) *Decisión y conclusión*. Aplicando la regla al valor obtenido y reformulando en términos de la hipótesis de investigación.

Recordemos que la forma alternativa de decidir sobre H_0 se basa en la comparación entre p y α. La manera de aplicarla depende de que el contraste sea bilateral o unilateral. Para hacerlo, se ignora el paso en el que se establece la regla de decisión y se calcula directamente el estadístico de contraste. Si se trata de un contraste bilateral se obtiene el área o probabilidad que deja en su cola más corta. El valor de p es igual a esa probabilidad multiplicada por 2. Se rechaza H_0 si $p < \alpha$. Si se trata de un contraste unilateral, el valor de p es directamente el área que deja el estadístico de contraste hacia el lado que señala H_1. De nuevo, se rechaza H_0 si $p < \alpha$.

Ejemplo de un contraste bilateral. Supongamos que queremos contrastar la hipótesis de que la media poblacional en una determinada variable, X, es igual a 100, sabiendo que la varianza poblacional es igual a 64 y que X es normal. Para ello extraemos una m.a.s. de 25 observaciones y calculamos su media aritmética en X, que resulta ser igual a 104; establecemos un nivel de significación (α) de 0,05.

Aplicamos el esquema de la siguiente forma:

a) *Hipótesis.* En el problema no se especifica nada sobre la dirección de la diferencia entre 100 y la media poblacional real, en caso de que H_0 sea falsa, por lo que se plantea como un contraste bilateral:

$$H_0: \mu = 100$$
$$H_1: \mu \neq 100$$

b) *Supuestos*:

— La población se distribuye normal.
— Se trata de una m.a.s.
— Conocemos la varianza poblacional ($\sigma^2 = 64$).

c) *Estadístico de contraste y distribución muestral:*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}; \quad Z \sim N(0; 1)$$

$$Z = \frac{104 - 100}{8/\sqrt{25}} = 2{,}50$$

d) *Regla de decisión*:

Rechazar si $Z \geq 1{,}96$ o $Z \leq -1{,}96$

No rechazar si $-1{,}96 < Z < 1{,}96$

e) *Decisión y conclusión.* Como $2{,}50 > 1{,}96$ *rechazamos* H_0. Concluimos que la evidencia aconseja rechazar, según la regla de decisión adoptada, la hipótesis de que la media poblacional es igual a 100.

Para aplicar la forma alternativa se obtiene el valor 2,50 y se comprueba en la tabla de la distribución normal que este valor deja a su derecha un área de 0,0062. Como se trata de un contraste bilateral, $p = 2 \cdot 0{,}0062 = 0{,}0124$. Dado que $p < \alpha$ $(0{,}0124 < 0{,}05)$ se rechaza H_0.

Un último comentario sobre esta técnica de contraste. Entre los supuestos se ha incluido la normalidad de la distribución en la población de la variable implicada. Por lo que ya hemos estudiado, sabemos que, aunque esto no fuera verdad, la distribución muestral de la media se aproxima a la normal a medida que se incrementa el tamaño de la muestra empleada (teorema central del límite). Es raro que sepamos con certeza la forma de la distribución de una variable y en cambio no conozcamos su media. Es más frecuente que desconozcamos ambos. Esta es la razón por la que se suele recomendar, para aplicar este contraste, que la muestra sobre la que se calcula la media sea de al menos $N = 30$. De esta forma, se podrá aplicar esta técnica sin preocuparse por la distribución de la variable de partida, ya que estaremos protegidos ante el eventual error de creer que la variable original es normal, cuando en realidad no es así.

15.2.2. Desconocida σ

Con mucha frecuencia nos encontraremos en una situación como la anterior, pero con la diferencia de que no conoceremos la varianza poblacional, σ^2. Es decir, queremos contrastar si la media poblacional es igual a un cierto valor y podemos asumir que la distribución muestral es normal (porque la población es normal o porque se trata de una muestra grande) y que la media se ha obtenido en una m.a.s. Si la única diferencia con el escenario anterior es que no conocemos la varianza poblacional (algo bastante razonable, dado que será raro que no conozcamos μ y en cambio conozcamos σ), entonces podemos recurrir a un

estadístico similar al anterior, pero en el que en lugar de aparecer σ en el denominador aparece su estimador S (la desviación típica de la muestra). La única consecuencia de sustituir σ por S es que el estadístico de contraste ya no se distribuye $N(0; 1)$, sino según la distribución t de Student, siendo los grados de libertad el tamaño de la muestra menos uno (t_{N-1}) (véase en Pardo, Ruiz y San Martín, 2009):

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{N}}; \quad T \sim t_{N-1}$$

Si lo que aparece en el denominador es el estimador sesgado de la desviación típica, S, entonces en la raíz aparece $(N - 1)$. En resumen, representando por S_N y S_{N-1} a los estimadores sesgado e insesgado, respectivamente, las fórmulas serían:

$$T = \frac{\bar{X} - \mu_0}{S_N/\sqrt{N-1}} \qquad \text{y} \qquad T = \frac{\bar{X} - \mu_0}{S_{N-1}/\sqrt{N}}$$

En ambos casos la distribución es la misma: t_{N-1}. El esquema es muy similar al del caso anterior:

a) *Hipótesis.* Si se trata de un contraste bilateral, éstas serán de la forma:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

De nuevo, si se trata de contrastes unilaterales H_0 se puede establecer como $\mu \leq \mu_0$ (derecho) o como $\mu \geq \mu_0$ (izquierdo); por supuesto, H_1 será $\mu > \mu_0$ en el primer caso y $\mu < \mu_0$ en el segundo.

b) *Supuestos.* Lo único que cambia es el tercer supuesto, relativo a σ:

— La población se distribuye $N(\mu; \sigma)$ o la muestra es suficientemente grande como para asumir la normalidad de la distribución muestral, basándonos en el teorema central del límite.
— La media muestral se ha obtenido sobre una m.a.s.
— Desconocemos σ.

c) *Estadístico de contraste y distribución muestral*. En la fórmula sólo cambia que σ se sustituye por su estimador y que su distribución ya no es la normal, sino la t de Student:

$$T = \frac{\bar{X} - \mu_0}{S_N/\sqrt{N-1}} \qquad \text{o} \qquad T = \frac{\bar{X} - \mu_0}{S_{N-1}/\sqrt{N}} \quad T \sim t_{N-1} \qquad [15.2]$$

d) *Regla de decisión*. Sólo cambia respecto al caso anterior en que los valores de la regla no proceden de la distribución normal, sino de la de *t*:

Bilateral	Rechazar si $T \geq {}_{1-\alpha/2}t_{N-1}$ o $T \leq {}_{\alpha/2}t_{N-1}$ No rechazar si ${}_{\alpha/2}t_{N-1} < T < {}_{1-\alpha/2}t_{N-1}$
Unilateral derecho	Rechazar si $T \geq {}_{1-\alpha}t_{N-1}$ No rechazar si $T < {}_{1-\alpha}t_{N-1}$
Unilateral izquierdo	Rechazar si $T \leq {}_{\alpha}t_{N-1}$ No rechazar si $T > {}_{\alpha}t_{N-1}$

e) *Decisión y conclusión*. Igual que el caso anterior.

La forma alternativa de decidir sobre H_0 también se aplica como en el caso anterior.

Ejemplo de un contraste unilateral izquierdo. Supongamos que sospechamos que la media poblacional en una determinada variable, *X*, que se distribuye según el modelo normal, es menor de 70. Para realizar un contraste que nos ayude a decidir al respecto, extraemos una m.a.s. de 81 observaciones y en ella obtenemos que su media es 65,8 y su varianza insesgada (S^2_{N-1}) es igual a 236; establecemos un nivel de significación (α) de 0,05.

Aplicamos el esquema de la siguiente forma:

a) *Hipótesis.* En el problema se especifica la dirección de la diferencia entre 70 y la media poblacional real. En caso de que H_0 sea falsa, se quiere decidir en el sentido de que la media poblacional es *menor* que ese valor, por lo que se plantea como un contraste unilateral izquierdo:

$$H_0: \mu \geq 70$$
$$H_1: \mu < 70$$

b) *Supuestos*:

— La población se distribuye normal.
— Se trata de una m.a.s.
— Desconocemos la varianza poblacional.

c) *Estadístico de contraste y distribución muestral.* Como disponemos del estimador insesgado de σ, el estadístico de contraste es:

$$T = \frac{\bar{X} - \mu_0}{S_{N-1}/\sqrt{N}} \quad T \sim t_{80}$$

La desviación típica muestral es $\sqrt{236} = 15{,}36$ y el estadístico de contraste:

$$T = \frac{65{,}8 - 70}{15{,}36/\sqrt{81}} = -2{,}461$$

d) *Regla de decisión*:

Rechazar si $T \leq -1{,}664$

No rechazar si $T > -1{,}664$

e) *Decisión y conclusión.* Como $-2{,}461 < -1{,}664$, *rechazamos* H_0. Concluimos que la evidencia aconseja rechazar, según la regla de decisión adoptada, la hipótesis de que la media poblacional sea igual o mayor que 70 y favoreciendo la hipótesis de que es menor de 70.

Para aplicar la forma alternativa, obtenemos el área que el valor del estadístico de contraste deja a su izquierda (porque es un contraste unilateral izquierdo). Según la tabla de *t*, esa área es aproximadamente $p = 0{,}01$. Por tanto, como $p < \alpha$ ($0{,}01 < 0{,}05$) rechazamos H_0.

15.3. CONTRASTE DE HIPÓTESIS SOBRE LA CORRELACIÓN (ρ)

El caso que exponemos aquí es única y exclusivamente el contraste de la hipótesis de independencia lineal: aquel en el que se contrasta si la correlación de Pearson paramétrica es 0. Los contrastes sobre cualquier otro valor exigen otros elementos que se expondrán en la asignatura de Análisis de Datos en Psicología II (Pardo y San Martín, 2010). No obstante, el contraste del valor 0 es, con mucho, el más interesante y el que con mayor frecuencia se aplica.

Contrastar la hipótesis de independencia lineal entre dos variables significa contrastar la hipótesis de que la correlación poblacional (ρ) es igual a 0. Para ello necesitamos especificar un escenario en el que podamos definir un estadístico de contraste con una distribución conocida con la que establecer la regla de decisión. El escenario buscado, que se basa en la distribución muestral de la correlación (apartado 13.4), es el que se resume en el siguiente esquema, en el que se llega a un estadístico de contraste que, bajo la hipótesis nula verdadera, se distribuye según el modelo *t* de Student con $N - 2$ grados de libertad (t_{N-2}).

a) *Hipótesis.* Si se trata de un contraste bilateral, éstas serán de la forma:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Si se trata de contrastes unilaterales, H_0 se puede establecer como $\rho \leq 0$ (derecho) o como $\rho \geq 0$ (izquierdo); H_1 será $\rho > 0$ (relación directa entre

las variables) en el primer caso y $\rho < 0$ (relación inversa entre las variables) en el segundo.

b) *Supuestos.* Tal y como vimos en el capítulo 13 en relación con la distribución muestral de la correlación, su distribución se ajustará a un modelo conocido (*t* de student) si ambas variables son normales. Los supuestos del contraste son los siguientes:

— Las dos variables, X e Y, se distribuyen según la normal.
— La correlación muestral, r_{xy}, se ha obtenido sobre una m.a.s. de pares de valores de X e Y.

c) *Estadístico de contraste y distribución muestral.* Bajo H_0 verdadera el estadístico de contraste y su distribución son:

$$T = \frac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r_{xy}^2}}; \quad T \sim t_{N-2} \qquad [15.3]$$

d) *Regla de decisión*. Basada en el nivel de significación adoptado (α), puede ser bilateral o unilateral. Dependiendo de ello, se formulará de la siguiente forma:

Bilateral	Rechazar si $T \geq {}_{1-\alpha/2}t_{N-2}$ o $T \leq {}_{\alpha/2}t_{N-2}$ No rechazar si ${}_{\alpha/2}t_{N-2} < T < {}_{1-\alpha/2}t_{N-2}$
Unilateral derecho	Rechazar si $T \geq {}_{1-\alpha}t_{N-2}$ No rechazar si $T < {}_{1-\alpha}t_{N-2}$
Unilateral izquierdo	Rechazar si $T \leq {}_{\alpha}t_{N-2}$ No rechazar si $T > {}_{\alpha}t_{N-2}$

e) *Decisión y conclusión*. Aplicando la regla al valor obtenido y reformulando en términos de la hipótesis de investigación.

La forma alternativa se aplica como en los casos anteriores.

Ejemplo de un contraste bilateral. Supongamos que queremos contrastar si a nivel poblacional las variables X e Y son linealmente independientes. Extraemos una m.a.s. de 62 observaciones, y en ella obtenemos una correlación de $r_{xy} = 0{,}23$. Por estudios anteriores sabemos que podemos asumir que se trata de variables normales; establecemos un nivel de significación (α) de 0,05.

Aplicamos el esquema de la siguiente forma:

a) *Hipótesis.* En el problema no se especifica nada sobre el sentido de la correlación, en caso de ser distinta de 0, por lo que se plantea como un contraste bilateral:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

b) *Supuestos:*

— Ambas variables se distribuyen normalmente en la población.
— Se trata de una m.a.s.

c) *Estadístico de contraste y distribución muestral*. El estadístico y su distribución bajo H_0 verdadera son:

$$T = \frac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1-r_{xy}^2}}; \quad T \sim t_{60}$$

$$T = \frac{0{,}23 \cdot \sqrt{62-2}}{\sqrt{1-0{,}23^2}} = 1{,}831$$

d) *Regla de decisión:*

Rechazar si $T \geq 2{,}00$ o $T \leq -2{,}00$

No rechazar si $-2{,}00 < T < 2{,}00$

e) *Decisión y conclusión*. Como 1,831 está entre −2,00 y 2,00, *mantenemos* H_0. Concluimos que la evidencia sugiere, según la regla de decisión adoptada, que la correlación en la muestra es (probabilísticamente) compatible con la hipótesis de que la correlación poblacional es 0 (que las variables son en realidad linealmente independientes).

Para aplicar la forma alternativa se obtiene el valor 1,831 y se comprueba en la tabla de *t* que este valor deja a su derecha aproximadamente un área de 0,04. Al tratarse de un contraste bilateral, $p = 2 \cdot 0{,}04 = 0{,}08$. Como $p > \alpha$ (0,08 > 0,05) no se rechaza H_0.

15.4. CONTRASTE DE HIPÓTESIS SOBRE LA PROPORCIÓN (π)

La técnica que describimos aquí se aplica cuando se quieren contrastar hipótesis relativas a proporciones poblacionales, pero también para contrastar hipótesis relativas a probabilidades de ocurrencia de un evento («propensiones»; véase el apartado 14.5.4). Ejemplos de lo primero serían su aplicación cuando queremos contrastar hipótesis respecto a la proporción (o porcentaje) de individuos de la población que cumplen cierta condición (tener una opinión favorable a una cierta cuestión, tener estudios superiores, estar vacunado, estar desempleado, etc.). Ejemplos de lo segundo serían su aplicación cuando queremos contrastar hipótesis relativas a si la probabilidad de que un estudiante acierte una pregunta es mayor que la del mero azar.

Por otro lado, el esquema que vamos a describir recoge, naturalmente, lo que hemos expuesto en el capítulo 13 respecto a la distribución muestral de la proporción, pero restringiéndonos al caso de muestras grandes del apartado 13.5.1 (para un caso de muestras pequeñas, véase el apéndice del capítulo presente). Allí veíamos que la tipificación de la proporción muestral se distribuye aproximadamente $N(0; 1)$. Aprovecharemos este hecho en la aplicación del siguiente esquema:

a) *Hipótesis.* Si se trata de un contraste bilateral respecto al valor π_0, éstas serán:

$$H_0: \pi = \pi_0$$
$$H_1: \pi \neq \pi_0$$

Si se trata de contrastes unilaterales, H_0 se puede establecer como $\pi \leq \pi_0$ (derecho) o como $\pi \geq \pi_0$ (izquierdo); de nuevo, H_1 será $\pi > \pi_0$ en el primer caso y $\pi < \pi_0$ en el segundo.

b) *Supuestos.* Este esquema se aplica cuando se cumplen las condiciones para que la proporción muestral se distribuya según el modelo binomial y el tamaño muestral permita emplear la aproximación al modelo normal. Es decir, al aplicar este esquema se indicarán los siguientes supuestos:

— La proporción se obtiene a partir de una variable dicotómica.
— La probabilidad de la condición no cambia (π constante).
— Las condiciones permiten asumir la aproximación del estadístico de contraste al modelo normal (siempre que $N \cdot \pi_0 \geq 5$).

c) *Estadístico de contraste y distribución muestral.* El estadístico y su distribución bajo H_0 verdadera son (la corrección por continuidad se suele ignorar, dado que con muestras grandes tiene un efecto muy pequeño en el cálculo):

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{N}}}; \quad Z \sim N(0; 1) \qquad [15.4]$$

d) *Regla de decisión.* Basada en el nivel de significación adoptado (α), puede ser bilateral o unilateral. Dependiendo de ello, se formulará de la siguiente forma:

Bilateral	Rechazar si $Z \geq z_{1-\alpha/2}$ o $Z \leq z_{\alpha/2}$ No rechazar si $z_{\alpha/2} < Z < z_{1-\alpha/2}$
Unilateral derecho	Rechazar si $Z \geq z_{1-\alpha}$ No rechazar si $Z < z_{1-\alpha}$
Unilateral izquierdo	Rechazar si $Z \leq z_{\alpha}$ No rechazar si $Z > z_{\alpha}$

e) *Decisión y conclusión*. Aplicando la regla al valor obtenido y reformulando en términos de la hipótesis de investigación.

La forma alternativa se aplica como en los casos anteriores.

Ejemplo de un contraste unilateral izquierdo. Supongamos que queremos *refutar* la hipótesis de que la mayoría de la población está a favor de una determinada reforma constitucional (nosotros creemos que los que están a favor son minoría). Para ello extraemos una m.a.s. de 250 ciudadanos y hacemos una consulta telefónica, en la que 115 dicen estar a favor de esa reforma; por tanto, la proporción de personas favorables en la muestra es $P = 115/250 = 0{,}46$. Establecemos un nivel de significación (α) de 0,05.

Aplicamos el esquema de la siguiente forma:

a) *Hipótesis.* En el problema sí se especifica la dirección de la diferencia que debería conducir al rechazo de la hipótesis nula. Serán mayoría los favorables a la reforma si el porcentaje de éstos es del 50 por 100 o más, mientras que concluiremos que no son mayoría si ese porcentaje es menor del 50 por 100. Expresamos esto como hipótesis en términos de proporciones:

$$H_0: \pi \geq 0{,}50$$
$$H_1: \pi < 0{,}50$$

b) *Supuestos:*

— La proporción se obtiene a partir de una variable dicotómica.
— La encuesta se ha realizado sobre una m.a.s. (π constante).
— Las condiciones permiten asumir la aproximación del estadístico de contraste al modelo normal ($N \cdot \pi_0 = 250 \cdot 0{,}50 = 125 > 5$).

c) *Estadístico de contraste y distribución muestral:*

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0 \cdot (1 - \pi_0)}{N}}}; \quad Z \sim N(0; 1)$$

$$Z = \frac{0{,}46 - 0{,}50}{\sqrt{\frac{0{,}50 \cdot 0{,}50}{250}}} = -1{,}265$$

d) *Regla de decisión*:

Rechazar si $Z \leq -1{,}64$

No rechazar si $Z > -1{,}64$

e) *Decisión y conclusión*. Como −1,265 > −1,64 *no rechazamos* H_0. Concluimos que la evidencia obtenida no es suficiente para rechazar la hipótesis de que los partidarios de la reforma sean mayoría.

Para aplicar la forma alternativa obtenemos el área que el valor del estadístico de contraste deja a su izquierda (porque es un contraste unilateral izquierdo). Según la tabla de la normal unitaria para el valor −1,265 esa área es aproximadamente $p = 0{,}102$. Por tanto, como $p > \alpha$ (0,102 > 0,05) no rechazamos H_0.

15.5. CONTRASTE DE LA HIPÓTESIS DE INDEPENDENCIA ENTRE VARIABLES CATEGÓRICAS

Ya estamos en condiciones de afrontar una cuestión que dejamos pendiente en el capítulo 8 (véase el apartado 8.1.3) y cuyo abordaje exige el empleo de lo que hemos aprendido sobre CH. Recordemos que en aquel apartado nos preguntábamos cómo interpretar o valorar unos resultados relacionados con la independencia de dos variables categóricas (o cualitativas). Para ello habíamos definido un estadístico que fue propuesto por Pearson y que se basa en la relación entre las frecuencias observadas en las casillas y las esperadas en caso de que las variables fueran independientes. Vamos a reproducir aquí aquel estadístico (que allí aparecía en la fórmula [8.2]) con otra numeración, propia de este capítulo, y representando por X^2 al estadístico, ya que será el estadístico de contraste (aunque en muchos textos se denota directamente como χ^2):

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \qquad [15.5]$$

En este estadístico, las n_{ij} son las frecuencias empíricas u observadas de las casillas y las m_{ij} son las estimaciones de las frecuencias esperadas bajo la hipótesis de independencia (véase en la fórmula [8.1]); I es el número de filas y J el número de columnas.

Recordemos también cuál es el contexto de aplicación. Supongamos que anotamos de una muestra de 200 encuestados dos valores, correspondientes a la variable sexo y a su acuerdo/desacuerdo con una propuesta de reforma constitucional (tabla 15.1).

TABLA 15.1

Ejemplo de tabla de contingencia entre el sexo y el acuerdo sobre una reforma constitucional

		Sexo		
		Varón (1)	*Mujer* (0)	
Reforma constitucional	*De acuerdo* (1)	48	78	126
	En desacuerdo (0)	32	42	74
		80	120	200

La proporción total de favorables a la reforma es 0,63 (126/200), pero cuando la calculamos sobre las distribuciones condicionales obtenemos que entre los hombres la proporción de favorables es 0,60 (48/80), mientras que entre las mujeres es 0,65 (78/120). Podemos preguntarnos si la opinión es independiente del sexo, empleando un procedimiento riguroso como el del CH, que ya hemos aplicado para otras situaciones.

Según vimos en el apartado 9.5.2 y en el cuadro 10.4, las variables son independientes si la probabilidad de obtener una opinión favorable entre los hombres es la misma que entre las mujeres [P(De acuerdo | Hombre) = P(De acuerdo | Mujer). Dicho de otra forma, si P(De acuerdo $\cap$ Hombre) = P(De acuerdo) $\cdot$ P(Hombre); esta condición se cumple para todas las casillas.

Sin embargo, aunque a nivel poblacional las variables sean independientes, es difícil que las frecuencias observadas en una muestra particular se ajusten aritméticamente a lo que se esperaría de dos variables independientes. Así, la diferencia entre las proporciones de favorables entre los hombres y las mujeres (0,60 y 0,65) podría ser el resultado de una mera fluctuación aleatoria (igual que una correlación muestral distinta de 0 puede ser compatible con una correlación poblacional nula, o igual que una proporción de caras diferente de 0,50 en una serie de 200 lanzamientos puede ser compatible con una moneda imparcial). Por tanto, el procedimiento que vamos a exponer permite responder a la pregunta de si una cierta discrepancia entre las frecuencias teóricas de una tabla de contingencia y las esperadas bajo la hipótesis de independencia son (probabilísticamente) compatibles.

Como se aprecia en la fórmula [15.5], el valor de X^2 crece a medida que las frecuencias observadas se separan de las esperadas, pero como las discrepancias de las casillas se elevan al cuadrado, el estadístico crece tanto si las discrepancias son por exceso como por defecto: este estadístico mide el grado de discrepancia en términos absolutos. Una consecuencia de esta característica es que la región crítica se concentra en una sola de las colas de la distribución de X^2, la derecha, de forma que los contrastes son siempre unilaterales derechos y los rechazos siempre conducen a rechazar la H_0 de independencia, sin especificar el sentido de la asociación entre las variables.

El esquema del procedimiento de contraste de hipótesis es el siguiente:

a) *Hipótesis:*

H_0: Las variables son independientes ($\pi_{ij} = \pi_{i+} \cdot \pi_{+j}$, para todo i y j).

H_1: Las variables no son independientes ($\pi_{ij} \neq \pi_{i+} \cdot \pi_{+j}$, para algún i y j).

b) *Supuestos*:

— Se hacen N observaciones, independientes entre sí (las probabilidades de clasificación en cada categoría de X e Y permanecen constantes para cada observación).

— Las frecuencias esperadas de todas las casillas son mayores de 5 ($N \cdot \pi_{ij} \geq 5$, para todo i y j) (la demostración de este supuesto excede el alcance de este libro; véase en Pardo, Ruiz y San Martín, 2009).

c) *Estadístico de contraste y distribución muestral*. El estadístico y su distribución bajo H_0 verdadera son:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}; \quad X^2 \sim \chi^2_{(I-1)(J-1)}$$

d) *Regla de decisión*. Basada en el nivel de significación adoptado (α), es siempre unilateral derecho. Se formulará siempre de la siguiente forma:

Rechazar si $X^2 \geq {}_{1-\alpha}\chi^2_{(I-1)(J-1)}$

No rechazar si $X^2 < {}_{1-\alpha}\chi^2_{(I-1)(J-1)}$

e) *Decisión y conclusión*. Aplicando la regla al valor obtenido y reformulando en términos de la hipótesis de investigación.

La forma alternativa se aplica como en los casos anteriores, pero teniendo en cuenta que el contraste es siempre unilateral derecho y que *p* será siempre igual al área que el estadístico deje a su derecha.

Ejemplo. Veamos como ejemplo lo que ocurre con la tabla de contingencia entre el sexo y la respuesta a la pregunta sobre la reforma constitucional (tabla 15.1). Hemos accedido a una m.a.s. de 200 ciudadanos y les hemos planteado las preguntas de nuestro cuestionario. En la sección de datos sociodemográficos aparece la pregunta sobre el sexo, y en la de opiniones la de si están o no de acuerdo con la reforma constitucional. La tabla de contingencia es la que hemos mostrado algo más arriba. Queremos estudiar si la posición respecto a la reforma constitucional (acuerdo/desacuerdo) es independiente del sexo o, más bien, si la posición es diferente entre hombres y mujeres; emplearemos el nivel de significación $\alpha = 0{,}05$.

Aplicamos el esquema de la siguiente forma:

a) *Hipótesis:*

H_0: La posición sobre la cuestión es independiente del sexo.
H_1: La posición sobre la cuestión no es independiente del sexo.

b) *Supuestos:*

— Se hacen 200 observaciones, independientes entre sí.
— Las frecuencias esperadas de todas las casillas son mayores de 5.

Comprobamos que se cumple el segundo supuesto obteniendo para cada casilla el resultado de $n_{i+} \cdot n_{+j}/N$, que no es más que la fórmula [8.1], donde n_{i+} y n_{+j} son las frecuencias marginales de la casilla; en la siguiente tabla reproducimos la tabla de contingencia, añadiendo entre paréntesis las frecuencias esperadas.

		Sexo		
		Varón (1)	*Mujer* (0)	
Reforma constitucional	*De acuerdo* (1)	48 (50,4)	78 (75,6)	126
	En desacuerdo (0)	32 (29,6)	42 (44,4)	74
		80	120	200

c) *Estadístico de contraste y distribución muestral.* Como la tabla tiene sólo dos filas y dos columnas, el estadístico tiene un único grado de libertad $[(I-1)(J-1)]$. Por tanto, el estadístico y su distribución bajo H_0 verdadera son:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}; \quad X^2 \sim \chi^2_{(I-1)(J-1)}$$

$$X^2 = \frac{(48-50{,}4)^2}{50{,}4} + \frac{(32-29{,}6)^2}{29{,}6} + \frac{(78-75{,}6)^2}{75{,}6} + \frac{(42-44{,}4)^2}{44{,}4} = 0{,}5148$$

d) *Regla de decisión*:

Rechazar si $X^2 \geq 3{,}841$

No rechazar si $X^2 < 3{,}841$

e) *Decisión y conclusión*. Como $0{,}5148 < 3{,}841$ *no rechazamos* H_0. Concluimos que la evidencia obtenida no es suficiente para rechazar que la opinión sobre la reforma constitucional es independiente del sexo.

Para aplicar la forma alternativa obtenemos el área que el valor del estadístico de contraste deja a su derecha. Según la tabla de su distribución, para el valor 0,5148 ese área es aproximadamente $p = 0{,}500$. Por tanto, como $p > \alpha$ $(0{,}500 > > 0{,}05)$ no rechazamos H_0.

PROBLEMAS Y EJERCICIOS

1. El *cociente intelectual, CI*, se distribuye $N(100; 15)$ en la población general. Un investigador toma una m.a.s. de 9 niños autistas y desea comprobar si la media es distinta en esta población. Encuentra que la media es 115. ¿Cuál será su conclusión con $\alpha = 0{,}05$?

2. Un equipo de psicólogos sociales está estudiando si los niños que ven habitualmente películas violentas presentan un mayor grado de agresividad. Se sabe que la agresividad de los niños de esa edad, a nivel poblacional, se distribuye $N(15; 5)$. Si el equipo selecciona una m.a.s. de 25 niños que habitualmente ven películas violentas y encuentra que la media en agresividad es 17, ¿a qué conclusión llegará con $\alpha = 0{,}01$?

3. Extraemos una m.a.s. de N sujetos de una población y les administramos un test y se obtiene una media muestral de 60,8. Sabiendo que las puntuaciones en el test siguen el modelo $N(51; 20)$, ¿cuál debería ser el tamaño muestral mínimo necesario para rechazar la H_0: $\mu = 51$, con $\alpha = 0{,}05$?

4. Queremos contrastar la hipótesis de que la media poblacional en una determinada variable, que sigue el modelo normal, es igual a 90 (con $\alpha = 0{,}05$). Para ello extraemos una m.a.s. de 61 observaciones y obtenemos que su media es 92,28 y su varianza $S_N^2 = 189$. ¿A qué conclusión se llegará?

5. En una empresa, el salario medio anual para las mujeres, que sigue el modelo normal, es de 28€ (expresado en miles). En una m.a.s. de 10 varones de la misma empresa se obtienen los siguientes salarios: 24, 27, 31, 21, 19, 26, 30, 22, 15, 36. ¿Existe evidencia para concluir que el salario medio es diferente en hombres y mujeres? ($\alpha = 0{,}01$).

6. En una población de estudiantes de bachillerato se sabe que las puntuaciones obtenidas en una prueba de matemáticas en el curso 2007-2008 siguen el modelo $N(5{,}8; \sigma)$. Se desea analizar si el promedio en dicha prueba se ha incrementado en el curso 2010-2011. Para ello, extraemos una m.a.s. de 51 estudiantes y se obtiene que su media es 6,23 y su varianza $S_N^2 = 10{,}9$. ¿A qué conclusión se llegará con $\alpha = 0{,}05$?

7. Medidas las variables X: Salario (€/semana) e Y: Absentismo (horas/año) en una m.a.s. de 5 sujetos, se encuentran los siguientes datos:

X:	300	400	350	320	420
Y:	200	406	272	250	452

Sabiendo que la correlación de Pearson es $r_{xy} = 0{,}989$ y asumiendo que X e Y son normales, someta a contraste la hipótesis de que existe una relación lineal de tipo directa entre salario y absentismo con $\alpha = 0{,}01$.

8. Un psicólogo escolar desea comprobar si la fatiga (X), medida como número de horas de estudio el día antes del examen, tiene una *relación lineal* con la calificación obtenida (Y). Para ello, toma una m.a.s. de cinco estudiantes y obtiene los siguientes datos. Asumiendo que X e Y son normales, ¿cuál será su conclusión con $\alpha = 0{,}05$?

X:	3,0	1,2	2,5	1,7	3,5
Y:	4,1	6,3	5,6	6,1	4,2

9. Retomemos los datos del ejercicio 6 del capítulo 5, donde se estudiaba la relación entre la motivación de logro con diferentes facetas de la satisfacción laboral en una muestra de 84 trabajadores y se obtenía la siguiente matriz de correlaciones:

$$\mathbf{R} = \begin{array}{c|ccccc} & SS & SH & SR & MC & OP \\ \hline SS & & 0{,}82 & 0{,}61 & 0{,}42 & -0{,}49 \\ SH & & & 0{,}35 & 0{,}15 & -0{,}15 \\ SR & & & & 0{,}75 & 0{,}31 \\ MC & & & & & 0{,}45 \\ OP & & & & & \end{array}$$

Asumiendo que *SS* (Satisfacción con el sueldo obtenido) y *OP* (Oportunidades de promoción) son normales, ¿se puede afirmar que la correlación entre ambas es estadísticamente significativa? ($\alpha = 0{,}05$).

10. Según estudios de audiencias realizados en años anteriores, el 80 por 100 de las personas prefieren seguir la retransmisión de un partido de fútbol por la televisión, mientras que el resto prefieren seguirlo por la radio. En el presente año se cree que dicho porcentaje ha disminuido. Para estudiar la posible reducción se ha tomado una m.a.s. de 200 personas que siguen retransmisiones de partidos de fútbol, obteniéndose que 150 prefieren ver un partido por la televisión. Si se establece un nivel de significación de 0,05, ¿a qué conclusión se llegará?

11. Se está realizando una investigación sobre la actitud de los españoles ante la reducción de penas a personas que han cometido delitos violentos. Hace cinco años se observó que el 60 por 100 de la población se mostraba contraria a la reducción. Se cree que actualmente dicho porcentaje ha aumentado. Para probar esta hipótesis se ha tomado una m.a.s. de 150 personas y se les ha preguntado si están a favor o en contra de la reducción de la pena. Se ha obtenido que 96 están en contra. ¿A qué conclusión se llegará? ($\alpha = 0{,}05$).

12. Un psicólogo social está investigando el interés que muestra la población de estudiantes universitarios acerca del comportamiento de los políticos. En estudios anteriores se concluyó que el 50 por 100 de dicha población estaba muy interesada. El psicólogo cree que este porcentaje ha cambiado. Para evaluar su hipótesis toma una m.a.s. de 80 universitarios españoles, de los cuales 30 afirman que están muy interesados. ¿A qué conclusión llegará el psicólogo social? ($\alpha = 0{,}01$).

13. Se está diseñando un experimento de psicofísica visual. En concreto, se trata de estudiar el mínimo nivel de contraste que ha de tener un estímulo, presentado en una pantalla, para ser percibido por un participante. Para ello se diseña un experimento en el que en cada ensayo se divide la pantalla en dos zonas, izquierda y derecha, presentándose, de forma aleatoria, el estímulo con un nivel de contraste fijo en una de las dos zonas. La tarea del participante es informar en cuál de las dos zonas se ha presentado el estímulo. Se considera que el estímulo ha sido percibido si, del total de ensayos, el participante da un porcentaje de respuestas correctas superior al 75 por 100. Tras pasar a un participante 150 ensayos, éste da 120 respuestas correctas. ¿Se puede concluir que dicho participante no ha percibido el estímulo? ($\alpha = 0{,}05$).

14. Siguiendo con la investigación planteada en el ejercicio 11, también se estudió si existía relación entre las variables sexo y la aceptación-rechazo a la reducción de la pena. Más abajo se presenta la tabla de contingencia obtenida. Según los datos, ¿a qué conclusión se llegará? ($\alpha = 0{,}01$).

		Actitud ante la reducción de la pena		
		A favor	En contra	
Sexo	Mujer	40	50	90
	Varón	10	50	60
		50	100	150

15. Un psicólogo clínico quiere estudiar si existe relación entre el nivel formativo (sin estudios, básico, superior) y la depresión (subclínica, depresión mayor). Para ello toma una m.a.s. de 200 personas y clasifica a cada una de ellas en su categoría respectiva, obteniendo la tabla de contingencia que se presenta más abajo. Según estos resultados, ¿a qué conclusión llegará el psicólogo? ($\alpha = 0{,}05$).

		Nivel de estudios			
		Sin estudios	Básico	Superior	
Depresión	Subclínica	20	35	15	70
	Depresión mayor	30	55	45	130
		50	90	60	200

16. Para el diseño de un nuevo juego de mesa se han tenido en cuenta dos características: el material en el que está hecho (plástico, tela y madera) y el número de colores utilizados (dos, tres y cinco). Para cada combinación de los niveles de dichas características se ha construido un prototipo, que es presentado a una m.a.s. de 300 adolescentes. La tarea de los participantes es informar de cuál prefieren. Los datos obtenidos quedan resumidos en la tabla de contingencia que se presenta más abajo. ¿Hay relación entre el tipo de material y el número de colores? ($\alpha = 0{,}05$).

		Colores			
		Dos	Tres	Cinco	
Material	Plástico	30	40	30	100
	Tela	23	39	8	70
	Madera	27	51	52	130
		80	130	90	300

17. Se está estudiando la posible relación entre el medio en el que se vive (rural-urbano) y padecer un tipo de trastorno de la alimentación (anorexia-bulimia). Se toma una m.a.s. de 200 personas que han sufrido uno de los trastornos y que viven en un medio rural o en un medio urbano. Los datos obtenidos se resumen en la tabla de contingencia que se presenta más adelante. ¿Se puede concluir que no existe relación entre el medio en el que se vive y el tipo de trastorno de la alimentación? ($\alpha = 0{,}01$).

		Trastorno		
		Anorexia	Bulimia	
Medio	Rural	43	37	80
	Urbano	87	33	120
		130	70	200

18. Volviendo a la matriz de correlaciones del ejercicio 9, calcule cuánto tiene que valer cualquier r_{xy} en esa matriz para que esa correlación sea considerada estadísticamente significativa en un contraste bilateral con $\alpha = 0{,}05$.

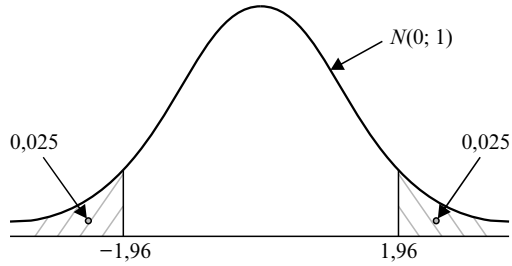
SOLUCIONES DE PROBLEMAS Y EJERCICIOS

1. Contraste de hipótesis de una media conocida σ:

1) *Hipótesis*: H_0: $\mu = 100$; H_1: $\mu \neq 100$.
2) *Supuestos*:
   - — Normalidad.
   - — m.a.s.
   - — Se conoce σ.
3) *Estadístico de contraste y distribución muestral:*

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} = \frac{115 - 100}{15/\sqrt{9}} = 3; \quad \text{donde } Z \sim N(0; 1).$$

4) *Regla de decisión*. Como $\alpha = 0{,}05$: $z_{0{,}025} = -1{,}96$ y $z_{0{,}975} = 1{,}96$, por tanto, los puntos críticos $Z \leq -1{,}96$ y $Z \geq 1{,}96$ delimitan la zona crítica. Es decir:



El nivel crítico es: $p = 2 \cdot P(z \geq 3{,}00) = 2 \cdot 0{,}0013 = 0{,}0026$.

5) *Decisión y conclusión*: como $3 > 1{,}96$ (es decir, el estadístico cae en la zona crítica) y $(p = 0{,}0026) < (\alpha = 0{,}05)$, se rechaza H_0. Por tanto, se concluye que la media en *CI* es distinta en la población de niños autistas.

2. H_0: $\mu \leq 15$; H_1: $\mu > 15$.

$Z = 2$; $p = 0{,}0228$.

No se puede rechazar H_0. Por tanto, los datos indican que no se puede concluir que los niños que ven habitualmente películas violentas sean más agresivos.

3. $N = 16$.

4. $H_0: \mu = 90$; $H_1: \mu \neq 90$.

$T = 1{,}285$; $p > 0{,}10$.

No se puede rechazar H_0. Por tanto, los datos son compatibles con la hipótesis de que la media poblacional es igual a 90.

5. $H_0: \mu = 28$; $H_1: \mu \neq 28$.

$T = -1{,}473$; $p > 0{,}10$.

No se puede rechazar H_0. Por tanto, no existe evidencia para concluir que el salario medio es diferente en hombres y mujeres.

6. $H_0: \mu \leq 5{,}8$; $H_1: \mu > 5{,}8$.

$T = 0{,}92$; $p > 0{,}10$.

No se puede rechazar H_0. Por tanto, no se puede concluir que el promedio obtenido en la prueba de matemáticas se haya incrementado en el curso 2010-2011.

7. $H_0: \rho \leq 0$; $H_1: \rho > 0$.

$T = 11{,}55$; $p < 0{,}005$.

Se rechaza H_0. Por tanto, existe una relación lineal de tipo directa entre salario y absentismo.

8. $H_0: \rho = 0$; $H_1: \rho \neq 0$.

$T = -4{,}60$; $p < 0{,}02$.

Se rechaza H_0. Por tanto, existe una relación lineal entre la fatiga y la calificación obtenida.

9. $H_0: \rho = 0$; $H_1: \rho \neq 0$.

$T = -5{,}09$; $p < 0{,}005$.

Se rechaza H_0. Por tanto, la correlación entre satisfacción con el sueldo obtenido y oportunidades de promoción es estadísticamente significativa.

10. $H_0: \pi \geq 0{,}80$; $H_1: \pi < 0{,}80$.

$Z = -1{,}77$; $p = 0{,}0384$.

Se rechaza H_0. Se puede concluir que la audiencia de televisión ha disminuido.

11. H_0: $\pi \leq 0{,}60$; H_1: $\pi > 0{,}60$.

$Z = 1{,}00$; $p = 0{,}1587$.

Se mantiene H_0. Los datos no permiten concluir que se haya producido un aumento del rechazo a la reducción de la pena.

12. H_0: $\pi = 0{,}50$; H_1: $\pi \neq 0{,}50$.

$Z = -2{,}24$; $p = 0{,}025$.

Se mantiene H_0. Los datos son compatibles con la hipótesis de que los trabajadores son indiferentes.

13. H_0: $\pi \leq 0{,}75$; H_1: $\pi > 0{,}75$.

$Z = 1{,}41$; $p = 0{,}0793$.

Se mantiene H_0. Los datos son compatibles con la hipótesis de que el participante no ha percibido el estímulo.

14. H_0: La actitud (aceptación-rechazo) es independiente del sexo.
H_1: La actitud (aceptación-rechazo) no es independiente del sexo.

$X^2 = 12{,}500$; $0{,}10 < p < 0{,}20$.

Se rechaza H_0. La actitud ante la reducción de la pena no es independiente del sexo.

15. H_0: La depresión es independiente del nivel de estudios.
H_1: La depresión no es independiente del nivel de estudios.

$X^2 = 3{,}785$; $0{,}10 < p < 0{,}20$.

Se mantiene H_0. Los datos son compatibles con la hipótesis de independencia entre depresión y nivel de estudios.

16. H_0: El material del juego y el número de colores son independientes.
H_1: El material del juego y el número de colores no son independientes.

$X^2 = 18{,}737$; $p < 0{,}001$.

Se rechaza H_0. Se puede concluir que existe relación entre material y color.

17. H_0: El tipo de trastorno es independiente del medio en el que se vive.
H_1: El tipo de trastorno no es independiente del medio en el que se vive.

$X^2 = 7{,}418$; $p < 0{,}01$.

Se rechaza H_0. Se puede concluir que existe relación entre trastorno y medio.

18. Para que sea significativa, r_{xy} debe ser mayor que 0,215 o menor que −0,215.

APÉNDICE

CH sobre proporciones con muestras pequeñas

La mayor parte de los CH sobre proporciones se harán en condiciones en las que es apropiado emplear la aproximación de la binomial a la normal, pero en algunos casos no será así. El CH en estas circunstancias tiene alguna complicación adicional, debido a la naturaleza discreta de la variable, pero precisamente porque se utiliza muy poco vamos a hacer sólo una descripción somera en este apéndice.

Supongamos que queremos contrastar la hipótesis de que una rata no ha sido enseñada en relación con un laberinto con dos salidas, una roja y otra azul. De ser verdadera esta hipótesis, las probabilidades de salir por cada lado serán 0,50. Supongamos que introducimos al animal ocho veces en el laberinto y anotamos las veces que sale por cada lado. ¿Cómo establecemos la regla de decisión, sabiendo que el estadístico «número de veces que sale por la salida roja» (igual podríamos hacerlo con la azul) es una variable aleatoria discreta que se distribuye según el modelo binomial?

Si la hipótesis de que no ha sido enseñada (H_0) es verdadera, la distribución de probabilidad de la variable X = «número de veces que sale por la roja» se distribuye $B(8; 0{,}50)$; según la tabla de la binomial, su distribución de probabilidad es:

X_i	0	1	2	3	4	5	6	7	8
$f(x_i)$	0,004	0,031	0,109	0,219	0,273	0,219	0,109	0,031	0,004

Supongamos que queremos contrastar H_0 frente a la alternativa de que ha sido reforzada para salir por la roja, con un nivel de significación de $\alpha = 0{,}05$. Se trata de un contraste unilateral derecho, dado que si la rata ha sido reforzada al salir por la roja, la alternativa es que la probabilidad de salir por la roja es *mayor* que la del azar (H_0: $\pi \leq 0{,}50$; H_1: $\pi > 0{,}50$).

Para establecer una regla de decisión con $\alpha = 0{,}05$ tendríamos que buscar un conjunto de valores extremos por el lado indicado en H_1, cuya probabilidad conjunta sea igual a α. Pero esto no se puede conseguir porque se trata de una variable discreta. Con la regla «rechazar si sale por la roja 7 veces o más» el nivel de significación será 0,035 (0,031 + 0,004) mientras que con la regla «rechazar si sale por la roja 6 veces o más» será igual a 0,144 (0,109 + 0,031 + 0,004). En un caso el valor real de α es mayor del deseado y en el otro es menor.

Aparte de este problema, el esquema del contraste sería igual a los que hemos ido describiendo en este capítulo.

Tabla resumen con esquemas para el CH

El estudiante que ha llegado hasta aquí ya se habrá hecho a la idea de que una de sus líneas de progreso en cursos posteriores consistirá en ir añadiendo esquemas como los expuestos en este capítulo, pero aplicados a escenarios analí-

ticos diferentes. Cada esquema será una valiosa herramienta que podrá aplicar llegado el momento adecuado. Para ayudar a que vaya elaborando su propio plan, hemos confeccionado la siguiente tabla. En ella hay casillas vacías que se refieren a casos que no hemos abordado en este capítulo. También le faltan las filas correspondientes a otras situaciones, de las que hemos incluido algunas de las que habrá que abordar en un futuro próximo.

CONTRASTES CON UNA MUESTRA		
Parámetro	**Supuestos y condiciones**	**Estadístico y distribución**
Media	— Normalidad o no normalidad con muestra grande — Conocida σ — m.a.s.	$Z = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}; \quad Z \sim N(0; 1)$
	— Normalidad o no normalidad con muestra grande — Desconocida σ — m.a.s.	$T = \dfrac{\bar{X} - \mu_0}{S_{N-1}/\sqrt{N}}; \quad T \sim t_{N-1}$ $T = \dfrac{\bar{X} - \mu_0}{S_N/\sqrt{N-1}}; \quad T \sim t_{N-1}$
Correlación	— Normalidad — Hipótesis de independencia lineal, $\rho = 0$ — m.a.s.	$T = \dfrac{r_{xy} \cdot \sqrt{N-2}}{\sqrt{1 - r_{xy}^2}}; \quad T \sim t_{N-2}$
	— Normalidad — Hipótesis sobre otros valores, diferentes de 0 — m.a.s.	**?**
Proporción	— Variable dicotómica — N observaciones independientes — Muestra grande ($N \cdot \pi > 5$)	$Z = \dfrac{P - \pi_0}{\sqrt{\dfrac{\pi_0 \cdot (1 - \pi_0)}{N}}}; \quad Z \sim N(0; 1)$
	— Variable dicotómica — N observaciones independientes — Muestra pequeña ($N \cdot \pi < 5$)	Véase el apéndice de este capítulo
Independencia de variables categóricas	— N observaciones independientes — Frecuencias esperadas mayores de 5	$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \dfrac{(n_{ij} - m_{ij})^2}{m_{ij}}; \quad X^2 \sim \chi^2_{(I-1)(J-1)}$
Varianza		**?**
Diferencia de medias	— Muestras independientes	**?**
	— Muestras relacionadas	**?**
Otras	**?**	**?**

APÉNDICE FINAL
Tablas estadísticas

TABLA I

Función de probabilidad binomial (tres primeros decimales)

$$B\,(N;\,\pi) = \binom{N}{x}\pi^{x}(1-\pi)^{N-x}$$

		π															
N	*x*	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	0,95	0,99	*x*
2	0	980	903	810	640	563	490	360	250	160	090	063	040	010	003	0+	0
	1	020	095	180	320	375	420	480	500	480	420	375	320	180	095	020	1
	2	0+	003	010	040	063	090	160	250	360	490	563	640	810	903	980	2
3	0	970	857	729	512	422	343	216	125	064	027	016	008	001	0+	0+	0
	1	029	135	243	384	422	441	432	375	288	189	141	096	027	007	0+	1
	2	0+	007	027	096	141	189	288	375	432	441	422	384	243	135	029	2
	3	0+	0+	001	008	016	027	064	125	216	343	422	512	729	857	970	3
4	0	961	815	656	410	316	240	130	063	026	008	004	002	0+	0+	0+	0
	1	039	171	292	410	422	412	346	250	154	076	047	026	004	0+	0+	1
	2	001	014	049	154	211	265	346	375	346	265	211	154	049	014	001	2
	3	0+	0+	004	026	047	076	154	250	346	412	422	410	292	171	039	3
	4	0+	0+	0+	002	004	008	026	063	130	240	316	410	656	815	961	4
5	0	951	774	590	328	237	168	078	031	010	002	001	0+	0+	0+	0+	0
	1	048	204	328	410	396	360	259	156	077	028	015	006	0+	0+	0+	1
	2	001	021	073	205	264	309	346	313	230	132	088	051	008	001	0+	2
	3	0+	001	008	051	088	132	230	313	346	309	264	205	073	021	001	3
	4	0+	0+	0+	006	015	028	077	156	259	360	396	410	328	204	048	4
	5	0+	0+	0+	0+	001	002	010	031	078	168	237	328	590	774	951	5
6	0	941	735	531	262	178	118	047	016	004	001	0+	0+	0+	0+	0+	0
	1	057	232	354	393	356	303	187	094	037	010	004	002	0+	0+	0+	1
	2	001	031	098	246	297	324	311	234	138	060	033	015	001	0+	0+	2
	3	0+	002	015	082	132	185	276	313	276	185	132	082	015	002	0+	3
	4	0+	0+	001	015	033	060	138	234	311	324	297	246	098	031	001	4
	5	0+	0+	0+	002	004	010	037	094	187	303	356	393	354	232	057	5
	6	0+	0+	0+	0+	0+	001	004	016	047	118	178	262	531	735	941	6
7	0	932	698	478	210	133	082	028	008	002	0+	0+	0+	0+	0+	0+	0
	1	066	257	372	367	311	247	131	055	017	004	001	0+	0+	0+	0+	1
	2	002	041	124	275	311	318	261	164	077	025	012	004	0+	0+	0+	2
	3	0+	004	023	115	173	227	290	273	194	097	058	029	003	0+	0+	3
	4	0+	0+	003	029	058	097	194	273	290	227	173	115	023	004	0+	4
	5	0+	0+	0+	004	012	025	077	164	261	318	311	275	124	041	002	5
	6	0+	0+	0+	0+	001	004	017	055	131	247	311	367	372	257	066	6
	7	0+	0+	0+	0+	0+	0+	002	008	028	082	133	210	478	698	932	7
8	0	923	663	430	168	100	058	017	004	001	0+	0+	0+	0+	0+	0+	0
	1	075	279	383	336	267	198	090	031	008	001	0+	0+	0+	0+	0+	1
	2	003	051	149	294	311	296	209	109	041	010	004	001	0+	0+	0+	2
	3	0+	005	033	147	208	254	279	219	124	047	023	009	0+	0+	0+	3
	4	0+	0+	005	046	087	136	232	273	232	136	087	046	005	0+	0+	4
	5	0+	0+	0+	009	023	047	124	219	279	254	208	147	033	005	0+	5
	6	0+	0+	0+	001	004	010	041	109	209	296	311	294	149	051	003	6
	7	0+	0+	0+	0+	0+	001	008	031	090	198	267	336	383	279	075	7
	8	0+	0+	0+	0+	0+	0+	001	004	017	058	100	168	430	663	923	8

TABLA I *(continuación)*

		π															
N	x	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	0,95	0,99	*x*
9	0	914	630	387	134	075	040	010	002	0+	0+	0+	0+	0+	0+	0+	0
	1	083	299	387	302	225	156	060	018	004	0+	0+	0+	0+	0+	0+	1
	2	003	063	172	302	300	267	161	070	021	004	001	0+	0+	0+	0+	2
	3	0+	008	045	176	234	267	251	164	074	021	009	003	0+	0+	0+	3
	4	0+	001	007	066	117	172	251	246	167	074	039	017	001	0+	0+	4
	5	0+	0+	001	017	039	074	167	246	251	172	117	066	007	001	0+	5
	6	0+	0+	0+	003	009	021	074	164	251	267	234	176	045	008	0+	6
	7	0+	0+	0+	0+	001	004	021	070	161	267	300	302	172	063	003	7
	8	0+	0+	0+	0+	0+	0+	004	018	060	156	225	302	387	299	083	8
	9	0+	0+	0+	0+	0+	0+	0+	002	010	040	075	134	387	630	914	9
10	0	904	599	349	107	056	028	006	001	0+	0+	0+	0+	0+	0+	0+	0
	1	091	315	387	268	188	121	040	010	002	0+	0+	0+	0+	0+	0+	1
	2	004	075	194	302	282	233	121	044	011	001	0+	0+	0+	0+	0+	2
	3	0+	010	057	201	250	267	215	117	042	009	003	001	0+	0+	0+	3
	4	0+	001	011	088	146	200	251	205	111	037	016	006	0+	0+	0+	4
	5	0+	0+	001	026	058	103	201	246	201	103	058	026	001	0+	0+	5
	6	0+	0+	0+	006	016	037	111	205	251	200	146	088	011	001	0+	6
	7	0+	0+	0+	001	003	009	042	117	215	267	250	201	057	010	0+	7
	8	0+	0+	0+	0+	0+	001	011	044	121	233	282	302	194	075	004	8
	9	0+	0+	0+	0+	0+	0+	002	010	040	121	188	268	387	315	091	9
	10	0+	0+	0+	0+	0+	0+	0+	001	006	028	056	107	349	599	904	10
11	0	895	569	314	086	042	020	004	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	099	329	384	236	155	093	027	005	001	0+	0+	0+	0+	0+	0+	1
	2	005	087	213	295	258	200	089	027	005	001	0+	0+	0+	0+	0+	2
	3	0+	014	071	221	258	257	177	081	023	004	001	0+	0+	0+	0+	3
	4	0+	001	016	111	172	220	236	161	070	017	006	002	0+	0+	0+	4
	5	0+	0+	002	039	080	132	221	226	147	057	027	010	0+	0+	0+	5
	6	0+	0+	0+	010	027	057	147	226	221	132	080	039	002	0+	0+	6
	7	0+	0+	0+	002	006	017	070	161	236	220	172	111	016	001	0+	7
	8	0+	0+	0+	0+	001	004	023	081	177	257	258	221	071	014	0+	8
	9	0+	0+	0+	0+	0+	001	005	027	089	200	258	295	213	087	005	9
	10	0+	0+	0+	0+	0+	0+	001	005	027	093	155	236	384	329	099	10
	11	0+	0+	0+	0+	0+	0+	0+	0+	004	020	042	086	314	569	895	11
12	0	886	540	282	069	032	014	002	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	107	341	377	206	127	071	017	003	0+	0+	0+	0+	0+	0+	0+	1
	2	006	099	230	283	232	168	064	016	002	0+	0+	0+	0+	0+	0+	2
	3	0+	017	085	236	258	240	142	054	012	001	0+	0+	0+	0+	0+	3
	4	0+	002	021	133	194	231	213	121	042	008	002	001	0+	0+	0+	4
	5	0+	0+	004	053	103	158	227	193	101	029	011	003	0+	0+	0+	5
	6	0+	0+	0+	016	040	079	177	226	177	079	040	016	0+	0+	0+	6
	7	0+	0+	0+	003	011	029	101	193	227	158	103	053	004	0+	0+	7
	8	0+	0+	0+	001	002	008	042	121	213	231	194	133	021	002	0+	8
	9	0+	0+	0+	0+	0+	001	012	054	142	240	258	236	085	017	0+	9
	10	0+	0+	0+	0+	0+	0+	002	016	064	168	232	283	230	099	006	10
	11	0+	0+	0+	0+	0+	0+	0+	003	017	071	127	206	377	341	107	11
	12	0+	0+	0+	0+	0+	0+	0+	0+	002	014	032	069	282	540	886	12

TABLA I *(continuación)*

		π															
N	*x*	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	0,95	0,99	*x*
13	0	878	513	254	055	024	010	001	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	115	351	367	179	103	054	011	002	0+	0+	0+	0+	0+	0+	0+	1
	2	007	111	245	268	206	139	045	010	001	0+	0+	0+	0+	0+	0+	2
	3	0+	021	100	246	252	218	111	035	006	001	0+	0+	0+	0+	0+	3
	4	0+	003	028	154	210	234	184	087	024	003	001	0+	0+	0+	0+	4
	5	0+	0+	006	069	126	180	221	157	066	014	005	001	0+	0+	0+	5
	6	0+	0+	001	023	056	103	197	209	131	044	019	006	0+	0+	0+	6
	7	0+	0+	0+	006	019	044	131	209	197	103	056	023	001	0+	0+	7
	8	0+	0+	0+	001	005	014	066	157	221	180	126	069	006	0+	0+	8
	9	0+	0+	0+	0+	001	003	024	087	184	234	210	154	028	003	0+	9
	10	0+	0+	0+	0+	0+	001	006	035	111	218	252	246	100	021	0+	10
	11	0+	0+	0+	0+	0+	0+	001	010	045	139	206	268	245	111	007	11
	12	0+	0+	0+	0+	0+	0+	0+	002	011	054	103	179	367	351	115	12
	13	0+	0+	0+	0+	0+	0+	0+	0+	001	010	024	055	254	513	878	13
14	0	869	488	229	044	018	007	001	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	123	359	356	154	083	041	007	001	0+	0+	0+	0+	0+	0+	0+	1
	2	008	123	257	250	180	113	032	006	001	0+	0+	0+	0+	0+	0+	2
	3	0+	026	114	250	240	194	085	022	003	0+	0+	0+	0+	0+	0+	3
	4	0+	004	035	172	220	229	155	061	014	001	0+	0+	0+	0+	0+	4
	5	0+	0+	008	086	147	196	207	122	041	007	002	0+	0+	0+	0+	5
	6	0+	0+	001	032	073	126	207	183	092	023	008	002	0+	0+	0+	6
	7	0+	0+	0+	009	028	062	157	209	157	062	028	009	0+	0+	0+	7
	8	0+	0+	0+	002	008	023	092	183	207	126	073	032	001	0+	0+	8
	9	0+	0+	0+	0+	002	007	041	122	207	196	147	086	008	0+	0+	9
	10	0+	0+	0+	0+	0+	001	014	061	155	229	220	172	035	004	0+	10
	11	0+	0+	0+	0+	0+	0+	003	022	085	194	240	250	114	026	0+	11
	12	0+	0+	0+	0+	0+	0+	001	006	032	113	180	250	257	123	008	12
	13	0+	0+	0+	0+	0+	0+	0+	001	007	041	083	154	356	359	123	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	001	007	018	044	229	488	869	14
15	0	860	463	206	035	013	005	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	130	366	343	132	067	031	005	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	009	135	267	231	156	092	022	003	0+	0+	0+	0+	0+	0+	0+	2
	3	0+	031	129	250	225	170	063	014	002	0+	0+	0+	0+	0+	0+	3
	4	0+	005	043	188	225	219	127	042	007	001	0+	0+	0+	0+	0+	4
	5	0+	001	010	103	165	206	186	092	024	003	001	0+	0+	0+	0+	5
	6	0+	0+	002	043	092	147	207	153	061	012	003	001	0+	0+	0+	6
	7	0+	0+	0+	014	039	081	177	196	118	035	013	003	0+	0+	0+	7
	8	0+	0+	0+	003	013	035	118	196	177	081	039	014	0+	0+	0+	8
	9	0+	0+	0+	001	003	012	061	153	207	147	092	043	002	0+	0+	9
	10	0+	0+	0+	0+	001	003	024	092	186	206	165	103	010	001	0+	10
	11	0+	0+	0+	0+	0+	001	007	042	127	219	225	188	043	005	0+	11
	12	0+	0+	0+	0+	0+	0+	002	014	063	170	225	250	129	031	0+	12
	13	0+	0+	0+	0+	0+	0+	0+	003	022	092	156	231	267	135	009	13
	14	0+	0+	0+	0+	0+	0+	0+	0+	005	031	067	132	343	366	130	14
	15	0+	0+	0+	0+	0+	0+	0+	0+	0+	005	013	035	206	463	860	15
16	0	851	440	185	028	010	003	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	138	371	329	113	053	023	003	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	010	146	275	211	134	073	015	002	0+	0+	0+	0+	0+	0+	0+	2

TABLA I *(continuación)*

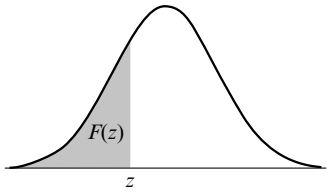
		π															
N	*x*	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	0,95	0,99	*x*
	3	0+	036	142	246	208	146	047	009	001	0+	0+	0+	0+	0+	0+	3
	4	0+	006	051	200	225	204	101	028	004	0+	0+	0+	0+	0+	0+	4
	5	0+	001	014	120	180	210	162	067	014	001	0+	0+	0+	0+	0+	5
	6	0+	0+	003	055	110	165	198	122	039	006	001	0+	0+	0+	0+	6
	7	0+	0+	0+	020	052	101	189	175	084	019	006	001	0+	0+	0+	7
	8	0+	0+	0+	006	020	049	142	196	142	049	020	006	0+	0+	0+	8
	9	0+	0+	0+	001	006	019	084	175	189	101	052	020	0+	0+	0+	9
	10	0+	0+	0+	0+	001	006	039	122	198	165	110	055	003	0+	0+	10
	11	0+	0+	0+	0+	0+	001	014	067	162	210	180	120	014	001	0+	11
	12	0+	0+	0+	0+	0+	0+	004	028	101	204	225	200	051	006	0+	12
	13	0+	0+	0+	0+	0+	0+	001	009	047	146	208	246	142	036	0+	13
	14	0+	0+	0+	0+	0+	0+	0+	002	015	073	134	211	275	146	010	14
	15	0+	0+	0+	0+	0+	0+	0+	0+	003	023	053	113	329	371	138	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	0+	003	010	028	185	440	851	16
17	0	843	418	167	023	008	002	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	145	374	315	096	043	017	002	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	012	158	280	191	114	058	010	001	0+	0+	0+	0+	0+	0+	0+	2
	3	001	041	156	239	189	125	034	005	0+	0+	0+	0+	0+	0+	0+	3
	4	0+	008	060	209	221	187	080	018	002	0+	0+	0+	0+	0+	0+	4
	5	0+	001	017	136	191	208	138	047	008	001	0+	0+	0+	0+	0+	5
	6	0+	0+	004	068	128	178	184	094	024	003	001	0+	0+	0+	0+	6
	7	0+	0+	001	027	067	120	193	148	057	009	002	0+	0+	0+	0+	7
	8	0+	0+	0+	008	028	064	161	185	107	028	009	002	0+	0+	0+	8
	9	0+	0+	0+	002	009	028	107	185	161	064	028	008	0+	0+	0+	9
	10	0+	0+	0+	0+	002	009	057	148	193	120	067	027	001	0+	0+	10
	11	0+	0+	0+	0+	001	003	024	094	184	178	128	068	004	0+	0+	11
	12	0+	0+	0+	0+	0+	001	008	047	138	208	191	136	017	001	0+	12
	13	0+	0+	0+	0+	0+	0+	002	018	080	187	221	209	060	008	0+	13
	14	0+	0+	0+	0+	0+	0+	0+	005	034	125	189	239	156	041	001	14
	15	0+	0+	0+	0+	0+	0+	0+	001	010	058	114	191	280	158	012	15
	16	0+	0+	0+	0+	0+	0+	0+	0+	002	017	043	096	315	374	145	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	008	023	167	418	843	17
18	0	835	397	150	018	006	002	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	152	376	300	081	034	013	001	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	013	168	284	172	096	046	007	001	0+	0+	0+	0+	0+	0+	0+	2
	3	001	047	168	230	170	105	025	003	0+	0+	0+	0+	0+	0+	0+	3
	4	0+	009	070	215	213	168	061	012	001	0+	0+	0+	0+	0+	0+	4
	5	0+	001	022	151	199	202	115	033	004	0+	0+	0+	0+	0+	0+	5
	6	0+	0+	005	082	144	187	166	071	015	001	0+	0+	0+	0+	0+	6
	7	0+	0+	001	035	082	138	189	121	037	005	001	0+	0+	0+	0+	7
	8	0+	0+	0+	012	038	081	173	167	077	015	004	001	0+	0+	0+	8
	9	0+	0+	0+	003	014	039	128	185	128	039	014	003	0+	0+	0+	9
	10	0+	0+	0+	001	004	015	077	167	173	081	038	012	0+	0+	0+	10
	11	0+	0+	0+	0+	001	005	037	121	189	138	082	035	001	0+	0+	11
	12	0+	0+	0+	0+	0+	001	015	071	166	187	144	082	005	0+	0+	12
	13	0+	0+	0+	0+	0+	0+	004	033	115	202	199	151	022	001	0+	13
	14	0+	0+	0+	0+	0+	0+	001	012	061	168	213	215	070	009	0+	14
	15	0+	0+	0+	0+	0+	0+	0+	003	025	105	170	230	168	047	001	15
	16	0+	0+	0+	0+	0+	0+	0+	001	007	046	096	172	284	168	013	16

TABLA I *(continuación)*

		π															
N	*x*	0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,75	0,80	0,90	0,95	0,99	*x*
	17	0+	0+	0+	0+	0+	0+	0+	0+	001	013	034	081	300	376	152	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	0+	002	006	018	150	397	835	18
19	0	826	377	135	014	004	001	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	159	377	285	068	027	009	001	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	014	179	285	154	080	036	005	0+	0+	0+	0+	0+	0+	0+	0+	2
	3	001	053	180	218	152	087	017	002	0+	0+	0+	0+	0+	0+	0+	3
	4	0+	011	080	218	202	149	047	007	001	0+	0+	0+	0+	0+	0+	4
	5	0+	002	027	164	202	192	093	022	002	0+	0+	0+	0+	0+	0+	5
	6	0+	0+	007	095	157	192	145	052	008	001	0+	0+	0+	0+	0+	6
	7	0+	0+	001	044	097	153	180	096	024	002	0+	0+	0+	0+	0+	7
	8	0+	0+	0+	017	049	098	180	144	053	008	002	0+	0+	0+	0+	8
	9	0+	0+	0+	005	020	051	146	176	098	022	007	001	0+	0+	0+	9
	10	0+	0+	0+	001	007	022	098	176	146	051	020	005	0+	0+	0+	10
	11	0+	0+	0+	0+	002	008	053	144	180	098	049	017	0+	0+	0+	11
	12	0+	0+	0+	0+	0+	002	024	096	180	153	097	044	001	0+	0+	12
	13	0+	0+	0+	0+	0+	001	008	052	145	192	157	095	007	0+	0+	13
	14	0+	0+	0+	0+	0+	0+	002	022	093	192	202	164	027	002	0+	14
	15	0+	0+	0+	0+	0+	0+	001	007	047	149	202	218	080	011	0+	15
	16	0+	0+	0+	0+	0+	0+	0+	002	017	087	152	218	180	053	001	16
	17	0+	0+	0+	0+	0+	0+	0+	0+	005	036	080	154	285	179	014	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	001	009	027	068	285	377	159	18
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	004	014	135	377	826	19
20	0	818	358	122	012	003	001	0+	0+	0+	0+	0+	0+	0+	0+	0+	0
	1	165	377	270	058	021	007	0+	0+	0+	0+	0+	0+	0+	0+	0+	1
	2	016	189	285	137	067	028	003	0+	0+	0+	0+	0+	0+	0+	0+	2
	3	001	060	190	205	134	072	012	001	0+	0+	0+	0+	0+	0+	0+	3
	4	0+	013	090	218	190	130	035	005	0+	0+	0+	0+	0+	0+	0+	4
	5	0+	002	032	175	202	179	075	015	001	0+	0+	0+	0+	0+	0+	5
	6	0+	0+	009	109	169	192	124	037	005	0+	0+	0+	0+	0+	0+	6
	7	0+	0+	002	055	112	164	166	074	015	001	0+	0+	0+	0+	0+	7
	8	0+	0+	0+	022	061	114	180	120	035	004	001	0+	0+	0+	0+	8
	9	0+	0+	0+	007	027	065	160	160	071	012	003	0+	0+	0+	0+	9
	10	0+	0+	0+	002	010	031	117	176	117	031	010	002	0+	0+	0+	10
	11	0+	0+	0+	0+	003	012	071	160	160	065	027	007	0+	0+	0+	11
	12	0+	0+	0+	0+	001	004	035	120	180	114	061	022	0+	0+	0+	12
	13	0+	0+	0+	0+	0+	001	015	074	166	164	112	055	002	0+	0+	13
	14	0+	0+	0+	0+	0+	0+	005	037	124	192	169	109	009	0+	0+	14
	15	0+	0+	0+	0+	0+	0+	001	015	075	179	202	175	032	002	0+	15
	16	0+	0+	0+	0+	0+	0+	0+	005	035	130	190	218	090	013	0+	16
	17	0+	0+	0+	0+	0+	0+	0+	001	012	072	134	205	190	060	001	17
	18	0+	0+	0+	0+	0+	0+	0+	0+	003	028	067	137	285	189	016	18
	19	0+	0+	0+	0+	0+	0+	0+	0+	0+	007	021	058	270	377	165	19
	20	0+	0+	0+	0+	0+	0+	0+	0+	0+	001	003	012	122	358	818	20

TABLA II

Función de distribución normal tipificada



$$F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{-\frac{x^2}{2}} \cdot dx$$

z	$F(z)$	z	$F(z)$	z	$F(z)$
-3,00	0,0013				
-2,99	0,0014	-2,64	0,0041	-2,29	0,0110
-2,98	0,0014	-2,63	0,0043	-2,28	0,0113
-2,97	0,0015	-2,62	0,0044	-2,27	0,0116
-2,96	0,0015	-2,61	0,0045	-2,26	0,0119
-2,95	0,0016	-2,60	0,0047	-2,25	0,0122
-2,94	0,0016	-2,59	0,0048	-2,24	0,0125
-2,93	0,0017	-2,58	0,0049	-2,23	0,0129
-2,92	0,0018	-2,57	0,0051	-2,22	0,0132
-2,91	0,0018	-2,56	0,0052	-2,21	0,0136
-2,90	0,0019	-2,55	0,0054	-2,20	0,0139
-2,89	0,0019	-2,54	0,0055	-2,19	0,0143
-2,88	0,0020	-2,53	0,0057	-2,18	0,0146
-2,87	0,0021	-2,52	0,0059	-2,17	0,0150
-2,86	0,0021	-2,51	0,0060	-2,16	0,0154
-2,85	0,0022	-2,50	0,0062	-2,15	0,0158
-2,84	0,0023	-2,49	0,0064	-2,14	0,0162
-2,83	0,0023	-2,48	0,0066	-2,13	0,0166
-2,82	0,0024	-2,47	0,0068	-2,12	0,0170
-2,81	0,0025	-2,46	0,0069	-2,11	0,0174
-2,80	0,0026	-2,45	0,0071	-2,10	0,0179
-2,79	0,0026	-2,44	0,0073	-2,09	0,0183
-2,78	0,0027	-2,43	0,0075	-2,08	0,0188
-2,77	0,0028	-2,42	0,0078	-2,07	0,0192
-2,76	0,0029	-2,41	0,0080	-2,06	0,0197
-2,75	0,0030	-2,40	0,0082	-2,05	0,0202
-2,74	0,0031	-2,39	0,0084	-2,04	0,0207
-2,73	0,0032	-2,38	0,0087	-2,03	0,0212
-2,72	0,0033	-2,37	0,0089	-2,02	0,0217
-2,71	0,0034	-2,36	0,0091	-2,01	0,0222
-2,70	0,0035	-2,35	0,0094	-2,00	0,0228
-2,69	0,0036	-2,34	0,0096	-1,99	0,0233
-2,68	0,0037	-2,33	0,0099	-1,98	0,0239
-2,67	0,0038	-2,32	0,0102	-1,97	0,0244
-2,66	0,0039	-2,31	0,0104	-1,96	0,0250
-2,65	0,0040	-2,30	0,0107	-1,95	0,0256

TABLA II *(continuación)*

z	$F(z)$	z	$F(z)$	z	$F(z)$
-1,94	0,0262	-1,49	0,0681	-1,04	0,1492
-1,93	0,0268	-1,48	0,0694	-1,03	0,1515
-1,92	0,0274	-1,47	0,0708	-1,02	0,1539
-1,91	0,0281	-1,46	0,0721	-1,01	0,1562
-1,90	0,0287	-1,45	0,0735	-1,00	0,1587
-1,89	0,0294	-1,44	0,0749	-0,99	0,1611
-1,88	0,0301	-1,43	0,0764	-0,98	0,1635
-1,87	0,0307	-1,42	0,0778	-0,97	0,1660
-1,86	0,0314	-1,41	0,0793	-0,96	0,1685
-1,85	0,0322	-1,40	0,0808	-0,95	0,1711
-1,84	0,0329	-1,39	0,0823	-0,94	0,1736
-1,83	0,0336	-1,38	0,0838	-0,93	0,1762
-1,82	0,0344	-1,37	0,0853	-0,92	0,1788
-1,81	0,0351	-1,36	0,0869	-0,91	0,1814
-1,80	0,0359	-1,35	0,0885	-0,90	0,1841
-1,79	0,0367	-1,34	0,0901	-0,89	0,1867
-1,78	0,0375	-1,33	0,0918	-0,88	0,1894
-1,77	0,0384	-1,32	0,0934	-0,87	0,1922
-1,76	0,0392	-1,31	0,0951	-0,86	0,1949
-1,75	0,0401	-1,30	0,0968	-0,85	0,1977
-1,74	0,0409	-1,29	0,0985	-0,84	0,2005
-1,73	0,0418	-1,28	0,1003	-0,83	0,2033
-1,72	0,0427	-1,27	0,1020	-0,82	0,2061
-1,71	0,0436	-1,26	0,1038	-0,81	0,2090
-1,70	0,0446	-1,25	0,1056	-0,80	0,2119
-1,69	0,0455	-1,24	0,1075	-0,79	0,2148
-1,68	0,0465	-1,23	0,1093	-0,78	0,2177
-1,67	0,0475	-1,22	0,1112	-0,77	0,2206
-1,66	0,0485	-1,21	0,1131	-0,76	0,2236
-1,65	0,0495	-1,20	0,1151	-0,75	0,2266
-1,64	0,0505	-1,19	0,1170	-0,74	0,2296
-1,63	0,0516	-1,18	0,1190	-0,73	0,2327
-1,62	0,0526	-1,17	0,1210	-0,72	0,2358
-1,61	0,0537	-1,16	0,1230	-0,71	0,2389
-1,60	0,0548	-1,15	0,1251	-0,70	0,2420
-1,59	0,0559	-1,14	0,1271	-0,69	0,2451
-1,58	0,0571	-1,13	0,1292	-0,68	0,2483
-1,57	0,0582	-1,12	0,1314	-0,67	0,2514
-1,56	0,0594	-1,11	0,1335	-0,66	0,2546
-1,55	0,0606	-1,10	0,1357	-0,65	0,2578
-1,54	0,0618	-1,09	0,1379	-0,64	0,2611
-1,53	0,0630	-1,08	0,1401	-0,63	0,2643
-1,52	0,0643	-1,07	0,1423	-0,62	0,2676
-1,51	0,0655	-1,06	0,1446	-0,61	0,2709
-1,50	0,0668	-1,05	0,1469	-0,60	0,2743

TABLA II *(continuación)*

z	$F(z)$	z	$F(z)$	z	$F(z)$
-0,59	0,2776	-0,14	0,4443	0,31	0,6217
-0,58	0,2810	-0,13	0,4483	0,32	0,6255
-0,57	0,2843	-0,12	0,4522	0,33	0,6293
-0,56	0,2877	-0,11	0,4562	0,34	0,6331
-0,55	0,2912	-0,10	0,4602	0,35	0,6368
-0,54	0,2946	-0,09	0,4641	0,36	0,6406
-0,53	0,2981	-0,08	0,4681	0,37	0,6443
-0,52	0,3015	-0,07	0,4721	0,38	0,6480
-0,51	0,3050	-0,06	0,4761	0,39	0,6517
-0,50	0,3085	-0,05	0,4801	0,40	0,6554
-0,49	0,3121	-0,04	0,4840	0,41	0,6591
-0,48	0,3156	-0,03	0,4880	0,42	0,6628
-0,47	0,3192	-0,02	0,4920	0,43	0,6664
-0,46	0,3228	-0,01	0,4960	0,44	0,6700
-0,45	0,3264	0,00	0,5000	0,45	0,6736
-0,44	0,3300	0,01	0,5040	0,46	0,6772
-0,43	0,3336	0,02	0,5080	0,47	0,6808
-0,42	0,3372	0,03	0,5120	0,48	0,6844
-0,41	0,3409	0,04	0,5160	0,49	0,6879
-0,40	0,3446	0,05	0,5199	0,50	0,6915
-0,39	0,3483	0,06	0,5239	0,51	0,6950
-0,38	0,3520	0,07	0,5279	0,52	0,6985
-0,37	0,3557	0,08	0,5319	0,53	0,7019
-0,36	0,3594	0,09	0,5359	0,54	0,7054
-0,35	0,3632	0,10	0,5398	0,55	0,7088
-0,34	0,3669	0,11	0,5438	0,56	0,7123
-0,33	0,3707	0,12	0,5478	0,57	0,7157
-0,32	0,3745	0,13	0,5517	0,58	0,7190
-0,31	0,3783	0,14	0,5557	0,59	0,7224
-0,30	0,3821	0,15	0,5596	0,60	0,7257
-0,29	0,3859	0,16	0,5636	0,61	0,7291
-0,28	0,3897	0,17	0,5675	0,62	0,7324
-0,27	0,3936	0,18	0,5714	0,63	0,7357
-0,26	0,3974	0,19	0,5753	0,64	0,7389
-0,25	0,4013	0,20	0,5793	0,65	0,7422
-0,24	0,4052	0,21	0,5832	0,66	0,7454
-0,23	0,4090	0,22	0,5871	0,67	0,7486
-0,22	0,4129	0,23	0,5910	0,68	0,7517
-0,21	0,4168	0,24	0,5948	0,69	0,7549
-0,20	0,4207	0,25	0,5987	0,70	0,7580
-0,19	0,4247	0,26	0,6026	0,71	0,7611
-0,18	0,4286	0,27	0,6064	0,72	0,7642
-0,17	0,4325	0,28	0,6103	0,73	0,7673
-0,16	0,4364	0,29	0,6141	0,74	0,7704
-0,15	0,4404	0,30	0,6179	0,75	0,7734

TABLA II *(continuación)*

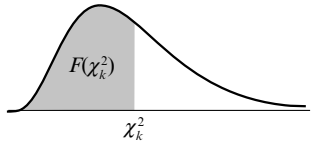
z	$F(z)$	z	$F(z)$	z	$F(z)$
0,76	0,7764	1,21	0,8869	1,66	0,9515
0,77	0,7794	1,22	0,8888	1,67	0,9525
0,78	0,7823	1,23	0,8907	1,68	0,9535
0,79	0,7852	1,24	0,8925	1,69	0,9545
0,80	0,7881	1,25	0,8944	1,70	0,9554
0,81	0,7910	1,26	0,8962	1,71	0,9564
0,82	0,7939	1,27	0,8980	1,72	0,9573
0,83	0,7967	1,28	0,8997	1,73	0,9582
0,84	0,7995	1,29	0,9015	1,74	0,9591
0,85	0,8023	1,30	0,9032	1,75	0,9599
0,86	0,8051	1,31	0,9049	1,76	0,9608
0,87	0,8078	1,32	0,9066	1,77	0,9616
0,88	0,8106	1,33	0,9082	1,78	0,9625
0,89	0,8133	1,34	0,9099	1,79	0,9633
0,90	0,8159	1,35	0,9115	1,80	0,9641
0,91	0,8186	1,36	0,9131	1,81	0,9649
0,92	0,8212	1,37	0,9147	1,82	0,9656
0,93	0,8238	1,38	0,9162	1,83	0,9664
0,94	0,8264	1,39	0,9177	1,84	0,9671
0,95	0,8289	1,40	0,9192	1,85	0,9678
0,96	0,8315	1,41	0,9207	1,86	0,9686
0,97	0,8340	1,42	0,9222	1,87	0,9693
0,98	0,8365	1,43	0,9236	1,88	0,9699
0,99	0,8389	1,44	0,9251	1,89	0,9706
1,00	0,8413	1,45	0,9265	1,90	0,9713
1,01	0,8438	1,46	0,9279	1,91	0,9719
1,02	0,8461	1,47	0,9292	1,92	0,9726
1,03	0,8485	1,48	0,9306	1,93	0,9732
1,04	0,8508	1,49	0,9319	1,94	0,9738
1,05	0,8531	1,50	0,9332	1,95	0,9744
1,06	0,8554	1,51	0,9345	1,96	0,9750
1,07	0,8577	1,52	0,9357	1,97	0,9756
1,08	0,8599	1,53	0,9370	1,98	0,9761
1,09	0,8621	1,54	0,9382	1,99	0,9767
1,10	0,8643	1,55	0,9394	2,00	0,9772
1,11	0,8665	1,56	0,9406	2,01	0,9778
1,12	0,8686	1,57	0,9418	2,02	0,9783
1,13	0,8708	1,58	0,9429	2,03	0,9788
1,14	0,8729	1,59	0,9441	2,04	0,9793
1,15	0,8749	1,60	0,9452	2,05	0,9798
1,16	0,8770	1,61	0,9463	2,06	0,9803
1,17	0,8790	1,62	0,9474	2,07	0,9808
1,18	0,8810	1,63	0,9484	2,08	0,9812
1,19	0,8830	1,64	0,9495	2,09	0,9817
1,20	0,8849	1,65	0,9505	2,10	0,9821

TABLA II *(continuación)*

z	$F(z)$	z	$F(z)$	z	$F(z)$
2,11	0,9826	2,41	0,9920	2,71	0,9966
2,12	0,9830	2,42	0,9922	2,72	0,9967
2,13	0,9834	2,43	0,9925	2,73	0,9968
2,14	0,9838	2,44	0,9927	2,74	0,9969
2,15	0,9842	2,45	0,9929	2,75	0,9970
2,16	0,9846	2,46	0,9931	2,76	0,9971
2,17	0,9850	2,47	0,9932	2,77	0,9972
2,18	0,9854	2,48	0,9934	2,78	0,9973
2,19	0,9857	2,49	0,9936	2,79	0,9974
2,20	0,9861	2,50	0,9938	2,80	0,9974
2,21	0,9864	2,51	0,9940	2,81	0,9975
2,22	0,9868	2,52	0,9941	2,82	0,9976
2,23	0,9871	2,53	0,9943	2,83	0,9977
2,24	0,9875	2,54	0,9945	2,84	0,9977
2,25	0,9878	2,55	0,9946	2,85	0,9978
2,26	0,9881	2,56	0,9948	2,86	0,9979
2,27	0,9884	2,57	0,9949	2,87	0,9979
2,28	0,9887	2,58	0,9951	2,88	0,9980
2,29	0,9890	2,59	0,9952	2,89	0,9981
2,30	0,9893	2,60	0,9953	2,90	0,9981
2,31	0,9896	2,61	0,9955	2,91	0,9982
2,32	0,9898	2,62	0,9956	2,92	0,9982
2,33	0,9901	2,63	0,9957	2,93	0,9983
2,34	0,9904	2,64	0,9959	2,94	0,9984
2,35	0,9906	2,65	0,9960	2,95	0,9984
2,36	0,9909	2,66	0,9961	2,96	0,9985
2,37	0,9911	2,67	0,9962	2,97	0,9985
2,38	0,9913	2,68	0,9963	2,98	0,9986
2,39	0,9916	2,69	0,9964	2,99	0,9986
2,40	0,9918	2,70	0,9965	3,00	0,9987

TABLA III

Función de distribución χ^2 de Pearson, χ^2_k



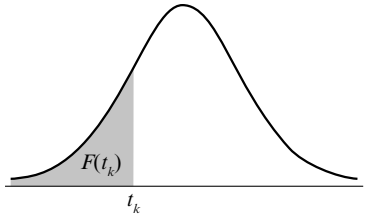
	$F(\chi^2_k)$										
k	0,005	0,010	0,020	0,025	0,050	0,100	0,200	0,250	0,300	0,400	0,500
1	0,04393	0,03157	0,03628	0,03982	0,00393	0,0158	0,0642	0,102	0,148	0,275	0,455
2	0,0100	0,0201	0,0404	0,0506	0,103	0,211	0,446	0,575	0,713	1,022	1,386
3	0,0717	0,115	0,185	0,216	0,352	0,584	1,005	1,213	1,424	1,869	2,366
4	0,207	0,297	0,429	0,484	0,711	1,064	1,649	1,923	2,195	2,753	3,357
5	0,412	0,554	0,752	0,831	1,145	1,610	2,343	2,675	3,000	3,656	4,351
6	0,676	0,872	1,134	1,237	1,635	2,204	3,070	3,455	3,828	4,570	5,348
7	0,989	1,239	1,564	1,690	2,167	2,833	3,822	4,255	4,671	5,493	6,346
8	1,344	1,647	2,032	2,180	2,733	3,490	4,594	5,071	5,527	6,423	7,344
9	1,735	2,088	2,532	2,700	3,325	4,168	5,380	5,899	6,393	7,357	8,343
10	2,156	2,558	3,059	3,247	3,940	4,865	6,179	6,737	7,267	8,295	9,342
11	2,603	3,053	3,609	3,816	4,575	5,578	6,989	7,584	8,148	9,237	10,341
12	3,074	3,571	4,178	4,404	5,226	6,304	7,807	8,438	9,034	10,182	11,340
13	3,565	4,107	4,765	5,009	5,892	7,041	8,634	9,299	9,926	11,129	12,340
14	4,075	4,660	5,368	5,629	6,571	7,790	9,467	10,165	10,821	12,078	13,339
15	4,601	5,229	5,985	6,262	7,261	8,547	10,307	11,037	11,721	13,030	14,339
16	5,142	5,812	6,614	6,908	7,962	9,312	11,152	11,912	12,624	13,983	15,338
17	5,697	6,408	7,255	7,564	8,672	10,085	12,002	12,792	13,531	14,937	16,338
18	6,265	7,015	7,906	8,231	9,390	10,865	12,857	13,675	14,440	15,893	17,338
19	6,844	7,633	8,567	8,907	10,117	11,651	13,716	14,562	15,352	16,850	18,338
20	7,434	8,260	9,237	9,591	10,851	12,443	14,578	15,452	16,266	17,809	19,337
21	8,034	8,897	9,915	10,283	11,591	13,240	15,445	16,344	17,182	18,768	20,337
22	8,643	9,542	10,600	10,982	12,338	14,041	16,314	17,240	18,101	19,729	21,337
23	9,260	10,196	11,293	11,689	13,091	14,848	17,187	18,137	19,021	20,690	22,337
24	9,886	10,856	11,992	12,401	13,848	15,659	18,062	19,037	19,943	21,652	23,337
25	10,520	11,524	12,697	13,120	14,611	16,473	18,940	19,939	20,867	22,616	24,337
26	11,160	12,198	13,409	13,844	15,379	17,292	19,820	20,843	21,792	23,579	25,336
27	11,808	12,878	14,125	14,573	16,151	18,114	20,703	21,749	22,719	24,544	26,336
28	12,461	13,565	14,847	15,308	16,928	18,939	21,588	22,657	23,647	25,509	27,336
29	13,121	14,256	15,574	16,047	17,708	19,768	22,475	23,567	24,577	26,475	28,336
30	13,787	14,953	16,306	16,791	18,493	20,599	23,364	24,478	25,508	27,442	29,336

TABLA III (*continuación)*

	$F(\chi^2_k)$										
k	0,600	0,700	0,750	0,800	0,900	0,950	0,975	0,980	0,990	0,995	0,999
1	0,708	1,074	1,323	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827
2	1,833	2,408	2,773	3,219	4,605	5,991	7,378	7,824	9,210	10,597	13,815
3	2,946	3,665	4,108	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,266
4	4,045	4,878	5,385	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,466
5	5,132	6,064	6,626	7,289	9,236	11,070	12,832	13,388	15,086	16,750	20,515
6	6,211	7,231	7,841	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457
7	7,283	8,383	9,037	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,321
8	8,351	9,524	10,219	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,124
9	9,414	10,656	11,389	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877
10	10,473	11,781	12,549	13,442	15,987	18,307	20,483	21,161	23,209	25,188	29,588
11	11,530	12,899	13,701	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264
12	12,584	14,011	14,845	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909
13	13,636	15,119	15,984	16,985	19,812	22,362	24,736	25,471	27,688	29,819	34,527
14	14,685	16,222	17,117	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,124
15	15,733	17,322	18,245	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,698
16	16,780	18,418	19,369	20,465	23,542	26,296	28,845	29,633	32,000	34,267	39,252
17	17,824	19,511	20,489	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,791
18	18,868	20,601	21,605	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312
19	19,910	21,689	22,718	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,819
20	20,951	22,775	23,828	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,314
21	21,992	23,858	24,935	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,796
22	23,031	24,939	26,039	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268
23	24,069	26,018	27,141	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728
24	25,106	27,096	28,241	29,553	33,196	36,415	39,364	40,270	42,980	45,558	51,179
25	26,143	28,172	29,339	30,675	34,382	37,652	40,646	41,566	44,314	46,928	52,619
26	27,179	29,246	30,435	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,051
27	28,214	30,319	31,528	32,912	36,741	40,113	43,195	44,140	46,963	49,645	55,475
28	29,249	31,391	32,620	34,027	37,916	41,337	44,461	45,419	48,278	50,994	56,892
29	30,283	32,461	33,711	35,139	39,087	42,557	45,722	46,693	49,588	52,335	58,301
30	31,316	33,530	34,800	36,250	40,256	43,773	46,979	47,962	50,892	53,672	59,702

TABLA IV

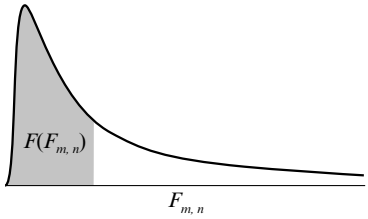
Función de distribución t de Student, t_k



	$F(t_k)$								
k	0,60	0,70	0,75	0,80	0,90	0,95	0,975	0,99	0,995
1	0,325	0,727	1,000	1,376	3,078	6,314	12,71	31,82	63,66
2	0,289	0,617	0,817	1,061	1,886	2,920	4,303	6,965	9,925
3	0,277	0,584	0,765	0,978	1,638	2,353	3,182	4,541	5,841
4	0,271	0,569	0,741	0,941	1,533	2,132	2,776	3,747	4,604
5	0,267	0,559	0,728	0,920	1,476	2,015	2,571	3,365	4,032
6	0,265	0,553	0,718	0,906	1,440	1,943	2,447	3,143	3,707
7	0,263	0,549	0,711	0,896	1,415	1,895	2,365	2,998	3,499
8	0,262	0,546	0,706	0,889	1,397	1,860	2,306	2,896	3,355
9	0,261	0,543	0,703	0,883	1,383	1,833	2,262	2,821	3,250
10	0,260	0,542	0,700	0,879	1,372	1,812	2,228	2,764	3,169
11	0,260	0,540	0,697	0,876	1,363	1,796	2,201	2,718	3,106
12	0,259	0,539	0,696	0,873	1,356	1,782	2,179	2,681	3,055
13	0,259	0,538	0,694	0,870	1,350	1,771	2,160	2,650	3,012
14	0,258	0,537	0,692	0,868	1,345	1,761	2,145	2,624	2,977
15	0,258	0,536	0,691	0,866	1,341	1,753	2,131	2,602	2,947
16	0,258	0,535	0,690	0,865	1,337	1,746	2,120	2,583	2,921
17	0,257	0,534	0,689	0,863	1,333	1,740	2,110	2,567	2,898
18	0,257	0,534	0,688	0,862	1,330	1,734	2,101	2,552	2,878
19	0,257	0,533	0,687	0,861	1,328	1,729	2,093	2,539	2,861
20	0,257	0,533	0,687	0,860	1,325	1,725	2,086	2,528	2,845
21	0,257	0,532	0,686	0,859	1,323	1,721	2,080	2,518	2,831
22	0,256	0,532	0,686	0,858	1,321	1,717	2,074	2,508	2,819
23	0,256	0,532	0,685	0,858	1,319	1,714	2,069	2,500	2,807
24	0,256	0,531	0,685	0,857	1,318	1,711	2,064	2,192	2,797
25	0,256	0,531	0,684	0,856	1,316	1,708	2,060	2,485	2,787
26	0,256	0,531	0,684	0,856	1,315	1,706	2,056	2,479	2,779
27	0,256	0,531	0,684	0,855	1,314	1,703	2,052	2,473	2,771
28	0,256	0,530	0,683	0,855	1,313	1,701	2,048	2,467	2,763
29	0,256	0,530	0,683	0,854	1,311	1,699	2,045	2,462	2,756
30	0,256	0,530	0,683	0,854	1,310	1,697	2,042	2,457	2,750
40	0,255	0,529	0,681	0,851	1,303	1,684	2,021	2,423	2,704
50	0,255	0,528	0,679	0,849	1,298	1,676	2,009	2,403	2,678
60	0,254	0,527	0,679	0,848	1,296	1,671	2,000	2,390	2,660
70	0,254	0,527	0,678	0,846	1,294	1,667	1,994	2,381	2,648
80	0,254	0,527	0,678	0,846	1,292	1,664	1,990	2,374	2,639
90	0,254	0,527	0,677	0,845	1,290	1,662	1,986	2,369	2,632
100	0,254	0,526	0,677	0,845	1,290	1,660	1,984	2,365	2,626
120	0,254	0,526	0,677	0,845	1,289	1,658	1,980	2,358	2,617
200	0,254	0,525	0,676	0,843	1,286	1,653	1,972	2,345	2,601
500	0,253	0,525	0,676	0,842	1,283	1,648	1,965	2,334	2,586
∞	0,253	0,524	0,674	0,842	1,282	1,645	1,960	2,326	2,576

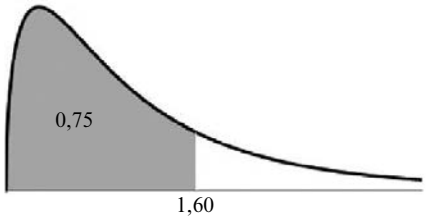
TABLA V

Función de distribución F de Snedecor, $F_{m,\,n}$

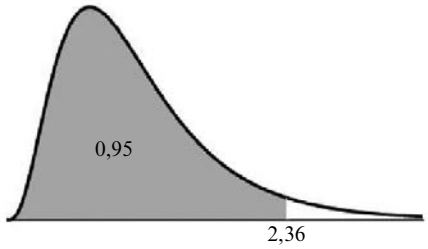


EJEMPLOS: Representación gráfica de cuatro valores, de distintas distribuciones *F,* con su probabilidad acumulada (o función de distribución) sombreada.

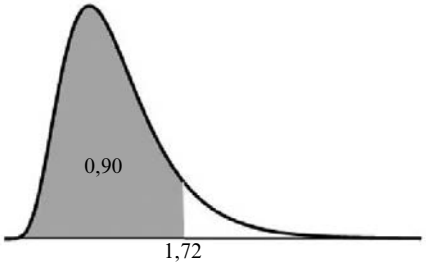
$m = 3$; $n = 10$; $p = 0{,}75$



$m = 8$; $n = 24$; $p = 0{,}95$



$m = 15$; $n = 30$; $p = 0{,}90$



$m = 24$; $n = 120$; $p = 0{,}99$

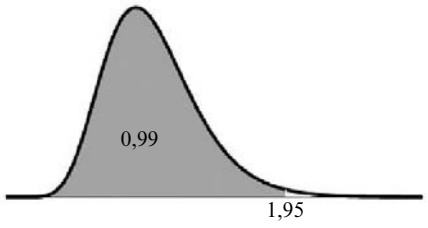


TABLA V (*continuación*)

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
1	0,0005	$0,0^{6}62$	$0,0^{2}50$	0,004	0,009	0,016	0,022	0,027	0,032	0,036	0,039	0,042	0,045	0,0005
	0,0010	$0,0^{5}25$	$0,0^{2}10$	0,006	0,013	0,021	0,028	0,034	0,039	0,044	0,048	0,051	0,054	0,0010
	0,0050	$0,0^{4}62$	$0,0^{2}51$	0,018	0,032	0,044	0,054	0,062	0,068	0,073	0,078	0,082	0,085	0,0050
	0,0100	$0,0^{3}25$	0,010	0,029	0,047	0,062	0,073	0,082	0,089	0,095	0,100	0,104	0,107	0,0100
	0,0250	$0,0^{2}15$	0,026	0,057	0,082	0,100	0,113	0,124	0,132	0,139	0,144	0,149	0,153	0,0250
	0,0500	0,0262	0,054	0,099	0,130	0,151	0,167	0,179	0,188	0,195	0,201	0,206	0,211	0,0500
	0,1000	0,025	0,117	0,181	0,220	0,246	0,265	0,279	0,289	0,298	0,304	0,310	0,315	0,1000
	0,2500	0,172	0,389	0,494	0,553	0,591	0,617	0,636	0,650	0,661	0,670	0,678	0,684	0,2500
	0,5000	1,000	1,500	1,709	1,823	1,894	1,942	1,977	2,004	2,025	2,042	2,056	2,067	0,5000
	0,7500	5,828	7,500	8,200	8,581	8,820	8,983	9,102	9,192	9,263	9,320	9,367	9,406	0,7500
	0,9000	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9	60,2	60,5	60,7	0,9000
	0,9500	161	200	216	225	230	234	237	239	241	242	243	244	0,9500
	0,9750	648	800	864	900	922	937	948	957	963	969	973	977	0,9750
	0,9900	405^{1}	500^{1}	540^{1}	562^{1}	576^{1}	586^{1}	593^{1}	598^{1}	602^{1}	606^{1}	608^{1}	611^{1}	0,9900
	0,9950	162^{2}	200^{2}	216^{2}	225^{2}	231^{2}	234^{2}	237^{2}	239^{2}	241^{2}	242^{2}	243^{2}	244^{2}	0,9950
	0,9990	406^{3}	500^{3}	540^{3}	562^{3}	576^{3}	586^{3}	593^{3}	598^{3}	602^{3}	606^{3}	609^{3}	611^{3}	0,9990
	0,9995	162^{4}	200^{4}	2156^{4}	225^{4}	230^{4}	234^{4}	237^{4}	239^{4}	241^{4}	242^{4}	243^{4}	244^{4}	0,9995
2	0,0005	$0,0^{6}50$	0,001	0,004	0,011	0,020	0,029	0,037	0,044	0,050	0,056	0,061	0,065	0,0005
	0,0010	$0,0^{5}20$	0,001	0,007	0,016	0,027	0,037	0,046	0,054	0,061	0,067	0,072	0,077	0,0010
	0,0050	$0,0^{4}50$	0,005	0,020	0,038	0,055	0,069	0,081	0,091	0,099	0,106	0,112	0,118	0,0050
	0,0100	$0,0^{3}20$	0,010	0,032	0,056	0,075	0,092	0,105	0,116	0,125	0,132	0,139	0,144	0,0100
	0,0250	0,001	0,026	0,062	0,094	0,119	0,138	0,153	0,165	0,175	0,183	0,190	0,196	0,0250
	0,0500	0,005	0,053	0,105	0,144	0,173	0,194	0,211	0,224	0,235	0,244	0,251	0,257	0,0500
	0,1000	0,020	0,111	0,183	0,231	0,265	0,289	0,307	0,321	0,333	0,342	0,350	0,356	0,1000
	0,2500	0,133	0,333	0,439	0,500	0,540	0,567	0,588	0,604	0,616	0,626	0,634	0,641	0,2500
	0,5000	0,667	1,000	1,135	1,207	1,252	1,282	1,305	1,321	1,334	1,345	1,354	1,361	0,5000
	0,7500	2,571	3,000	3,153	3,232	3,280	3,312	3,335	3,353	3,366	3,377	3,386	3,393	0,7500
	0,9000	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,401	9,408	0,9000
	0,9500	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	0,9500
	0,9750	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	0,9750
	0,9900	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	0,9900
	0,9950	198	199	199	199	199	199	199	199	199	199	199	199	0,9950
	0,9990	998	999	999	999	999	999	999	999	999	999	999	999	0,9990
	0,9995	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	200^{1}	0,9995
3	0,0005	$0,0^{6}46$	0,001	0,004	0,012	0,023	0,033	0,043	0,052	0,060	0,067	0,073	0,079	0,0005
	0,0010	$0,0^{5}19$	0,001	0,007	0,018	0,030	0,042	0,053	0,063	0,072	0,080	0,087	0,093	0,0010
	0,0050	$0,0^{4}46$	0,005	0,021	0,041	0,060	0,077	0,092	0,104	0,115	0,124	0,132	0,138	0,0050
	0,0100	$0,0^{3}19$	0,010	0,034	0,060	0,083	0,102	0,118	0,132	0,143	0,153	0,161	0,168	0,0100
	0,0250	0,001	0,026	0,065	0,100	0,129	0,152	0,170	0,185	0,197	0,207	0,216	0,224	0,0250
	0,0500	0,005	0,052	0,108	0,152	0,185	0,210	0,230	0,246	0,259	0,270	0,279	0,287	0,0500
	0,1000	0,019	0,109	0,186	0,239	0,276	0,304	0,325	0,342	0,356	0,367	0,376	0,384	0,1000
	0,2500	0,122	0,317	0,425	0,489	0,531	0,560	0,582	0,599	0,613	0,624	0,633	0,641	0,2500
	0,5000	0,585	0,881	1,000	1,063	1,102	1,129	1,148	1,163	1,174	1,183	1,191	1,197	0,5000
	0,7500	2,024	2,280	2,356	2,390	2,409	2,422	2,430	2,436	2,441	2,445	2,448	2,450	0,7500
	0,9000	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,222	5,216	0,9000
	0,9500	10,1	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,763	8,745	0,9500
	0,9750	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,4	14,3	0,9750
	0,9900	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1	0,9900
	0,9950	55,6	49,8	47,5	46,2	45,4	44,8	44,4	44,1	43,9	43,7	43,5	43,4	0,9950
	0,9990	167	149	141	137	135	133	132	131	130	129	129	128	0,9990
	0,9995	267	237	225	218	214	211	209	208	207	206	204	204	0,9995

TABLA V (*continuación)*

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,051	0,058	0,062	0,066	0,070	0,072	0,074	0,077	0,078	0,080	0,081	0,083	0,0005	1
0,001	0,060	0,067	0,071	0,075	0,079	0,082	0,084	0,087	0,088	0,090	0,091	0,092	0,001	
0,005	0,093	0,101	0,105	0,109	0,113	0,116	0,118	0,121	0,122	0,124	0,126	0,127	0,005	
0,01	0,115	0,124	0,128	0,132	0,137	0,139	0,141	0,145	0,146	0,148	0,150	0,151	0,01	
0,025	0,161	0,170	0,175	0,180	0,184	0,187	0,189	0,193	0,194	0,196	0,198	0,199	0,025	
0,05	0,220	0,230	0,235	0,240	0,245	0,248	0,250	0,254	0,255	0,257	0,259	0,260	0,05	
0,10	0,325	0,336	0,342	0,347	0,353	0,356	0,358	0,363	0,364	0,366	0,368	0,370	0,10	
0,25	0,698	0,712	0,720	0,727	0,734	0,738	0,741	0,747	0,748	0,751	0,754	0,756	0,25	
0,50	2,093	2,119	2,132	2,145	2,158	2,166	2,172	2,182	2,185	2,190	2,195	2,198	0,50	
0,75	9,493	9,581	9,625	9,670	9,714	9,741	9,759	9,795	9,804	9,822	9,838	9,850	0,75	
0,90	61,2	61,7	62,0	62,3	62,5	62,7	62,8	63,0	63,1	63,2	63,3	63,3	0,90	
0,95	246	248	249	250	251	252	252	253	253	254	254	254	0,95	
0,975	985	993	997	100[1]	101[1]	101[1]	101[1]	101[1]	101[1]	102[1]	1021	1021	0,975	
0,99	616[1]	621[1]	623[1]	626[1]	629[1]	630[1]	631[1]	633[1]	634[1]	635[1]	6361	6371	0,99	
0,995	246[2]	248[2]	249[2]	250[2]	251[2]	252[2]	252[2]	253[2]	254[2]	254[2]	2542	2552	0,995	
0,999	616[3]	621[3]	624[3]	626[3]	628[3]	630[3]	631[3]	633[3]	634[3]	635[3]	6363	6373	0,999	
0,9995	246[4]	248[4]	249[4]	250[4]	251[4]	252[4]	253[4]	253[4]	253[4]	2544	2544	2554	0,9995	
0,0005	0,076	0,088	0,094	0,101	0,108	0,113	0,116	0,122	0,123	0,127	0,130	0,132	0,0005	2
0,001	0,088	0,100	0,107	0,114	0,121	0,126	0,129	0,135	0,137	0,140	0,143	0,145	0,001	
0,005	0,130	0,143	0,150	0,157	0,165	0,169	0,173	0,179	0,181	0,184	0,187	0,189	0,005	
0,01	0,157	0,171	0,178	0,186	0,193	0,198	0,201	0,207	0,209	0,212	0,215	0,217	0,01	
0,025	0,210	0,224	0,232	0,239	0,247	0,252	0,255	0,261	0,263	0,266	0,269	0,271	0,025	
0,05	0,272	0,286	0,294	0,302	0,309	0,314	0,317	0,324	0,326	0,329	0,332	0,334	0,05	
0,10	0,371	0,386	0,394	0,402	0,410	0,415	0,418	0,424	0,426	0,429	0,432	0,434	0,10	
0,25	0,657	0,673	0,680	0,689	0,697	0,702	0,705	0,711	0,713	0,716	0,719	0,721	0,25	
0,50	1,377	1,393	1,401	1,410	1,418	1,423	1,426	1,433	1,434	1,438	1,441	1,443	0,50	
0,75	3,410	3,426	3,435	3,443	3,451	3,456	3,459	3,466	3,468	3,471	3,474	3,480	0,75	
0,90	9,425	9,441	9,450	9,458	9,466	9,471	9,475	9,481	9,483	9,486	9,489	9,491	0,90	
0,95	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5	19,5	19,5	19,5	0,95	
0,975	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5	39,5	39,5	39,5	39,5	0,975	
0,99	99,4	99,5	99, 5	99,5	99,5	99,5	99,5	99,5	99,5	99,5	99,5	99,5	0,99	
0,995	199	199	199	199	199	199	199	199	199	199	200	200	0,995	
0,999	999	999	999	999	999	999	999	999	999	999	999	999	0,999	
0,9995	200[1]	200[1]	200[1]	200[1]	200[1]	200[1]	200[1]	200[1]	200[1]	200[1]	200V	200[1]	0,9995	
0,0005	0,093	0,109	0,117	0,127	0,136	0,143	0,147	0,155	0,158	0,162	0,166	0,169	0,0005	3
0,001	0,107	0,123	0,132	0,142	0,152	0,158	0,162	0,171	0,173	0,177	0,182	0,184	0,001	
0,005	0,154	0,172	0,181	0,191	0,201	0,207	0,211	0,220	0,222	0,227	0,231	0,234	0,005	
0,01	0,185	0,203	0,212	0,222	0,232	0,238	0,242	0,251	0,253	0,258	0,262	0,264	0,01	
0,025	0,241	0,259	0,269	0,279	0,289	0,295	0,299	0,308	0,310	0,314	0,318	0,321	0,025	
0,05	0,304	0,323	0,332	0,342	0,352	0,358	0,363	0,371	0,373	0,377	0,381	0,384	0,05	
0,10	0,402	0,420	0,430	0,439	0,449	0,455	0,459	0,467	0,469	0,474	0,477	0,480	0,10	
0,25	0,658	0,675	0,684	0,693	0,702	0,708	0,712	0,719	0,721	0,725	0,728	0,730	0,25	
0,50	1,211	1,225	1,232	1,239	1,246	1,251	1,254	1,259	1,261	1,264	1,266	1,268	0,50	
0,75	2,455	2,460	2,463	2,465	2,467	2,469	2,470	2,471	2,472	2,473	2,474	2,470	0,75	
0,90	5,200	5,184	5,176	5,168	5,160	5,155	5,151	5,144	5,143	5,139	5,136	5,130	0,90	
0,95	8,703	8,660	8,638	8,617	8,594	8,581	8,572	8,554	8,549	8,540	8,532	8,526	0,95	
0,975	14,3	14,2	14,1	14,1	14,0	14,0	14,0	14,0	13,9	13,9	13,9	13,9	0,975	
0,99	26,9	26,7	26,6	26,5	26,4	26, 4	26,3	26,2	26,2	26,2	26,1	26,1	0,99	
0,995	43,1	42,8	42,6	42,5	42,3	42,2	42,2	42,0	42, 0	42,0	41,9	41,9	0,995	
0,999	127	126	126	125	125	125	124	124	124	124	124	123	0,999	
0,9995	203	201	200	200	199	198	198	197	197	197	197	196	0,9995	

TABLA V (*continuación)*

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
4	0,0005	$0,0^{6}44$	0,001	0,005	0,013	0,024	0,036	0,047	0,057	0,066	0,075	0,082	0,089	0,0005
	0,0010	$0,0^{5}18$	0,001	0,007	0,019	0,032	0,046	0,058	0,069	0,080	0,089	0,097	0,104	0,0010
	0,0050	$0,0^{4}44$	0,005	0,022	0,043	0,064	0,083	0,099	0,114	0,126	0,136	0,145	0,153	0,0050
	0,0100	$0,0^{3}18$	0,010	0,035	0,063	0,088	0,109	0,127	0,143	0,156	0,167	0,176	0,185	0,0100
	0,0250	0,001	0,025	0,066	0,104	0,135	0,161	0,181	0,198	0,212	0,224	0,234	0,243	0,0250
	0,0500	0,004	0,052	0,110	0,157	0,193	0,221	0,243	0,261	0,275	0,288	0,298	0,307	0,0500
	0,1000	0,018	0,108	0,187	0,243	0,284	0,314	0,338	0,356	0,371	0,384	0,394	0,403	0,1000
	0,2500	0,117	0,309	0,418	0,484	0,528	0,560	0,583	0,601	0,615	0,627	0,637	0,645	0,2500
	0,5000	0,549	0,828	0,941	1,000	1,037	1,062	1,080	1,093	1,104	1,113	1,120	1,126	0,5000
	0,7500	1,807	2,000	2,047	2,064	2,072	2,077	2,079	2,080	2,081	2,082	2,082	2,083	0,7500
	0,9000	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,907	3,896	0,9000
	0,9500	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	0,9500
	0,9750	12,2	10,6	9,979	9,604	9,364	9,197	9,074	8,980	8,905	8,844	8,794	8,751	0,9750
	0,9900	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14, 4	0,9900
	0,9950	31,3	26,3	24,3	23,2	22,5	22,0	21,6	21,4	21,1	21,0	20,8	20,7	0,9950
	0,9990	74,1	61,2	56,2	53,4	51,7	50,5	49,7	49,0	48,5	48,0	47,7	47,4	0,9990
	0,9995	106	87,4	80,1	76,1	73,6	71,9	70,6	69,7	68,9	68,3	67,8	67,4	0,9995
5	0,0005	$0,0^{6}43$	0,001	0,005	0,014	0,025	0,038	0,050	0,061	0,071	0,080	0,089	0,097	0,0005
	0,0010	$0,0^{5}17$	0,001	0,007	0,019	0,034	0,048	0,062	0,074	0,085	0,095	0,104	0,112	0,0010
	0,0050	$0,0^{4}43$	0,005	0,022	0,045	0,067	0,087	0,105	0,120	0,134	0,146	0,156	0,165	0,0050
	0,0100	$0,0^{3}17$	0,010	0,035	0,064	0,091	0,114	0,134	0,151	0,165	0,177	0,188	0,197	0,0100
	0,0250	0,001	0,025	0,067	0,107	0,140	0,167	0,189	0,208	0,223	0,236	0,247	0,257	0,0250
	0,0500	0,004	0,052	0,111	0,160	0,198	0,228	0,252	0,271	0,287	0,301	0,312	0,322	0,0500
	0,1000	0,017	0,108	0,188	0,247	0,290	0,322	0,347	0,367	0,383	0,397	0,408	0,418	0,1000
	0,2500	0,113	0,305	0,415	0,483	0,528	0,560	0,584	0,603	0,618	0,631	0,641	0,650	0,2500
	0,5000	0,528	0,799	0,907	0,965	1,000	1,024	1,041	1,055	1,065	1,073	1,080	1,085	0,5000
	0,7500	1,692	1,853	1,884	1,893	1,895	1,894	1,894	1,892	1,891	1,890	1,889	1,888	0,7500
	0,9000	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,282	3,268	0,9000
	0,9500	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	0,9500
	0,9750	10,0	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619	6,568	6,525	0,9750
	0,9900	16,3	13,3	12,1	11,4	11,0	10, 7	10,5	10,3	10,2	10, 1	9,963	9,888	0,9900
	0,9950	22,8	18,3	16,5	15, 6	14,9	14,5	14,2	14,0	13,8	13,6	13,5	13,4	0,9950
	0,9990	47,2	37,1	33,2	31,1	29,7	28,8	28,2	27,6	27,2	26,9	26,6	26,4	0,9990
	0,9995	63,6	49,8	44,4	41,5	39,7	38, 5	37,6	36,9	36,4	35,9	35,6	35,2	0,9995
6	0,0005	$0,0^{6}43$	0,001	0,005	0,014	0,026	0,039	0,052	0,064	0,075	0,085	0,094	0,103	0,0005
	0,0010	$0,0^{5}17$	0,001	0,008	0,020	0,035	0,050	0,064	0,078	0,090	0,101	0,111	0,119	0,0010
	0,0050	$0,0^{4}43$	0,005	0,022	0,046	0,069	0,090	0,109	0,126	0,140	0,153	0,164	0,174	0,0050
	0,0100	$0,0^{3}17$	0,010	0,036	0,066	0,094	0,118	0,139	0,157	0,172	0,186	0,197	0,207	0,0100
	0,0250	0,001	0,025	0,068	0,109	0,143	0,172	0,195	0,215	0,231	0,246	0,258	0,268	0,0250
	0,0500	0,004	0,052	0,112	0,162	0,202	0,233	0,259	0,279	0,296	0,311	0,323	0,334	0,0500
	0,1000	0,017	0,107	0,189	0,249	0,294	0,327	0,354	0,375	0,392	0,406	0,419	0,429	0,1000
	0,2500	0,111	0,302	0,413	0,482	0,528	0,561	0,586	0,606	0,621	0,634	0,645	0,654	0,2500
	0,5000	0,515	0,780	0,886	0,942	0,977	1,000	1,017	1,030	1,040	1,048	1,054	1,060	0,5000
	0,7500	1,621	1,762	1,784	1,787	1,785	1,782	1,779	1,776	1,773	1,771	1,769	1,767	0,7500
	0,9000	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,920	2,905	0,9000
	0,9500	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	0,9500
	0,9750	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461	5,410	5,366	0,9750
	0,9900	13,7	10,9	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,790	7,718	0,9900
	0,9950	18,6	14,5	12,9	12,0	11,5	11,1	10,8	10,6	10,4	10,2	10,1	10,0	0,9950
	0,9990	35,5	27,0	23,7	21,9	20,8	20,0	19,5	19,0	18,7	18,4	18,2	18,0	0,9990
	0,9995	46, 1	34,8	30,5	28,1	26,6	25,6	24, 9	24,3	23,9	23,5	23,2	23,0	0,9995

TABLA V (*continuación*)

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,106	0,125	0,135	0,147	0,159	0,166	0,172	0,183	0,185	0,191	0,196	0,200	0,0005	4
0,0010	0,121	0,141	0,152	0,163	0,175	0,183	0,188	0,199	0,202	0,208	0,213	0,217	0,0010	
0,0050	0,172	0,193	0,205	0,216	0,229	0,236	0,242	0,252	0,255	0,261	0,266	0,269	0,0050	
0,0100	0,204	0,226	0,237	0,249	0,261	0,269	0,274	0,285	0,287	0,293	0,298	0,301	0,0100	
0,0250	0,263	0,285	0,296	0,308	0,320	0,327	0,332	0,343	0,346	0,351	0,356	0,359	0,0250	
0,0500	0,327	0,349	0,360	0,372	0,384	0,391	0,396	0,406	0,409	0,414	0,418	0,422	0,0500	
0,1000	0,423	0,445	0,456	0,467	0,478	0,485	0,490	0,500	0,502	0,507	0,511	0,514	0,1000	
0,2500	0,664	0,683	0,692	0,702	0,712	0,718	0,722	0,730	0,732	0,737	0,740	0,743	0,2500	
0,5000	1,139	1,152	1,158	1,165	1,172	1,176	1,178	1,184	1,185	1,188	1,190	1,192	0,5000	
0,7500	2,083	2,083	2,083	2,082	2,082	2,082	2,082	2,081	2,081	2,081	2,081	2,080	0,7500	
0,9000	3,870	3,844	3,831	3,817	3,804	3,795	3,790	3,778	3,775	3,769	3,764	3,760	0,9000	
0,9500	5,858	5,803	5,774	5,746	5,717	5,699	5,688	5,664	5,658	5,646	5,635	5,628	0,9500	
0,9750	8,657	8,560	8,511	8,461	8,411	8,381	8,360	8,319	8,309	8,288	8,270	8,257	0,9750	
0,9900	14,2	14,0	13,9	13,8	13,7	13,7	13,6	13,6	13,6	13,5	13,5	13,5	0,9900	
0,9950	20,4	20,2	20,0	19,9	19,8	19, 7	19,6	19,5	19,5	19,4	19,4	19,3	0,9950	
0,9990	46,8	46,1	45,8	45,4	45,1	44,9	44,7	44,5	44,4	44,3	44,1	44,0	0,9990	
0,9995	66,5	65,5	65,1	64,6	64,1	63,8	63,6	63,2	63,1	62,9	62,7	62,6	0,9995	
0,0005	0,115	0,137	0,150	0,163	0,177	0,186	0,192	0,205	0,209	0,216	0,222	0,226	0,0005	5
0,0010	0,132	0,155	0,167	0,181	0,195	0,204	0,210	0,223	0,226	0,233	0,239	0,244	0,0010	
0,0050	0,186	0,210	0,223	0,237	0,251	0,260	0,266	0,279	0,282	0,288	0,294	0,299	0,0050	
0,0100	0,220	0,244	0,257	0,270	0,285	0,293	0,300	0,312	0,315	0,322	0,327	0,331	0,0100	
0,0250	0,280	0,304	0,317	0,330	0,344	0,353	0,359	0,371	0,374	0,380	0,386	0,390	0,0250	
0,0500	0,345	0,369	0,382	0,395	0,408	0,417	0,422	0,434	0,437	0,443	0,448	0,452	0,0500	
0,1000	0,440	0,463	0,476	0,488	0,501	0,509	0,514	0,525	0,527	0,533	0,538	0,541	0,1000	
0,2500	0,669	0,690	0,700	0,711	0,721	0,728	0,732	0,741	0,743	0,748	0,752	0,755	0,2500	
0,5000	1,098	1,111	1,117	1,123	1,130	1,134	1,136	1,141	1,143	1,145	1,147	1,150	0,5000	
0,7500	1,885	1,882	1,880	1,878	1,876	1,875	1,874	1,872	1,872	1,871	1,870	1,870	0,7500	
0,9000	3,238	3,207	3,191	3,174	3,157	3,147	3,140	3,126	3,123	3,116	3,109	3,100	0,9000	
0,9500	4,619	4,558	4,527	4,496	4,464	4,444	4,431	4,405	4,398	4,385	4,373	4,365	0,9500	
0,9750	6,428	6,329	6,278	6,227	6,175	6,144	6,123	6,080	6,069	6,048	6,028	6,015	0,9750	
0,9900	9,722	9,553	9,466	9,379	9,291	9,238	9,202	9,130	9,112	9,075	9,042	9,020	0,9900	
0,9950	13,1	12,9	12,8	12,7	12,5	12,5	12,4	12,3	12,3	12,2	12,2	12,1	0,9950	
0,9990	25,9	25,4	25,1	24,9	24,6	24,4	24,3	24,1	24,1	24,0	23,9	23,8	0,9990	
0,9995	34,5	33,8	33,4	33, 1	32,7	32,5	32,4	32,1	32,0	31,8	31,7	31,6	0,9995	
0,0005	0,123	0,148	0,162	0,177	0,193	0,203	0,210	0,225	0,229	0,237	0,244	0,249	0,0005	6
0,0010	0,141	0,166	0,180	0,195	0,211	0,222	0,229	0,243	0,247	0,255	0,262	0,267	0,0010	
0,0050	0,197	0,224	0,238	0,253	0,269	0,279	0,286	0,301	0,304	0,312	0,319	0,323	0,0050	
0,0100	0,232	0,258	0,273	0,288	0,304	0,314	0,321	0,335	0,338	0,346	0,352	0,357	0,0100	
0,0250	0,293	0,320	0,334	0,349	0,364	0,374	0,381	0,394	0,398	0,405	0,411	0,415	0,0250	
0,0500	0,358	0,385	0,399	0,413	0,428	0,437	0,444	0,456	0,460	0,466	0,472	0,477	0,0500	
0,1000	0,453	0,478	0,491	0,505	0,519	0,528	0,533	0,545	0,548	0,554	0,560	0,564	0,1000	
0,2500	0,675	0,696	0,707	0,718	0,730	0,737	0,741	0,751	0,753	0,758	0,762	0,765	0,2500	
0,5000	1,072	1,084	1,091	1,097	1,103	1,107	1,109	1,114	1,116	1,118	1,120	1,122	0,5000	
0,7500	1,762	1,757	1,754	1,751	1,748	1,746	1,744	1,741	1,741	1,739	1,738	1,740	0,7500	
0,9000	2,871	2,836	2,818	2,800	2,781	2,770	2,762	2,746	2,742	2,734	2,727	2,720	0,9000	
0,9500	3,938	3,874	3,841	3,808	3,774	3,754	3,740	3,712	3,705	3,690	3,678	3,669	0,9500	
0,9750	5,269	5,168	5,117	5,065	5,012	4,980	4,959	4,915	4,904	4,882	4,862	4,849	0,9750	
0,9900	7,559	7,396	7,313	7,229	7,143	7,091	7,057	6,987	6,969	6,934	6,901	6,880	0,9900	
0,9950	9,814	9,589	9,474	9,358	9,241	9,170	9,122	9,026	9,001	8,953	8,909	8,879	0,9950	
0,9990	17,6	17,1	16,9	16,7	16,4	16,3	16,2	16,0	16,0	15,9	15,8	15,7	0,9990	
0,9995	22,4	21,8	21,5	21,2	20,9	20,8	20,6	20,4	20,3	20,2	20,1	20,0	0,9995	

TABLA V (*continuación*)

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
7	0,0005	0,0642	0,001	0,005	0,014	0,027	0,040	0,054	0,066	0,078	0,089	0,099	0,108	0,0005
	0,0010	0,0517	0,001	0,008	0,020	0,036	0,051	0,067	0,081	0,093	0,105	0,116	0,125	0,0010
	0,0050	0,0442	0,005	0,023	0,046	0,070	0,093	0,113	0,130	0,145	0,159	0,171	0,181	0,0050
	0,0100	0,0317	0,010	0,036	0,067	0,096	0,121	0,143	0,162	0,178	0,192	0,205	0,216	0,0100
	0,0250	0,001	0,025	0,068	0,110	0,146	0,176	0,200	0,221	0,238	0,253	0,266	0,277	0,0250
	0,0500	0,004	0,052	0,113	0,164	0,205	0,238	0,264	0,286	0,304	0,319	0,332	0,343	0,0500
	0,1000	0,017	0,107	0,190	0,251	0,297	0,332	0,359	0,381	0,399	0,414	0,427	0,438	0,1000
	0,2500	0,110	0,300	0,411	0,481	0,528	0,562	0,588	0,608	0,624	0,637	0,649	0,658	0,2500
	0,5000	0,506	0,767	0,871	0,926	0,960	0,983	1,000	1,013	1,022	1,030	1,037	1,042	0,5000
	0,7500	1,573	1,701	1,717	1,716	1,711	1,706	1,701	1,697	1,693	1,690	1,687	1,684	0,7500
	0,9000	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,684	2,668	0,9000
	0,9500	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	0,9500
	0,9750	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761	4,709	4,666	0,9750
	0,9900	12,2	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,538	6,469	0,9900
	0,9950	16,2	12,4	10,9	10,1	9,522	9,155	8,885	8,678	8,514	8,380	8,270	8,176	0,9950
	0,9990	29,2	21,7	18,8	17,2	16,2	15,5	15,0	14,6	14,3	14,1	13, 9	13,7	0,9990
	0,9995	37,0	27,2	23,5	21,4	20,2	19,3	18,7	18,2	17,8	17,5	17,2	17,0	0,9995
8	0,0005	0,0642	0,001	0,005	0,014	0,027	0,041	0,055	0,068	0,081	0,092	0,102	0,112	0,0005
	0,0010	0,0517	0,001	0,008	0,020	0,036	0,053	0,068	0,083	0,096	0,109	0,120	0,130	0,0010
	0,0050	0,0442	0,005	0,023	0,047	0,072	0,095	0,115	0,133	0,149	0,164	0,176	0,187	0,0050
	0,0100	0,0317	0,010	0,036	0,068	0,097	0,123	0,146	0,166	0,183	0,198	0,211	0,222	0,0100
	0,0250	0,001	0,025	0,069	0,111	0,148	0,179	0,204	0,226	0,244	0,259	0,273	0,285	0,0250
	0,0500	0,004	0,052	0,113	0,166	0,208	0,241	0,268	0,291	0,310	0,326	0,339	0,351	0,0500
	0,1000	0,017	0,107	0,190	0,253	0,299	0,335	0,363	0,386	0,405	0,421	0,434	0,446	0,1000
	0,2500	0,109	0,298	0,410	0,481	0,528	0,563	0,589	0,610	0,627	0,640	0,652	0,661	0,2500
	0,5000	0,499	0,757	0,860	0,915	0,948	0,971	0,988	1,000	1,010	1,018	1,024	1,029	0,5000
	0,7500	1,538	1,657	1,668	1,664	1,658	1,651	1,645	1,640	1,635	1,631	1,627	1,624	0,7500
	0,9000	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,519	2,502	0,9000
	0,9500	5,318	4,459	4,066	3,838	3,688	3,581	3,500	3,438	3,388	3,347	3,313	3,284	0,9500
	0,9750	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295	4,243	4,200	0,9750
	0,9900	11,3	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,734	5,667	0,9900
	0,9950	14,7	11,0	9,597	8,805	8,302	7,952	7,694	7,496	7,339	7,211	7,105	7,015	0,9950
	0,9990	25,4	18,5	15,8	14,4	13,5	12,9	12,4	12,0	11,8	11,5	11,4	11,2	0,9990
	0,9995	31,6	22,8	19,4	17,6	16,4	15,7	15,1	14,6	14,3	14,0	13,8	13,6	0,9995
9	0,0005	0,000	0,001	0,005	0,015	0,028	0,042	0,056	0,070	0,083	0,095	0,106	0,115	0,0005
	0,0010	0,000	0,001	0,008	0,021	0,037	0,054	0,070	0,085	0,099	0,112	0,123	0,134	0,0010
	0,0050	0,000	0,005	0,023	0,047	0,073	0,096	0,117	0,136	0,153	0,168	0,181	0,192	0,0050
	0,0100	0,000	0,010	0,037	0,068	0,098	0,125	0,149	0,169	0,187	0,202	0,216	0,228	0,0100
	0,0250	0,001	0,025	0,069	0,112	0,150	0,181	0,207	0,230	0,248	0,265	0,279	0,291	0,0250
	0,0500	0,004	0,052	0,113	0,167	0,210	0,244	0,272	0,295	0,315	0,331	0,345	0,358	0,0500
	0,1000	0,017	0,107	0,191	0,254	0,302	0,338	0,367	0,390	0,410	0,426	0,440	0,452	0,1000
	0,2500	0,108	0,297	0,410	0,480	0,529	0,564	0,591	0,612	0,629	0,643	0,654	0,664	0,2500
	0,5000	0,494	0,749	0,852	0,906	0,939	0,962	0,978	0,990	1,000	1,008	1,014	1,019	0,5000
	0,7500	1,512	1,624	1,632	1,625	1,617	1,609	1,602	1,596	1,591	1,586	1,582	1,579	0,7500
	0,9000	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,396	2,379	0,9000
	0,9500	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	0,9500
	0,9750	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964	3,912	3,868	0,9750
	0,9900	10,6	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,178	5,111	0,9900
	0,9950	13,6	10,1	8,717	7,956	7,471	7,134	6,885	6,693	6,541	6,417	6,314	6,227	0,9950
	0,9990	22,9	16,4	13,9	12,6	11,7	11,1	10,7	10,4	10,1	9,894	9,719	9,570	0,9990
	0,9995	28,0	19,9	16,8	15,1	14,1	13,3	12,8	12,4	12,1	11,8	11,6	11,4	0,9995

TABLA V (*continuación*)

	m													
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,130	0,157	0,172	0,188	0,206	0,218	0,225	0,242	0,246	0,255	0,263	0,269	0,0005	7
0,0010	0,148	0,176	0,191	0,208	0,225	0,237	0,245	0,261	0,265	0,274	0,282	0,288	0,0010	
0,0050	0,206	0,235	0,251	0,267	0,285	0,296	0,304	0,320	0,324	0,332	0,340	0,345	0,0050	
0,0100	0,241	0,270	0,286	0,303	0,320	0,331	0,339	0,354	0,358	0,366	0,374	0,379	0,0100	
0,0250	0,304	0,333	0,348	0,364	0,381	0,392	0,399	0,414	0,418	0,425	0,432	0,437	0,0250	
0,0500	0,369	0,398	0,413	0,428	0,445	0,455	0,462	0,476	0,479	0,486	0,493	0,498	0,0500	
0,1000	0,463	0,490	0,504	0,519	0,534	0,543	0,550	0,563	0,566	0,572	0,578	0,582	0,1000	
0,2500	0,679	0,702	0,713	0,725	0,737	0,744	0,749	0,759	0,762	0,767	0,771	0,775	0,2500	
0,5000	1,054	1,066	1,072	1,079	1,085	1,088	1,091	1,096	1,097	1,099	1,102	1,103	0,5000	
0,7500	1,678	1,671	1,667	1,663	1,659	1,657	1,655	1,651	1,650	1,648	1,646	1,645	0,7500	
0,9000	2,632	2,595	2,575	2,555	2,535	2,523	2,514	2,497	2,493	2,484	2,476	2,470	0,9000	
0,9500	3,511	3,445	3,410	3,376	3,340	3,319	3,304	3,275	3,267	3,252	3,239	3,230	0,9500	
0,9750	4,568	4,467	4,415	4,362	4,309	4,276	4,254	4,210	4,199	4,176	4,156	4,142	0,9750	
0,9900	6,314	6,155	6,074	5,992	5,908	5,858	5,824	5,755	5,737	5,702	5,671	5,650	0,9900	
0,9950	7,968	7,754	7,645	7,534	7,422	7,354	7,309	7,217	7,193	7,147	7,104	7,076	0,9950	
0,9990	13,324	12,931	12,733	12,529	12,325	12,202	12,118	11,951	11,909	11,823	11,747	11,696	0,9990	
0,9995	16,502	16,003	15,752	15,494	15,236	15,076	14,974	14,759	14,705	14,599	14,501	14,435	0,9995	
0,0005	0,136	0,164	0,181	0,198	0,218	0,230	0,239	0,257	0,262	0,272	0,281	0,287	0,0005	8
0,0010	0,155	0,184	0,200	0,218	0,238	0,250	0,259	0,277	0,282	0,291	0,300	0,306	0,0010	
0,0050	0,214	0,244	0,261	0,279	0,299	0,311	0,319	0,336	0,341	0,350	0,359	0,364	0,0050	
0,0100	0,250	0,281	0,297	0,315	0,334	0,346	0,354	0,371	0,376	0,384	0,393	0,398	0,0100	
0,0250	0,313	0,343	0,360	0,377	0,395	0,407	0,415	0,431	0,435	0,443	0,451	0,456	0,0250	
0,0500	0,379	0,409	0,425	0,441	0,459	0,469	0,477	0,492	0,496	0,504	0,511	0,516	0,0500	
0,1000	0,472	0,500	0,515	0,531	0,547	0,557	0,563	0,577	0,581	0,588	0,594	0,599	0,1000	
0,2500	0,683	0,707	0,718	0,731	0,743	0,751	0,756	0,767	0,769	0,775	0,780	0,783	0,2500	
0,5000	1,041	1,053	1,059	1,065	1,071	1,075	1,077	1,082	1,083	1,086	1,088	1,089	0,5000	
0,7500	1,617	1,609	1,604	1,600	1,595	1,591	1,589	1,585	1,584	1,581	1,579	1,578	0,7500	
0,9000	2,464	2,425	2,404	2,383	2,361	2,348	2,339	2,321	2,316	2,307	2,298	2,293	0,9000	
0,9500	3,218	3,150	3,115	3,079	3,043	3,020	3,005	2,975	2,967	2,951	2,937	2,930	0,9500	
0,9750	4,101	3,999	3,947	3,894	3,840	3,807	3,784	3,739	3,728	3,705	3,684	3,670	0,9750	
0,9900	5,515	5,359	5,279	5,198	5,116	5,065	5,032	4,963	4,946	4,911	4,880	4,859	0,9900	
0,9950	6,814	6,608	6,503	6,396	6,288	6,222	6,177	6,087	6,065	6,019	5,978	5,951	0,9950	
0,9990	10,8	10,5	10,3	10,1	9,919	9,804	9,728	9,572	9,532	9,453	9,382	9,333	0,9990	
0,9995	13,1	12,7	12,5	12,2	12,0	11,8	11,8	11,6	11,5	11,4	11,3	11,3	0,9995	
0,0005	0,141	0,171	0,188	0,207	0,228	0,242	0,251	0,271	0,276	0,287	0,297	0,303	0,0005	9
0,0010	0,160	0,191	0,208	0,228	0,248	0,262	0,271	0,291	0,296	0,306	0,316	0,323	0,0010	
0,0050	0,220	0,253	0,271	0,290	0,310	0,323	0,332	0,351	0,356	0,366	0,375	0,382	0,0050	
0,0100	0,257	0,289	0,307	0,326	0,346	0,359	0,368	0,386	0,391	0,400	0,409	0,415	0,0100	
0,0250	0,320	0,353	0,370	0,388	0,408	0,420	0,428	0,446	0,450	0,459	0,467	0,473	0,0250	
0,0500	0,386	0,418	0,435	0,452	0,471	0,482	0,490	0,506	0,511	0,519	0,527	0,532	0,0500	
0,1000	0,479	0,509	0,525	0,541	0,558	0,568	0,575	0,590	0,594	0,601	0,608	0,613	0,1000	
0,2500	0,687	0,711	0,723	0,736	0,749	0,757	0,762	0,773	0,776	0,782	0,787	0,791	0,2500	
0,5000	1,031	1,043	1,049	1,055	1,061	1,064	1,067	1,072	1,073	1,075	1,077	1,080	0,5000	
0,7500	1,570	1,561	1,556	1,551	1,545	1,541	1,539	1,534	1,533	1,530	1,527	1,530	0,7500	
0,9000	2,340	2,298	2,277	2,255	2,232	2,218	2,208	2,189	2,184	2,174	2,165	2,160	0,9000	
0,9500	3,006	2,936	2,900	2,864	2,826	2,803	2,787	2,756	2,748	2,731	2,717	2,707	0,9500	
0,9750	3,769	3,667	3,614	3,560	3,505	3,472	3,449	3,403	3,392	3,368	3,347	3,333	0,9750	
0,9900	4,962	4,808	4,729	4,649	4,567	4,517	4,483	4,415	4,398	4,363	4,332	4,311	0,9900	
0,9950	6,032	5,832	5,729	5,625	5,519	5,454	5,410	5,322	5,300	5,255	5,215	5,188	0,9950	
0,9990	9,239	8,898	8,724	8,547	8,368	8,260	8,186	8,038	8,002	7,926	7,858	7,813	0,9990	
0,9995	11,0	10,6	10,4	10,2	9,943	9,808	9,719	9,539	9,491	9,401	9,317	9,262	0,9995	

TABLA V (*continuación)*

		m												
n	*p*	1	2	3	4	5	6	, 7	8	9	10	11	12	*p*
10	0,0005	0,0641	0,001	0,005	0,015	0,028	0,043	0,057	0,071	0,085	0,097	0,108	0,119	0,0005
	0,0010	0,0517	0,001	0,008	0,021	0,037	0,054	0,071	0,087	0,101	0,114	0,126	0,137	0,0010
	0,0050	0,0441	0,005	0,023	0,048	0,073	0,098	0,119	0,139	0,156	0,171	0,185	0,197	0,0050
	0,0100	0,0317	0,010	0,037	0,069	0,099	0,127	0,151	0,172	0,190	0,206	0,220	0,233	0,0100
	0,0250	0,001	0,025	0,069	0,113	0,151	0,183	0,210	0,233	0,252	0,269	0,284	0,296	0,0250
	0,0500	0,004	0,052	0,114	0,168	0,211	0,246	0,275	0,299	0,319	0,336	0,350	0,363	0,0500
	0,1000	0,017	0,106	0,191	0,255	0,303	0,340	0,370	0,394	0,414	0,431	0,445	0,457	0,1000
	0,2500	0,107	0,296	0,409	0,480	0,529	0,565	0,592	0,613	0,630	0,645	0,657	0,667	0,2500
	0,5000	0,490	0,743	0,845	0,899	0,932	0,954	0,971	0,983	0,992	1,000	1,006	1,012	0,5000
	0,7500	1,491	1,598	1,603	1,595	1,585	1,576	1,569	1,562	1,556	1,551	1,547	1,543	0,7500
	0,9000	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,302	2,284	0,9000
	0,9500	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	0,9500
	0,9750	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717	3,665	3,621	0,9750
	0,9900	10,0	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,772	4,706	0,9900
	0,9950	12,8	9,427	8,081	7,343	6,872	6,545	6,303	6,116	5,968	5,847	5,746	5,661	0,9950
	0,9990	21,0	14,9	12,6	11,3	10,5	9,926	9,517	9,204	8,956	8,754	8,587	8,446	0,9990
	0,9995	25,5	17,9	15,0	13,4	12,4	11,7	11,2	10,9	10,6	10,3	10,1	9,943	0,9995
11	0,0005	0,0641	0,001	0,005	0,015	0,028	0,043	0,058	0,073	0,086	0,099	0,111	0,121	0,0005
	0,0010	0,0516	0,001	0,008	0,021	0,038	0,055	0,072	0,088	0,103	0,116	0,129	0,140	0,0010
	0,0050	0,0440	0,005	0,023	0,048	0,074	0,099	0,121	0,141	0,158	0,174	0,188	0,200	0,0050
	0,0100	0,0316	0,010	0,037	0,069	0,100	0,128	0,153	0,174	0,193	0,210	0,224	0,237	0,0100
	0,0250	0,001	0,025	0,070	0,114	0,152	0,185	0,212	0,236	0,256	0,273	0,288	0,301	0,0250
	0,0500	0,004	0,052	0,114	0,168	0,213	0,248	0,278	0,302	0,322	0,340	0,355	0,368	0,0500
	0,1000	0,017	0,106	0,191	0,256	0,305	0,343	0,373	0,397	0,417	0,434	0,449	0,462	0,1000
	0,2500	0,107	0,295	0,409	0,480	0,529	0,565	0,593	0,614	0,632	0,646	0,659	0,669	0,2500
	0,5000	0,486	0,739	0,840	0,893	0,926	0,948	0,964	0,977	0,986	0,994	1,000	1,005	0,5000
	0,7500	1,475	1,577	1,580	1,570	1,560	1,550	1,542	1,535	1,528	1,523	1,518	1,514	0,7500
	0,9000	3,225	2,860	2,660	2,536	2,451	2,389	2,342	2,304	2,274	2,248	2,227	2,209	0,9000
	0,9500	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	0,9500
	0,9750	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588	3,526	3,474	3,430	0,9750
	0,9900	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,462	4,397	0,9900
	0,9950	12,2	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537	5,418	5,320	5,236	0,9950
	0,9990	19,7	13,8	11,6	10,3	9,579	9,047	8,655	8,355	8,116	7,923	7,762	7,625	0,9990
	0,9995	23,7	16,4	13,7	12,2	11,2	10,6	10,1	9,764	9,477	9,242	9,048	8,884	0,9995
12	0,0005	0,0641	0,001	0,005	0,015	0,028	0,044	0,059	0,074	0,088	0,101	0,113	0,124	0,0005
	0,0010	0,0516	0,001	0,008	0,021	0,038	0,056	0,073	0,089	0,104	0,118	0,131	0,143	0,0010
	0,0050	0,0439	0,005	0,023	0,048	0,075	0,100	0,122	0,143	0,161	0,177	0,191	0,204	0,0050
	0,0100	0,0316	0,010	0,037	0,070	0,101	0,130	0,155	0,176	0,196	0,213	0,227	0,241	0,0100
	0,0250	0,001	0,025	0,070	0,114	0,153	0,186	0,214	0,238	0,259	0,276	0,292	0,305	0,0250
	0,0500	0,004	0,052	0,114	0,169	0,214	0,250	0,280	0,305	0,325	0,343	0,359	0,372	0,0500
	0,1000	0,016	0,106	0,192	0,257	0,306	0,344	0,375	0,400	0,420	0,438	0,453	0,466	0,1000
	0,2500	0,106	0,295	0,408	0,480	0,530	0,566	0,594	0,616	0,633	0,648	0,660	0,671	0,2500
	0,5000	0,484	0,735	0,835	0,888	0,921	0,943	0,959	0,972	0,981	0,989	0,995	1,000	0,5000
	0,7500	1,461	1,560	1,561	1,550	1,539	1,529	1,520	1,512	1,505	1,500	1,495	1,490	0,7500
	0,9000	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,166	2,147	0,9000
	0,9500	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	0,9500
	0,9750	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436	3,374	3,321	3,277	0,9750
	0,9900	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,220	4,155	0,9900
	0,9950	11,8	8,510	7,226	6,521	6,071	5,757	5,524	5,345	5,202	5,085	4,988	4,906	0,9950
	0,9990	18,6	13,0	10,8	9,633	8,892	8,378	8,001	7,711	7,480	7,292	7,136	7,005	0,9990
	0,9995	22,2	15,3	12,7	11,2	10,4	9,739	9,284	8,935	8,658	8,435	8,247	8,091	0,9995

TABLA V (*continuación*)

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,145	0,177	0,195	0,215	0,237	0,252	0,262	0,283	0,289	0,300	0,311	0,318	0,0005	10
0,0010	0,164	0,197	0,216	0,236	0,258	0,272	0,282	0,303	0,309	0,320	0,331	0,338	0,0010	
0,0050	0,226	0,260	0,279	0,299	0,321	0,335	0,344	0,364	0,370	0,380	0,390	0,397	0,0050	
0,0100	0,263	0,297	0,316	0,336	0,357	0,371	0,380	0,399	0,405	0,415	0,424	0,431	0,0100	
0,0250	0,327	0,361	0,379	0,398	0,419	0,432	0,440	0,459	0,464	0,473	0,482	0,488	0,0250	
0,0500	0,393	0,426	0,444	0,462	0,481	0,494	0,502	0,519	0,523	0,532	0,541	0,546	0,0500	
0,1000	0,486	0,516	0,533	0,550	0,567	0,578	0,586	0,601	0,605	0,613	0,621	0,626	0,1000	
0,2500	0,690	0,715	0,727	0,740	0,754	0,762	0,768	0,779	0,782	0,788	0,793	0,797	0,2500	
0,5000	1,023	1,035	1,041	1,047	1,053	1,056	1,059	1,063	1,064	1,067	1,069	1,070	0,5000	
0,7500	1,534	1,523	1,518	1,512	1,506	1,502	1,499	1,493	1,492	1,489	1,486	1,484	0,7500	
0,9000	2,244	2,201	2,178	2,155	2,132	2,117	2,107	2,087	2,082	2,071	2,062	2,055	0,9000	
0,9500	2,845	2,774	2,737	2,700	2,661	2,637	2,621	2,588	2,580	2,563	2,548	2,538	0,9500	
0,9750	3,522	3,419	3,365	3,311	3,255	3,221	3,198	3,152	3,140	3,116	3,094	3,080	0,9750	
0,9900	4,558	4,405	4,327	4,247	4,165	4,115	4,082	4,014	3,996	3,962	3,930	3,909	0,9900	
0,9950	5,471	5,274	5,173	5,071	4,966	4,902	4,859	4,772	4,750	4,706	4,666	4,639	0,9950	
0,9990	8,129	7,803	7,638	7,469	7,297	7,192	7,122	6,980	6,944	6,872	6,807	6,762	0,9990	
0,9995	9,561	9,164	8,964	8,760	8,551	8,426	8,340	8,167	8,125	8,038	7,958	7,905	0,9995	
0,0005	0,149	0,182	0,201	0,222	0,246	0,261	0,272	0,294	0,300	0,313	0,324	0,332	0,0005	11
0,0010	0,168	0,202	0,222	0,243	0,267	0,282	0,292	0,315	0,321	0,333	0,344	0,352	0,0010	
0,0050	0,231	0,266	0,286	0,307	0,330	0,345	0,355	0,376	0,382	0,393	0,404	0,411	0,0050	
0,0100	0,268	0,304	0,323	0,344	0,367	0,381	0,391	0,411	0,417	0,428	0,438	0,445	0,0100	
0,0250	0,332	0,368	0,387	0,407	0,428	0,442	0,451	0,471	0,476	0,486	0,495	0,502	0,0250	
0,0500	0,399	0,433	0,451	0,470	0,491	0,504	0,512	0,530	0,535	0,544	0,553	0,559	0,0500	
0,1000	0,491	0,523	0,540	0,557	0,576	0,587	0,595	0,611	0,615	0,624	0,632	0,637	0,1000	
0,2500	0,693	0,718	0,731	0,744	0,758	0,767	0,773	0,784	0,787	0,794	0,799	0,803	0,2500	
0,5000	1,017	1,028	1,034	1,040	1,046	1,050	1,052	1,057	1,058	1,060	1,062	1,064	0,5000	
0,7500	1,504	1,493	1,487	1,481	1,474	1,469	1,466	1,460	1,459	1,455	1,452	1,450	0,7500	
0,9000	2,167	2,123	2,100	2,076	2,052	2,036	2,026	2,005	2,000	1,989	1,979	1,970	0,9000	
0,9500	2,719	2,646	2,609	2,570	2,531	2,507	2,490	2,457	2,448	2,431	2,415	2,404	0,9500	
0,9750	3,330	3,226	3,173	3,118	3,061	3,027	3,004	2,956	2,944	2,920	2,898	2,883	0,9750	
0,9900	4,251	4,099	4,021	3,941	3,860	3,810	3,776	3,708	3,690	3,656	3,624	3,602	0,9900	
0,9950	5,049	4,855	4,756	4,654	4,551	4,488	4,445	4,359	4,337	4,293	4,252	4,226	0,9950	
0,9990	7,321	7,008	6,848	6,684	6,517	6,416	6,348	6,210	6,175	6,105	6,041	5,998	0,9990	
0,9995	8,518	8,142	7,949	7,753	7,554	7,432	7,351	7,185	7,143	7,059	6,983	6,932	0,9995	
0,0005	0,152	0,186	0,206	0,229	0,253	0,269	0,280	0,305	0,311	0,324	0,336	0,345	0,0005	12
0,0010	0,172	0,207	0,228	0,250	0,275	0,290	0,302	0,325	0,332	0,344	0,356	0,365	0,0010	
0,0050	0,235	0,272	0,292	0,315	0,339	0,354	0,365	0,387	0,393	0,405	0,416	0,424	0,0050	
0,0100	0,273	0,309	0,330	0,352	0,375	0,390	0,401	0,422	0,428	0,440	0,450	0,458	0,0100	
0,0250	0,337	0,374	0,394	0,415	0,437	0,451	0,461	0,481	0,487	0,497	0,507	0,514	0,0250	
0,0500	0,404	0,439	0,458	0,478	0,499	0,512	0,522	0,540	0,545	0,555	0,564	0,571	0,0500	
0,1000	0,496	0,528	0,546	0,564	0,583	0,595	0,603	0,620	0,625	0,633	0,641	0,647	0,1000	
0,2500	0,695	0,721	0,734	0,748	0,762	0,771	0,777	0,789	0,792	0,799	0,804	0,808	0,2500	
0,5000	1,012	1,023	1,029	1,035	1,041	1,044	1,046	1,051	1,052	1,055	1,057	1,058	0,5000	
0,7500	1,480	1,468	1,461	1,454	1,447	1,443	1,439	1,433	1,431	1,428	1,424	1,420	0,7500	
0,9000	2,105	2,060	2,036	2,011	1,986	1,970	1,960	1,938	1,932	1,921	1,911	1,904	0,9000	
0,9500	2,617	2,544	2,505	2,466	2,426	2,401	2,384	2,350	2,341	2,323	2,307	2,296	0,9500	
0,9750	3,177	3,073	3,019	2,963	2,906	2,871	2,848	2,800	2,787	2,763	2,740	2,725	0,9750	
0,9900	4,010	3,858	3,780	3,701	3,619	3,569	3,535	3,467	3,449	3,414	3,382	3,361	0,9900	
0,9950	4,721	4,530	4,431	4,331	4,228	4,165	4,123	4,037	4,015	3,971	3,931	3,904	0,9950	
0,9990	6,709	6,405	6,249	6,090	5,928	5,829	5,763	5,627	5,593	5,524	5,462	5,420	0,9990	
0,9995	7,738	7,374	7,189	6,999	6,807	6,690	6,610	6,450	6,410	6,328	6,254	6,204	0,9995	

TABLA V (*continuación)*

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
15	0,0005	0,0641	0,001	0,005	0,015	0,029	0,045	0,061	0,076	0,091	0,105	0,117	0,129	0,0005
	0,0010	0,0516	0,001	0,008	0,021	0,039	0,057	0,075	0,092	0,108	0,123	0,137	0,149	0,0010
	0,0050	0,0439	0,005	0,023	0,049	0,076	0,102	0,126	0,147	0,166	0,183	0,198	0,212	0,0050
	0,0100	0,0316	0,010	0,037	0,070	0,103	0,132	0,158	0,181	0,202	0,219	0,235	0,249	0,0100
	0,0250	0,001	0,025	0,070	0,116	0,156	0,190	0,219	0,244	0,265	0,284	0,300	0,315	0,0250
	0,0500	0,004	0,051	0,115	0,171	0,217	0,254	0,285	0,311	0,333	0,351	0,368	0,382	0,0500
	0,1000	0,016	0,106	0,192	0,258	0,309	0,348	0,380	0,406	0,427	0,446	0,461	0,475	0,1000
	0,2500	0,105	0,293	0,407	0,480	0,530	0,567	0,596	0,618	0,637	0,652	0,665	0,676	0,2500
	0,5000	0,478	0,726	0,826	0,878	0,911	0,933	0,949	0,960	0,970	0,977	0,983	0,989	0,5000
	0,7500	1,432	1,523	1,520	1,507	1,494	1,482	1,472	1,463	1,456	1,449	1,443	1,438	0,7500
	0,9000	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,037	2,017	0,9000
	0,9500	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	0,9500
	0,9750	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123	3,060	3,008	2,963	0,9750
	0,9900	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,730	3,666	0,9900
	0,9950	10,8	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536	4,424	4,329	4,250	0,9950
	0,9990	16,6	11,3	9,335	8,253	7,567	7,091	6,741	6,471	6,256	6,081	5,935	5,812	0,9990
	0,9995	19,5	13,2	10,8	9,475	8,662	8,098	7,683	7,365	7,112	6,905	6,734	6,589	0,9995
20	0,0005	0,0640	0,001	0,005	0,015	0,030	0,046	0,062	0,079	0,094	0,109	0,123	0,136	0,0005
	0,0010	0,0516	0,001	0,008	0,022	0,039	0,058	0,077	0,095	0,112	0,128	0,143	0,156	0,0010
	0,0050	0,0439	0,005	0,023	0,050	0,077	0,104	0,129	0,151	0,171	0,190	0,206	0,221	0,0050
	0,0100	0,0316	0,010	0,037	0,071	0,105	0,135	0,162	0,187	0,208	0,227	0,244	0,259	0,0100
	0,0250	0,001	0,025	0,071	0,117	0,158	0,193	0,224	0,250	0,273	0,293	0,310	0,325	0,0250
	0,0500	0,004	0,051	0,115	0,172	0,219	0,258	0,290	0,317	0,341	0,360	0,378	0,393	0,0500
	0,1000	0,016	0,106	0,193	0,260	0,312	0,353	0,385	0,412	0,435	0,454	0,471	0,486	0,1000
	0,2500	0,104	0,292	0,406	0,480	0,531	0,569	0,598	0,622	0,641	0,656	0,670	0,681	0,2500
	0,5000	0,472	0,718	0,816	0,868	0,900	0,922	0,938	0,950	0,959	0,966	0,972	0,977	0,5000
	0,7500	1,404	1,487	1,481	1,465	1,450	1,437	1,425	1,415	1,407	1,399	1,393	1,387	0,7500
	0,9000	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,913	1,892	0,9000
	0,9500	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,310	2,278	0,9500
	0,9750	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837	2,774	2,721	2,676	0,9750
	0,9900	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	3,294	3,231	0,9900
	0,9950	9,944	6,987	5,818	5,174	4,762	4,472	4,257	4,090	3,956	3,847	3,756	3,678	0,9950
	0,9990	14,8	9,953	8,098	7,096	6,461	6,019	5,692	5,440	5,239	5,075	4,939	4,823	0,9990
	0,9995	17,2	11,4	9,195	8,018	7,274	6,759	6,378	6,085	5,852	5,662	5,503	5,370	0,9995
24	0,0005	0,0640	0,001	0,005	0,015	0,030	0,046	0,063	0,080	0,096	0,112	0,126	0,139	0,0005
	0,0010	0,0516	0,001	0,008	0,022	0,040	0,059	0,079	0,097	0,115	0,131	0,146	0,160	0,0010
	0,0050	0,0440	0,005	0,023	0,050	0,078	0,106	0,131	0,154	0,175	0,193	0,210	0,226	0,0050
	0,0100	0,0316	0,010	0,038	0,072	0,106	0,137	0,165	0,189	0,211	0,231	0,249	0,265	0,0100
	0,0250	0,001	0,025	0,071	0,117	0,159	0,195	0,226	0,253	0,277	0,297	0,315	0,331	0,0250
	0,0500	0,004	0,051	0,116	0,173	0,221	0,260	0,293	0,321	0,345	0,365	0,383	0,399	0,0500
	0,1000	0,016	0,106	0,193	0,261	0,313	0,355	0,388	0,416	0,439	0,459	0,476	0,491	0,1000
	0,2500	0,104	0,291	0,406	0,480	0,532	0,570	0,600	0,623	0,643	0,659	0,673	0,684	0,2500
	0,5000	0,469	0,714	0,812	0,863	0,895	0,917	0,932	0,944	0,953	0,961	0,967	0,972	0,5000
	0,7500	1,390	1,470	1,462	1,445	1,428	1,414	1,402	1,392	1,383	1,375	1,368	1,362	0,7500
	0,9000	2,927	2,538	2,327	2,195	2,103	2,035	1,983	1,941	1,906	1,877	1,853	1,832	0,9000
	0,9500	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255	2,216	2,183	0,9500
	0,9750	5,717	4,319	3,721	3,379	3,155	2,995	2,874	2,779	2,703	2,640	2,586	2,541	0,9750
	0,9900	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256	3,168	3,094	3,032	0,9900
	0,9950	9,551	6,661	5,519	4,890	4,486	4,202	3,991	3,826	3,695	3,587	3,497	3,420	0,9950
	0,9990	14,0	9,340	7,554	6,589	5,977	5,551	5,235	4,991	4,797	4,638	4,505	4,393	0,9990
	0,9995	16,2	10,6	8,515	7,389	6,678	6,183	5,818	5,537	5,312	5,130	4,977	4,848	0,9995

TABLA V (*continuación*)

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,160	0,197	0,220	0,244	0,272	0,290	0,303	0,331	0,338	0,353	0,368	0,378	0,0005	15
0,0010	0,181	0,219	0,242	0,266	0,294	0,312	0,325	0,352	0,359	0,374	0,388	0,398	0,0010	
0,0050	0,246	0,286	0,308	0,333	0,360	0,377	0,389	0,415	0,421	0,435	0,448	0,457	0,0050	
0,0100	0,284	0,324	0,346	0,370	0,397	0,413	0,425	0,450	0,456	0,470	0,482	0,491	0,0100	
0,0250	0,349	0,389	0,410	0,433	0,458	0,474	0,485	0,508	0,514	0,526	0,538	0,546	0,0250	
0,0500	0,416	0,454	0,474	0,496	0,520	0,534	0,545	0,566	0,571	0,583	0,593	0,600	0,0500	
0,1000	0,507	0,542	0,561	0,581	0,602	0,615	0,624	0,642	0,647	0,657	0,666	0,672	0,1000	
0,2500	0,701	0,728	0,742	0,757	0,772	0,782	0,788	0,801	0,805	0,812	0,818	0,822	0,2500	
0,5000	1,000	1,011	1,017	1,023	1,029	1,032	1,034	1,039	1,040	1,043	1,045	1,046	0,5000	
0,7500	1,426	1,413	1,405	1,397	1,389	1,383	1,380	1,372	1,370	1,366	1,362	1,359	0,7500	
0,9000	1,972	1,924	1,899	1,873	1,845	1,828	1,817	1,793	1,787	1,774	1,763	1,760	0,9000	
0,9500	2,403	2,328	2,288	2,247	2,204	2,178	2,160	2,123	2,114	2,095	2,078	2,066	0,9500	
0,9750	2,862	2,756	2,701	2,644	2,585	2,549	2,524	2,474	2,461	2,435	2,411	2,395	0,9750	
0,9900	3,522	3,372	3,294	3,214	3,132	3,081	3,047	2,977	2,959	2,923	2,891	2,868	0,9900	
0,9950	4,070	3,883	3,786	3,687	3,585	3,523	3,480	3,394	3,372	3,328	3,287	3,260	0,9950	
0,9990	5,535	5,249	5,101	4,950	4,796	4,702	4,638	4,508	4,475	4,408	4,348	4,307	0,9990	
0,9995	6,264	5,927	5,754	5,578	5,398	5,287	5,212	5,061	5,023	4,945	4,875	4,827	0,9995	
0,0005	0,169	0,210	0,235	0,263	0,295	0,316	0,331	0,364	0,373	0,391	0,409	0,421	0,0005	20
0,0010	0,191	0,233	0,258	0,286	0,318	0,339	0,354	0,386	0,395	0,412	0,430	0,441	0,0010	
0,0050	0,258	0,301	0,327	0,354	0,385	0,405	0,419	0,449	0,457	0,474	0,489	0,500	0,0050	
0,0100	0,297	0,340	0,365	0,392	0,422	0,441	0,455	0,484	0,492	0,507	0,522	0,532	0,0100	
0,0250	0,363	0,406	0,430	0,456	0,484	0,502	0,514	0,541	0,548	0,562	0,576	0,585	0,0250	
0,0500	0,430	0,471	0,493	0,518	0,544	0,560	0,572	0,597	0,603	0,616	0,628	0,637	0,0500	
0,1000	0,520	0,557	0,578	0,600	0,623	0,638	0,648	0,669	0,675	0,686	0,697	0,704	0,1000	
0,2500	0,708	0,736	0,751	0,767	0,784	0,794	0,801	0,816	0,820	0,827	0,835	0,839	0,2500	
0,5000	0,989	1,000	1,006	1,011	1,017	1,020	1,023	1,027	1,029	1,031	1,033	1,034	0,5000	
0,7500	1,374	1,358	1,349	1,340	1,330	1,324	1,319	1,310	1,307	1,302	1,298	1,290	0,7500	
0,9000	1,845	1,794	1,767	1,738	1,708	1,690	1,677	1,650	1,643	1,629	1,616	1,610	0,9000	
0,9500	2,203	2,124	2,082	2,039	1,994	1,966	1,946	1,907	1,896	1,875	1,856	1,840	0,9500	
0,9750	2,573	2,464	2,408	2,349	2,287	2,249	2,223	2,170	2,156	2,128	2,103	2,085	0,9750	
0,9900	3,088	2,938	2,859	2,778	2,695	2,643	2,608	2,535	2,517	2,479	2,445	2,421	0,9900	
0,9950	3,502	3,318	3,222	3,123	3,022	2,959	2,916	2,828	2,806	2,760	2,719	2,690	0,9950	
0,9990	4,562	4,290	4,149	4,005	3,856	3,765	3,703	3,576	3,544	3,478	3,418	3,378	0,9990	
0,9995	5,067	4,753	4,591	4,425	4,254	4,149	4,078	3,932	3,895	3,820	3,752	3,705	0,9995	
0,0005	0,174	0,218	0,244	0,274	0,309	0,332	0,349	0,385	0,395	0,415	0,435	0,449	0,0005	24
0,0010	0,196	0,241	0,268	0,298	0,332	0,355	0,371	0,407	0,416	0,436	0,456	0,469	0,0010	
0,0050	0,264	0,310	0,337	0,367	0,400	0,421	0,437	0,470	0,479	0,497	0,515	0,527	0,0050	
0,0100	0,304	0,350	0,376	0,405	0,437	0,458	0,473	0,504	0,513	0,530	0,547	0,558	0,0100	
0,0250	0,370	0,415	0,441	0,468	0,498	0,518	0,531	0,561	0,568	0,584	0,599	0,610	0,0250	
0,0500	0,437	0,480	0,504	0,530	0,558	0,576	0,588	0,615	0,622	0,636	0,650	0,659	0,0500	
0,1000	0,527	0,566	0,588	0,611	0,635	0,651	0,662	0,685	0,691	0,703	0,715	0,723	0,1000	
0,2500	0,712	0,741	0,757	0,773	0,791	0,801	0,809	0,825	0,829	0,837	0,845	0,850	0,2500	
0,5000	0,983	0,994	1,000	1,006	1,011	1,015	1,017	1,022	1,023	1,025	1,027	1,028	0,5000	
0,7500	1,347	1,331	1,321	1,311	1,300	1,293	1,289	1,278	1,275	1,270	1,264	1,260	0,7500	
0,9000	1,783	1,730	1,702	1,672	1,641	1,621	1,607	1,579	1,571	1,556	1,542	1,530	0,9000	
0,9500	2,108	2,027	1,984	1,939	1,892	1,863	1,842	1,800	1,790	1,768	1,747	1,733	0,9500	
0,9750	2,437	2,327	2,269	2,209	2,146	2,107	2,080	2,024	2,010	1,981	1,954	1,935	0,9750	
0,9900	2,889	2,738	2,659	2,577	2,492	2,440	2,403	2,329	2,310	2,271	2,235	2,211	0,9900	
0,9950	3,246	3,062	2,967	2,868	2,765	2,702	2,658	2,569	2,546	2,500	2,457	2,428	0,9950	
0,9990	4,139	3,873	3,735	3,593	3,447	3,356	3,295	3,168	3,136	3,070	3,010	2,969	0,9990	
0,9995	4,556	4,251	4,094	3,932	3,764	3,661	3,591	3,447	3,410	3,336	3,267	3,221	0,9995	

TABLA V (*continuación*)

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
30	0,0005	0,0640	0,001	0,005	0,015	0,030	0,047	0,065	0,082	0,098	0,114	0,129	0,143	0,0005
	0,0010	0,0516	0,001	0,008	0,022	0,040	0,060	0,080	0,099	0,117	0,134	0,150	0,164	0,0010
	0,0050	0,0440	0,005	0,024	0,050	0,079	0,107	0,133	0,156	0,178	0,197	0,215	0,231	0,0050
	0,0100	0,0316	0,010	0,038	0,072	0,107	0,138	0,167	0,192	0,215	0,235	0,254	0,270	0,0100
	0,0250	0,001	0,025	0,071	0,118	0,161	0,197	0,229	0,257	0,281	0,302	0,321	0,337	0,0250
	0,0500	0,004	0,051	0,116	0,174	0,222	0,263	0,296	0,325	0,349	0,370	0,389	0,405	0,0500
	0,1000	0,016	0,106	0,193	0,262	0,315	0,357	0,391	0,420	0,444	0,464	0,482	0,497	0,1000
	0,2500	0,103	0,290	0,406	0,480	0,532	0,571	0,601	0,625	0,645	0,661	0,675	0,688	0,2500
	0,5000	0,466	0,709	0,807	0,858	0,890	0,912	0,927	0,939	0,948	0,955	0,961	0,966	0,5000
	0,7500	1,376	1,452	1,443	1,424	1,407	1,392	1,380	1,369	1,359	1,351	1,343	1,337	0,7500
	0,9000	2,881	2,489	2,276	2,142	2,049	1,980	1,927	1,884	1,849	1,819	1,794	1,773	0,9000
	0,9500	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	2,126	2,092	0,9500
	0,9750	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575	2,511	2,458	2,412	0,9750
	0,9900	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067	2,979	2,906	2,843	0,9900
	0,9950	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,451	3,344	3,255	3,179	0,9950
	0,9990	13,3	8,773	7,054	6,125	5,534	5,122	4,817	4,582	4,393	4,239	4,110	4,001	0,9990
	0,9995	15,2	9,897	7,894	6,817	6,135	5,661	5,311	5,040	4,825	4,648	4,501	4,376	0,9995
40	0,0005	0,0640	0,001	0,005	0,016	0,031	0,048	0,066	0,083	0,101	0,117	0,132	0,147	0,0005
	0,0010	0,0516	0,001	0,008	0,022	0,041	0,061	0,081	0,101	0,119	0,137	0,153	0,169	0,0010
	0,0050	0,0440	0,005	0,024	0,051	0,080	0,108	0,135	0,159	0,181	0,201	0,220	0,237	0,0050
	0,0100	0,0316	0,010	0,038	0,073	0,108	0,140	0,169	0,195	0,219	0,240	0,259	0,276	0,0100
	0,0250	0,001	0,025	0,071	0,119	0,162	0,200	0,232	0,260	0,285	0,307	0,327	0,344	0,0250
	0,0500	0,004	0,051	0,116	0,175	0,224	0,265	0,299	0,329	0,354	0,376	0,395	0,412	0,0500
	0,1000	0,016	0,106	0,194	0,263	0,317	0,360	0,394	0,423	0,448	0,469	0,487	0,503	0,1000
	0,2500	0,103	0,290	0,405	0,480	0,533	0,572	0,603	0,627	0,647	0,664	0,679	0,691	0,2500
	0,5000	0,463	0,705	0,802	0,854	0,885	0,907	0,922	0,934	0,943	0,950	0,956	0,961	0,5000
	0,7500	1,363	1,435	1,424	1,404	1,386	1,371	1,357	1,345	1,335	1,327	1,319	1,312	0,7500
	0,9000	2,835	2,440	2,226	2,091	1,997	1,927	1,873	1,829	1,793	1,763	1,737	1,715	0,9000
	0,9500	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077	2,038	2,003	0,9500
	0,9750	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452	2,388	2,334	2,288	0,9750
	0,9900	7,314	5,178	4,313	3,828	3,514	3,291	3,124	2,993	2,888	2,801	2,727	2,665	0,9900
	0,9950	8,828	6,066	4,976	4,374	3,986	3,713	3,509	3,350	3,222	3,117	3,028	2,953	0,9950
	0,9990	12,6	8,251	6,595	5,698	5,128	4,731	4,436	4,207	4,024	3,874	3,749	3,643	0,9990
	0,9995	14,4	9,248	7,329	6,296	5,643	5,188	4,852	4,591	4,384	4,214	4,071	3,950	0,9995
60	0,0005	0,0640	0,001	0,005	0,016	0,031	0,048	0,067	0,085	0,103	0,120	0,136	0,151	0,0005
	0,0010	0,0516	0,001	0,008	0,022	0,041	0,062	0,083	0,103	0,122	0,140	0,158	0,174	0,0010
	0,0050	0,0440	0,005	0,024	0,051	0,081	0,110	0,137	0,162	0,185	0,206	0,225	0,243	0,0050
	0,0100	0,0316	0,010	0,038	0,073	0,109	0,142	0,172	0,199	0,223	0,245	0,265	0,283	0,0100
	0,0250	0,001	0,025	0,071	0,120	0,163	0,202	0,235	0,264	0,290	0,313	0,333	0,351	0,0250
	0,0500	0,004	0,051	0,117	0,176	0,226	0,267	0,303	0,333	0,359	0,382	0,402	0,419	0,0500
	0,1000	0,016	0,106	0,194	0,264	0,318	0,362	0,398	0,428	0,453	0,475	0,494	0,510	0,1000
	0,2500	0,102	0,289	0,405	0,480	0,534	0,573	0,604	0,629	0,650	0,667	0,682	0,695	0,2500
	0,5000	0,460	0,701	0,798	0,849	0,880	0,901	0,917	0,928	0,937	0,945	0,951	0,956	0,5000
	0,7500	1,349	1,419	1,405	1,385	1,366	1,349	1,335	1,323	1,312	1,303	1,294	1,287	0,7500
	0,9000	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,680	1,657	0,9000
	0,9500	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,952	1,917	0,9500
	0,9750	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334	2,270	2,216	2,169	0,9750
	0,9900	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,559	2,496	0,9900
	0,9950	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008	2,904	2,817	2,742	0,9950
	0,9990	12,0	7,768	6,171	5,307	4,757	4,372	4,086	3,865	3,687	3,542	3,419	3,315	0,9990
	0,9995	13,5	8,651	6,812	5,823	5,196	4,759	4,436	4,185	3,984	3,820	3,683	3,566	0,9995

TABLA V (*continuación)*

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,179	0,226	0,254	0,287	0,325	0,350	0,368	0,409	0,420	0,444	0,467	0,483	0,0005	30
0,0010	0,202	0,250	0,278	0,311	0,348	0,373	0,391	0,431	0,442	0,465	0,487	0,502	0,0010	
0,0050	0,271	0,320	0,349	0,381	0,416	0,440	0,457	0,494	0,504	0,525	0,545	0,559	0,0050	
0,0100	0,311	0,360	0,388	0,419	0,454	0,477	0,493	0,528	0,538	0,557	0,576	0,589	0,0100	
0,0250	0,378	0,426	0,453	0,482	0,515	0,536	0,551	0,583	0,592	0,610	0,627	0,639	0,0250	
0,0500	0,445	0,490	0,516	0,543	0,573	0,593	0,606	0,636	0,643	0,659	0,675	0,685	0,0500	
0,1000	0,534	0,575	0,598	0,622	0,649	0,666	0,678	0,703	0,710	0,723	0,736	0,746	0,1000	
0,2500	0,716	0,746	0,763	0,780	0,798	0,810	0,818	0,835	0,839	0,848	0,856	0,862	0,2500	
0,5000	0,978	0,989	0,994	1,000	1,006	1,009	1,011	1,016	1,017	1,019	1,021	1,023	0,5000	
0,7500	1,321	1,303	1,293	1,282	1,270	1,263	1,257	1,245	1,242	1,236	1,230	1,226	0,7500	
0,9000	1,722	1,667	1,638	1,606	1,573	1,552	1,538	1,507	1,499	1,482	1,467	1,460	0,9000	
0,9500	2,015	1,932	1,887	1,841	1,792	1,761	1,740	1,695	1,683	1,660	1,637	1,620	0,9500	
0,9750	2,307	2,195	2,136	2,074	2,009	1,968	1,940	1,882	1,866	1,835	1,806	1,787	0,9750	
0,9900	2,700	2,549	2,469	2,386	2,299	2,245	2,208	2,131	2,111	2,070	2,032	2,006	0,9900	
0,9950	3,006	2,823	2,727	2,628	2,524	2,459	2,415	2,323	2,300	2,251	2,207	2,176	0,9950	
0,9990	3,753	3,493	3,357	3,217	3,072	2,981	2,920	2,792	2,760	2,693	2,631	2,589	0,9990	
0,9995	4,094	3,798	3,644	3,486	3,321	3,219	3,149	3,005	2,969	2,894	2,824	2,777	0,9995	
0,0005	0,185	0,235	0,266	0,301	0,343	0,371	0,392	0,439	0,452	0,479	0,506	0,526	0,0005	40
0,0010	0,209	0,259	0,290	0,326	0,367	0,395	0,415	0,461	0,473	0,500	0,526	0,545	0,0010	
0,0050	0,279	0,331	0,362	0,396	0,436	0,462	0,481	0,523	0,534	0,559	0,582	0,599	0,0050	
0,0100	0,319	0,371	0,401	0,435	0,473	0,498	0,517	0,556	0,567	0,590	0,612	0,628	0,0100	
0,0250	0,387	0,437	0,466	0,498	0,533	0,557	0,573	0,610	0,620	0,640	0,660	0,674	0,0250	
0,0500	0,454	0,502	0,529	0,558	0,591	0,612	0,627	0,660	0,669	0,687	0,705	0,717	0,0500	
0,1000	0,542	0,585	0,610	0,636	0,664	0,683	0,696	0,724	0,731	0,747	0,762	0,772	0,1000	
0,2500	0,720	0,752	0,769	0,787	0,807	0,819	0,828	0,846	0,851	0,861	0,870	0,877	0,2500	
0,5000	0,972	0,983	0,989	0,994	1,000	1,003	1,006	1,010	1,011	1,014	1,016	1,017	0,5000	
0,7500	1,295	1,276	1,265	1,253	1,240	1,231	1,225	1,212	1,208	1,201	1,193	1,190	0,7500	
0,9000	1,662	1,605	1,574	1,541	1,506	1,483	1,467	1,434	1,425	1,406	1,389	1,380	0,9000	
0,9500	1,924	1,839	1,793	1,744	1,693	1,660	1,637	1,589	1,577	1,551	1,526	1,509	0,9500	
0,9750	2,182	2,068	2,007	1,943	1,875	1,832	1,803	1,741	1,724	1,691	1,659	1,637	0,9750	
0,9900	2,522	2,369	2,288	2,203	2,114	2,058	2,019	1,938	1,917	1,874	1,833	1,805	0,9900	
0,9950	2,781	2,598	2,502	2,401	2,296	2,230	2,184	2,088	2,064	2,012	1,965	1,932	0,9950	
0,9990	3,400	3,145	3,011	2,872	2,727	2,636	2,574	2,444	2,410	2,341	2,277	2,233	0,9990	
0,9995	3,677	3,388	3,238	3,081	2,918	2,816	2,747	2,602	2,564	2,487	2,415	2,366	0,9995	
0,0005	0,192	0,245	0,278	0,318	0,364	0,397	0,421	0,476	0,491	0,525	0,559	0,584	0,0005	60
0,0010	0,216	0,270	0,304	0,343	0,389	0,420	0,444	0,498	0,513	0,545	0,578	0,602	0,0010	
0,0050	0,287	0,343	0,376	0,414	0,458	0,488	0,510	0,559	0,572	0,602	0,631	0,653	0,0050	
0,0100	0,328	0,383	0,416	0,453	0,495	0,524	0,545	0,591	0,604	0,632	0,659	0,679	0,0100	
0,0250	0,396	0,450	0,481	0,515	0,555	0,581	0,600	0,642	0,654	0,678	0,703	0,720	0,0250	
0,0500	0,463	0,514	0,543	0,575	0,611	0,635	0,652	0,689	0,700	0,722	0,743	0,759	0,0500	
0,1000	0,550	0,596	0,622	0,650	0,682	0,702	0,717	0,749	0,757	0,776	0,794	0,806	0,1000	
0,2500	0,725	0,758	0,776	0,795	0,816	0,830	0,839	0,860	0,865	0,877	0,888	0,896	0,2500	
0,5000	0,967	0,978	0,983	0,989	0,994	0,998	1,000	1,004	1,006	1,008	1,010	1,011	0,5000	
0,7500	1,269	1,248	1,236	1,223	1,208	1,198	1,191	1,176	1,172	1,163	1,154	1,147	0,7500	
0,9000	1,603	1,543	1,511	1,476	1,437	1,413	1,395	1,358	1,348	1,326	1,306	1,290	0,9000	
0,9500	1,836	1,748	1,700	1,649	1,594	1,559	1,534	1,481	1,467	1,438	1,409	1,390	0,9500	
0,9750	2,061	1,944	1,882	1,815	1,744	1,699	1,667	1,599	1,581	1,543	1,507	1,482	0,9750	
0,9900	2,352	2,198	2,115	2,028	1,936	1,877	1,836	1,749	1,726	1,678	1,633	1,601	0,9900	
0,9950	2,570	2,387	2,290	2,187	2,079	2,010	1,962	1,861	1,834	1,779	1,726	1,689	0,9950	
0,9990	3,078	2,826	2,694	2,555	2,409	2,316	2,252	2,118	2,082	2,009	1,939	1,890	0,9990	
0,9995	3,299	3,017	2,869	2,714	2,551	2,449	2,378	2,228	2,189	2,108	2,031	1,978	0,9995	

TABLA V (*continuación)*

		m												
n	*p*	1	2	3	4	5	6	7	8	9	10	11	12	*p*
120	0,0005	0,0640	0,001	0,005	0,016	0,031	0,049	0,068	0,087	0,105	0,123	0,140	0,156	0,0005
	0,0010	0,0516	0,001	0,008	0,023	0,042	0,063	0,084	0,105	0,125	0,144	0,162	0,179	0,0010
	0,0050	0,0439	0,005	0,024	0,051	0,081	0,111	0,139	0,165	0,189	0,211	0,231	0,249	0,0050
	0,0100	0,0316	0,010	0,038	0,074	0,110	0,143	0,174	0,202	0,227	0,250	0,271	0,290	0,0100
	0,0250	0,001	0,025	0,072	0,120	0,165	0,204	0,238	0,268	0,295	0,318	0,340	0,359	0,0250
	0,0500	0,004	0,051	0,117	0,177	0,227	0,270	0,306	0,337	0,364	0,388	0,408	0,427	0,0500
	0,1000	0,016	0,105	0,194	0,265	0,320	0,365	0,401	0,432	0,458	0,480	0,500	0,518	0,1000
	0,2500	0,102	0,288	0,405	0,480	0,534	0,574	0,606	0,631	0,653	0,670	0,686	0,699	0,2500
	0,5000	0,458	0,697	0,793	0,844	0,875	0,896	0,912	0,923	0,932	0,939	0,945	0,950	0,5000
	0,7500	1,336	1,402	1,387	1,365	1,345	1,328	1,313	1,300	1,289	1,279	1,270	1,262	0,7500
	0,9000	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,625	1,601	0,9000
	0,9500	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,869	1,834	0,9500
	0,9750	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222	2,157	2,102	2,055	0,9750
	0,9900	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,399	2,336	0,9900
	0,9950	8,179	5,539	4,497	3,921	3,548	3,285	3,087	2,933	2,808	2,705	2,618	2,544	0,9950
	0,9990	11,4	7,321	5,781	4,947	4,416	4,044	3,767	3,552	3,379	3,237	3,118	3,016	0,9990
	0,9995	12,8	8,104	6,341	5,392	4,790	4,370	4,059	3,817	3,623	3,464	3,331	3,217	0,9995
∞	0,0005	0,0639	0,001	0,005	0,016	0,032	0,050	0,069	0,089	0,108	0,127	0,144	0,161	0,0005
	0,0010	0,0516	0,001	0,008	0,023	0,042	0,064	0,085	0,107	0,128	0,148	0,167	0,185	0,0010
	0,0050	0,0439	0,005	0,024	0,052	0,082	0,113	0,141	0,168	0,193	0,216	0,237	0,256	0,0050
	0,0100	0,0316	0,010	0,038	0,074	0,111	0,145	0,177	0,206	0,232	0,256	0,278	0,298	0,0100
	0,0250	0,001	0,025	0,072	0,121	0,166	0,206	0,241	0,272	0,300	0,325	0,347	0,367	0,0250
	0,0500	0,004	0,051	0,117	0,178	0,229	0,273	0,310	0,342	0,369	0,394	0,416	0,436	0,0500
	0,1000	0,016	0,105	0,195	0,266	0,322	0,367	0,405	0,436	0,463	0,487	0,507	0,525	0,1000
	0,2500	0,102	0,288	0,404	0,481	0,535	0,576	0,608	0,634	0,655	0,674	0,689	0,703	0,2500
	0,5000	0,455	0,693	0,789	0,839	0,870	0,891	0,907	0,918	0,927	0,934	0,940	0,945	0,5000
	0,7500	1,323	1,386	1,369	1,346	1,325	1,307	1,291	1,277	1,265	1,255	1,246	1,237	0,7500
	0,9000	2,706	2,303	2,084	1,945	1,847	1,774	1,717	1,670	1,632	1,599	1,570	1,546	0,9000
	0,9500	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880	1,831	1,789	1,752	0,9500
	0,9750	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114	2,048	1,993	1,945	0,9750
	0,9900	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407	2,321	2,248	2,185	0,9900
	0,9950	7,879	5,298	4,279	3,715	3,350	3,091	2,897	2,744	2,621	2,519	2,432	2,358	0,9950
	0,9990	10,8	6,908	5,422	4,617	4,103	3,743	3,474	3,266	3,098	2,959	2,842	2,742	0,9990
	0,9995	12,1	7,601	5,910	4,999	4,421	4,017	3,717	3,483	3,296	3,142	3,012	2,902	0,9995

TABLA V (*continuación)*

							m							
p	15	20	24	30	40	50	60	100	120	200	500	∞	*p*	*n*
0,0005	0,199	0,257	0,293	0,337	0,390	0,428	0,457	0,525	0,545	0,590	0,638	0,676	0,0005	120
0,0010	0,223	0,282	0,319	0,362	0,415	0,452	0,480	0,547	0,566	0,609	0,655	0,691	0,0010	
0,0050	0,297	0,356	0,393	0,435	0,485	0,519	0,545	0,605	0,623	0,661	0,702	0,733	0,0050	
0,0100	0,338	0,397	0,433	0,474	0,522	0,555	0,579	0,636	0,652	0,688	0,726	0,755	0,0100	
0,0250	0,406	0,464	0,498	0,536	0,580	0,610	0,632	0,683	0,698	0,730	0,763	0,788	0,0250	
0,0500	0,473	0,527	0,559	0,594	0,634	0,662	0,682	0,727	0,740	0,768	0,797	0,819	0,0500	
0,1000	0,560	0,609	0,636	0,667	0,702	0,725	0,742	0,780	0,791	0,814	0,838	0,856	0,1000	
0,2500	0,730	0,765	0,784	0,805	0,828	0,843	0,854	0,877	0,884	0,898	0,912	0,923	0,2500	
0,5000	0,961	0,972	0,978	0,983	0,989	0,992	0,994	0,999	1,000	1,002	1,004	1,006	0,5000	
0,7500	1,243	1,220	1,207	1,192	1,175	1,164	1,156	1,137	1,131	1,120	1,108	1,099	0,7500	
0,9000	1,545	1,482	1,447	1,409	1,368	1,340	1,320	1,277	1,265	1,239	1,212	1,193	0,9000	
0,9500	1,750	1,659	1,608	1,554	1,495	1,457	1,429	1,369	1,352	1,316	1,280	1,254	0,9500	
0,9750	1,945	1,825	1,760	1,690	1,614	1,565	1,530	1,454	1,433	1,388	1,343	1,310	0,9750	
0,9900	2,191	2,035	1,950	1,860	1,763	1,700	1,656	1,559	1,533	1,477	1,421	1,381	0,9900	
0,9950	2,373	2,188	2,089	1,984	1,871	1,798	1,747	1,636	1,606	1,541	1,478	1,431	0,9950	
0,9990	2,783	2,534	2,402	2,262	2,113	2,017	1,950	1,806	1,767	1,684	1,603	1,543	0,9990	
0,9995	2,958	2,681	2,534	2,379	2,214	2,109	2,035	1,877	1,834	1,744	1,655	1,590	0,9995	
0,0005	0,207	0,270	0,311	0,360	0,423	0,469	0,506	0,599	0,629	0,703	0,805	0,999	0,0005	∞
0,0010	0,232	0,296	0,337	0,386	0,448	0,493	0,529	0,619	0,648	0,719	0,816	0,999	0,0010	
0,0050	0,307	0,372	0,412	0,460	0,518	0,560	0,592	0,673	0,699	0,761	0,845	0,999	0,0050	
0,0100	0,349	0,413	0,452	0,498	0,554	0,594	0,625	0,701	0,724	0,782	0,859	0,999	0,0100	
0,0250	0,417	0,480	0,517	0,560	0,611	0,647	0,675	0,742	0,763	0,814	0,880	0,999	0,0250	
0,0500	0,484	0,543	0,577	0,616	0,663	0,695	0,720	0,779	0,798	0,841	0,898	0,999	0,0500	
0,1000	0,570	0,622	0,652	0,687	0,726	0,754	0,774	0,824	0,839	0,874	0,920	0,999	0,1000	
0,2500	0,736	0,773	0,793	0,816	0,842	0,859	0,872	0,901	0,910	0,931	0,957	1,000	0,2500	
0,5000	0,956	0,967	0,972	0,978	0,983	0,987	0,989	0,993	0,994	0,997	0,999	1,000	0,5000	
0,7500	1,216	1,191	1,177	1,160	1,140	1,127	1,116	1,091	1,084	1,066	1,042	1,000	0,7500	
0,9000	1,487	1,421	1,383	1,342	1,295	1,263	1,240	1,185	1,169	1,130	1,082	1,001	0,9000	
0,9500	1,666	1,571	1,517	1,459	1,394	1,350	1,318	1,243	1,221	1,170	1,106	1,001	0,9500	
0,9750	1,833	1,708	1,640	1,566	1,484	1,428	1,388	1,296	1,268	1,205	1,128	1,001	0,9750	
0,9900	2,039	1,878	1,791	1,696	1,592	1,523	1,473	1,358	1,325	1,247	1,153	1,001	0,9900	
0,9950	2,187	2,000	1,898	1,789	1,669	1,590	1,533	1,402	1,364	1,276	1,170	1,001	0,9950	
0,9990	2,513	2,266	2,132	1,990	1,835	1,733	1,660	1,494	1,447	1,338	1,207	1,001	0,9990	
0,9995	2,648	2,375	2,228	2,072	1,902	1,791	1,712	1,532	1,480	1,362	1,221	1,001	0,9995	

Referencias bibliográficas

Abelson, R. P. (1995): *Statistics as a principled argument*, Hillsdale, NJ: LEA.
American Psychological Association (2010): *Manual de publicaciones de la American Psychological Association,* México: El Manual Moderno.
Amón, J. (1993): *Estadística para psicólogos I. Estadística descriptiva,* Madrid: Pirámide.
Amón, J. (1996): *Estadística para psicólogos II. Probabilidad y estadística inferencial,* Madrid: Pirámide.
Ato, M. y Vallejo, G. (2007): *Diseños experimentales en psicología,* Madrid: Pirámide.
Azorín, F. y Sánchez-Crespo, J. L. (1986): *Métodos y aplicaciones del muestreo,* Madrid: Alianza.
Bernoulli, J. (1713): *Ars Conjectandi,* Basil: Thurnisiorum.
Borges, A., San Luis, C., Sánchez-Bruno, J. A. y Cañadas, I. (2001): «El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa», *Psicothema, 13*, 173-178.
Botella, J., León, O., San Martín, R. y Barriopedro, M. I. (2001): *Análisis de datos en psicología I,* Madrid: Pirámide.
Carrobles, J. A., Remor, E. y Rodríguez-Alzamora, L. (2003): «Afrontamiento, apoyo social percibido y distrés emocional en pacientes con infección por VIH», *Psicothema, 15*(3), 420-426.
Clairin, R. y Brion, P. (2001): *Manual de muestreo*, Madrid: La Muralla.
Cohen, J. y Cohen, P. (1983): *Applied multiple regression/correlation analysis for the Behavioral Sciences* (2.ª ed.), Hillsdale, N. J.: Lawrence Erlbaum Associates.
Cowles, M. (1989): *Statistics in Psychology: an Historical Perspective,* Hillsdale, N. J.: Lawrence Erlbaum Associates.
De Moivre, A. (1733): *Approximatio ad Summam Terminorum Binomii* $(a + b)^n$ *in Seriem Expansi,* Londres.
Deaño, A. (1999): *Introducción a la lógica formal,* Madrid: Alianza editorial.
Etxeberría, J. (1999): *Regresión múltiple,* Madrid: La Muralla.
Fang-peng, G. y Dong, Y. (2010): A Study on College Students' Anxiety to Spoken English. *Canadian Social Science, 6*(2), 95-101.
Fisher, R. A. (1918): «The correlation between Relations on the Supposition of Mendelian Inheritance», *Transactions of the Royal Society of Edinburgh,* LII.
Frick, R. W. (1998): «Interpreting statistical testing: Process and propensity, not population and random sampling», *Behavior Research Methods, 30,* 527-535.
Galton, F. (1883): *Inquiries into Human Faculty and its Development,* Londres y Nueva York.
Galton, F. (1885): «Regression towards mediocrity in Heredity Stature», *Journal of the Anthropological Institute, XV,* 246-264.

García-Jiménez, M. V., Alvarado, J. M. y Jiménez-Blanco, A. (2000): «La predicción del rendimiento académico: regresión lineal versus regresión logística», *Psicothema, 12*, Supl. n.º 2, 248-252.

Gigerenzer, G. (1998): «We need statistical thinking, no statistical rituals», *Behavioral and Brain Sciences, 21*, 199-200.

Gigerenzer G., Swijtnik, Z., Porter, T., Daston, L., Beatty, J. y Kruger, L. (1989): *The Empire of Chance. How probability changed science and every day life,* Cambridge. Cambridge University Press.

Glass, G. V. y Stanley, J. C. (1970): *Statistical Methods in Education and Psychology,* Englewood Cliffs, N. J.: Prentice-Hall (trad. al español en 1980, Madrid: Prentice-Hall).

Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997): *What if there were no significance tests?* Mahwah, NJ: LAU.

Hays, W. L. (1988): *Statistics* (4.ª ed.), Nueva York: Holt Rinehart and Wiston Inc.

Henry, G. T. (1995): *Graphing data: techniques for display and analysis,* Beverly-Hills, CA: Sage.

Howard, G. S., Maxwell, S. E. y Fleming, K. J. (2000): «The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis», *Psychological Methods, 3*, 315-332.

Howell, D. C. (2009): *Statistical Methods for Psychology* (7.ª ed.), Belmont, CA: Wadsworth Cengage, Leaving.

Huff, D. (1954): *How to Lie with Statistics,* Nueva York: Norton & Co Ltd.

Jáñez, L. (1989): *Fundamentos de psicología matemática,* Madrid: Pirámide.

Kelley, T. L. (1914): «Comparable Measures», *Journal of Educational Psychology, V,* 589-595.

Kruskall, W. H. (1974): «Estadística: su objeto», en *Enciclopedia Internacional de las Ciencias Sociales,* Madrid: Aguilar.

Laplace, P. (1812): *Théorie Analytique des probabilités,* París: Courcier.

León, O. G. y Montero, I. (2003): *Métodos de investigación en psicología y educación* (3.ª ed.), Madrid: McGraw-Hill.

Martin, M. (1977): «Reading while listening: a linear model of selective attention», *Journal of Verbal Learning and Verbal Behavior, 16,* 453-463.

Martín-Baró, I. (1985): «El hacinamiento residencial: ideologización y verdad de un problema real», *Revista de Psicología Social, 0,* 31-50.

Nickerson, R. S. (2000): «Null hypothesis significance testing: a review of an old and continuing controversy», *Psychological Methods, 5*, 241-301.

Norusis, M. J. (2011): *IBM SPSS Statistics 19 guide to data analysis*. Inc. SPSS.

O'Connor, M. J., Frankel, F., Paley, B., Schonfeld, A. M., Carpenter, E., Laugeson, E. A. y Marquardt, R. (2006): «A Controlled Social Skills Training for Children With Fetal Alcohol Spectrum Disorders», *Journal of Consulting and Clinical Psychology, 74*(4), 639-648.

Palmer, A. L. (1999): *Análisis de datos. Etapa exploratoria,* Madrid: Pirámide.

Pardo, A. y San Martín, R. (2010): *Análisis de Datos en Ciencias Sociales y de la Salud II,* Madrid: Síntesis.

Pardo, A., Ruiz, M. A. y San Martín, R. (2009): *Análisis de Datos en Ciencias Sociales y de la Salud I,* Madrid: Síntesis.

Pearson, K. (1894): «Contributions to the Mathematical Theory of Evolution I. On the Dissection of Asymmetrical Frequency Curves», *Philosophical Transactions,* A, CLXXXV, parte I.

Pearson, K. (1896): «Regression, Heredity and Panmixia», *Philosophical Transactions,* A, CLXXXVII.

Pearson, K. (1906): «Skew Variation, A Rejoinder», *Biometrika, IV,* 173.

Peña, D. (1987): *Estadística: Modelos y métodos,* Madrid: Alianza Universidad.
Perea, M. (1999): Tiempos de reacción y psicología: dos procedimientos para evitar el sesgo debido al temaño muestral, *Psicológica, 20,* 13-21.
Pew, R. W. (1969): «The speed-accuracy operation characteristic», Attention and Performance II. *Acta Psychologica, 30,* 16-26.
Redzuan, M., Juhari, R. B., Yousefi, F., Mansor, M. B. y Talib, M. A. (2010): «The Relationship between Gender, Age, Depression and Academic Achievement», *Current Research in Psychology, 1*(1), 61-66.
Ríos-Rísquez, M. I., Sánchez-Meca, J. y Godoy-Fernández, C. (2010): «Personalidad resistente, autoeficacia y estado general de salud en profesionales de Enfermería de cuidados intensivos y urgencias», *Psicothema, 22*(4), 600-605.
Rivadulla, A. (1991): *Probabilidad e inferencia científica,* Barcelona: Anthropos.
Runyon, R. P. y Haber, A. (1971): *Fundamentals of behavioral statistics*, Reading, MA: Addison Wesley.
Sánchez-Carrión, J. J. (1999): *Manual de análisis estadístico de los datos* (2.ª ed.), Madrid: Alianza editorial.
Schweizera, K. y Moosbrugger, H. (2004): «Attention and working memory as predictors of intelligence», *Intelligence, 32,* 329-347.
Sheppard, W. F. (1899): «On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation», *Philosophical Transactions,* A, *CXCII,* 105-106.
Sheppard, W. F. (1902): «New Tables of the Probability Integral», *Biometrika, II,* 174-190.
Solanas, A., Salafranca, L., Fauquet, J. y Núñez, M. I. (2005): *Estadística descriptiva en ciencias del comportamiento,* Madrid: Thomson.
Stern, W. (1912): *Psychologischen Methoden Der Intelligenzprüfung,* Leipzig.
Stevens, S. S. (1946): «On the theory of scales of measurement», *Science, 103,* 677-680.
Stevens, S. S. (1975): *Psychophysics,* Nueva York: Wiley.
Tatsuoka, M. (1976): *The use of Multiple Regression Equations,* Champaign, Il.: IPAT.
Treisman, A. y Gelade, G. (1980): «A feature-integration theory of attention», *Cognitive Psychology, 12,* 97-136.
Tukey, J. W. (1962): «The future of data analysis», *The Annals of Mathematical Statistics, 33*, 1-67.
Tukey, J. W. (1977): *Exploratory Data Analysis, Reading,* MA: Addison Wesley.
Walker, H. M. (1975): *Studies in the History of Statistical Method,* Nueva York: Arno Press.
Ximénez, C. y Revuelta, J. (2010): «Factorial invariance in a repeated measures design: an application to the study of person-organization fit», *Spanish Journal of Psychology, 13*(1), 485-493.
Ximénez, C. y Revuelta, J. (2011): *Cuaderno de prácticas de análisis de datos con SPSS,* Madrid: UAM ediciones.
Yerkes, R. M. y Dodson, J. D. (1908): «The relation of strength of stimulus to rapidity of habit-formation», *Journal of Comparative and Neurological Psychology, 18,* 459-482.

TÍTULOS RELACIONADOS

Análisis de datos. Etapa exploratoria, *A. L. Palmer Pol.*

Análisis de datos en psicología, *F. J. Pérez Santamaría, V. Manzano Arrondo y H. Fazeli Khalili.*

Análisis de datos en psicología I, *J. Botella Ausina, M. Suero Suñe y C. Ximénez Gómez.*

Análisis de datos en psicología II, *A. Pardo Merino y R. San Martín Castellanos.*

Diseños experimentales en psicología, *M. Ato García y G. Vallejo Seco.*

Estadística para psicólogos I. Estadística descriptiva, *J. Amón Hortelano.*

Estadística para psicólogos II. Probabilidad. Estadística inferencial, *J. Amón Hortelano.*

Fundamentos metodológicos en psicología y ciencias afines, *R. Moreno Rodríguez, R. J. Martínez Cervantes y S. Chacón Moscoso.*

Introducción a la teoría de respuesta a los ítems, *J. Muñiz Fernández.*

Introducción a los métodos de investigación de la psicología, *A. R. Delgado González y G. Prieto Adánez.*

Métodos de investigación y análisis de datos en ciencias sociales y de la salud, *S. Cubo Delgado, B. Martín Marín y J. L. Ramos Sánchez (coords.).*

Problemas resueltos de análisis de datos, *F. J. Pérez Santamaría, V. Manzano Arrondo y H. Fazeli Khalili.*

Teoría clásica de los tests, *J. Muñiz Fernández.*