Estadística descriptiva y cálculo de probabilidades

Isabel Castillo Manrique
Marta Guijarro Garvi

Prólogo:
José Luis Rojo García

PEARSON
Prentice Hall

ESTADÍSTICA DESCRIPTIVA Y CÁLCULO DE PROBABILIDADES

ESTADÍSTICA DESCRIPTIVA Y CÁLCULO DE PROBABILIDADES

Isabel Castillo Manrique
Marta Guijarro Garvi
Profesoras del Departamento de Economía
Universidad de Cantabria

Prólogo
José Luis Rojo García
Catedrático de Economía Aplicada
Universidad de Valladolid

PEARSON
Prentice Hall

Madrid • México • Santafé de Bogotá • Buenos Aires • Caracas • Lima •
Montevideo • San Juan • San José • Santiago • São Paulo • White Plains

Datos de catalogación bibliográfica

Isabel Castillo Manrique-Marta Guijarro Garvi
Estadística descriptiva y cálculo de probabilidades

PEARSON EDUCACIÓN, S.A., Madrid, 2006

ISBN: 978-84-832-2209-6
MATERIA: Estadística matemática 519.2

Formato: 170 × 240 mm Páginas: 440

Marta Guijarro Garvi-Isabel Castillo Manrique
Estadística descriptiva y cálculo de probabilidades

ISBN: 84-205-4806-5
Depósito Legal: M.

Equipo editorial
Editor: Juan Luis Posadas
Técnico editorial: Elena Bazaco
Equipo de producción:
Director: José Antonio Clares
Técnico: José Antonio Hernán
Diseño de cubierta: Equipo de diseño de PEARSON EDUCACIÓN, S.A.
Composición: JOSUR TRATAMIENTOS DE TEXTOS, S.L.
Impreso por:

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Este libro ha sido impreso con papel y tinta ecológicos

Índice

Prólogo

Los que llevamos ya bastantes años impartiendo clases de descripción estadística de datos, también llamada estadística descriptiva, recordamos con cariño la obra de Gérard Calot, *Cours de Statistique Descriptive* (Dunod, París, 1965) que algunos conocimos ya en su versión castellana, *Curso de Estadística Descriptiva* (Paraninfo, Madrid, 1974).

Se trataba de un libro que conjugaba la precisión en el empleo de los términos estadísticos con una sencillez en la argumentación, sencillez que no estaba reñida con el rigor en las demostraciones matemáticas.

Porque, en aquellos tiempos, la estadística descriptiva no se solía enseñar en las licenciaturas de Matemáticas, pues se consideraba una derivación menor, más bien correspondiente a la Sociología, la Psicología o la Economía.

Mucho han cambiado las cosas desde entonces, y hoy día el tratamiento estadístico de la información ocupa un lugar de honor, no sólo en el campo de las aplicaciones estadísticas sino de la propia estadística matemática.

De forma paralela, ha ido cambiando el propio panorama bibliográfico, incrementándose tanto la oferta de producción nacional como (más escasamente) las traducciones de obras extranjeras, en general anglosajonas. Este incremento se ha orientado, en general, a cubrir dos lagunas. Por un lado, la inmersión de la estadística descriptiva en el seno de otras ramas del conocimiento; y por otro, la difusión de las posibilidades del software estadístico y econométrico en cuanto al tratamiento de los datos y a las derivaciones inferenciales de dicho tratamiento.

Por ello, la aparición del libro de las profesoras Castillo y Guijarro llena, sin duda alguna, un vacío bibliográfico de libros precisos en las definiciones y en su desarrollo, un libro en el que los lectores no encontrarán ni imprecisiones ni incorrecciones.

Pero la mayor innovación que se aprecia en la obra es su formato, que corresponde al de los denominados «libros de problemas». Así, las autoras no apabullan al lector (al estudiante) con una impactante y densa enumeración exhaustiva de los resultados y sus demostraciones. La presentación de los temas se realiza a través de un breve y bien organizado resumen que aborda únicamente los conceptos centrales en estudio. Las ampliaciones se presentan dentro de los problemas, a través de sucesivos ejercicios que siguen el esquema de definición-ejemplo-resultados complementarios.

Este estilo disminuye la aridez de los desarrollos, facilitando la incorporación de los estudiantes a los contenidos propuestos. Además, permite realizar diversas lecturas de los materiales, desde una más básica, que de cada tema extrae los rasgos más elementales, hasta la más sofisticada, para la que se definen conceptos más elaborados y se demuestran resultados formales de cierta complejidad, si bien ello se realiza, como se ha dicho más arriba, a través de la presentación de ejercicios que consecutivamente sitúan los conceptos como ampliaciones de materiales más elementales.

Como las profesoras indican en su presentación, los temas tratados cubren las necesidades de la docencia en descripción estadística de datos que forman parte de los programas de las asignaturas de Introducción a la Estadística de las titulaciones de Ciencias Sociales (Administración y Dirección de Empresas, Economía, Empresariales, Sociología, Relaciones laborales o Sociología, por citar las más notables). Incluso se aborda un capítulo dedicado al cálculo de probabilidades, material que las distintas programaciones docentes sitúan indistintamente al final de las disciplinas introductorias o en el inicio de las disciplinas dedicadas al estudio de las distribuciones estadísticas y de los procedimientos inferenciales.

Pero, aunque su motivación responde a las necesidades docentes en Ciencias Sociales, la posibilidad de realizar lecturas a distintos niveles hace que este libro pueda ser utilizado también para un curso semestral de Introducción a la Estadística en carreras más técnicas, como las diplomaturas o licenciaturas en Ciencias y Técnicas estadísticas o las diplomaturas en Informática de Gestión o de Sistemas, entre otras.

Cada profesional de la estadística tiene en la cabeza su libro, como proyecto o como declaración de intenciones, y no conozco dos de estos proyectos que coincidan al cien por cien. Así que no sorprenderá que eche en falta algunas cuestiones, como serían una incursión por el análisis exploratorio de datos, o un mayor desarrollo de las medidas de asociación para atributos que sigan escalas nominales u ordinales. Cierto es que ello incrementaría notablemente el volumen y (el precio) del libro, y perdería parcialmente el atractivo que posee en su versión actual.

En fin, no me cabe ninguna duda de que espera a este libro una fructífera singladura (por utilizar un símil marinero de los que tanto gustan a las autoras) de la que seremos beneficiarios docentes y profesionales de Estadística. Mi enhorabuena.

José Luis Rojo García
Catedrático de Economía aplicada
Universidad de Valladolid

Introducción

La obra que presentamos a continuación contiene las nociones fundamentales de estadística descriptiva, así como los conceptos introductorios de cálculo de probabilidades.

La estructura del trabajo permite entender los contenidos de la materia como un todo, en el cual teoría y práctica son indivisibles: no es un libro de teoría —aunque al inicio de cada capítulo haya una presentación de los principales conceptos y resultados—, tampoco un libro de ejercicios —aunque tenga más de 250 problemas resueltos y comentados—, es un libro de estadística descriptiva e introducción al cálculo de probabilidades. Este hecho es fundamental, si se tiene en cuenta que el alumno tiende a rechazar los aspectos teóricos de las disciplinas de naturaleza matemática, y a pensar que «no tienen relación» con las aplicaciones prácticas. Con este libro pretendemos ayudar a desmontar estas expectativas.

En la obra, por tanto, no sólo se enseña la herramienta estadística, sino que, prioritariamente, se muestra el modo de utilizarla. En la actualidad, con la generalización del uso de programas informáticos, el empleo de procedimientos estadísticos puede ser peligroso si se desconoce cómo, cuándo, dónde y por qué hay que aplicarlos; así, el libro consta de problemas sencillos que introducen en el conocimiento de las técnicas, y de otros, basados en la realidad que se pretende analizar, que permiten aprender los conceptos presentados.

A pesar de que este texto hará posible el aprendizaje individualizado de cualquier lector con cierta madurez, pues se describe y analiza cada concepto de manera sencilla, la claridad en la exposición no está exenta de rigor: un rigor que hemos procurado no sólo en los aspectos más teóricos, sino también en la elección de los supuestos prácticos que ayudarán al lector a interpretar la realidad en términos estadísticos.

En el primer capítulo se estudian las distribuciones de frecuencias unidimensionales, desde la presentación y representación de las mismas, hasta el análisis de sus principales medidas de resumen (posición, dispersión, forma y concentración). El capítulo segundo versa sobre las dis-

tribuciones de frecuencias bidimensionales con especial empeño en el análisis de las distribuciones de frecuencias condicionadas y en el estudio de la regresión y la correlación entre variables.

El análisis estadístico de atributos es el objetivo del tercer capítulo, estando una gran parte del mismo dedicado a la asociación entre caracteres.

Los capítulos cuarto y quinto desarrollan, respectivamente, números índices (índices simples y compuestos, cambio de base de series de índices, deflación de series estadísticas, etc.) y tasas de variación (absolutas, relativas y acumulativas), y análisis clásico de series temporales (descripción de sus componentes), conceptos clave para el conocimiento de la evolución de una variable a través del tiempo.

Por último, en el capítulo sexto, se realiza una introducción al cálculo de probabilidades, partiendo de la definición axiomática de probabilidad que permitirá, utilizando el concepto de probabilidad condicionada y los teoremas de la probabilidad total y de Bayes, la obtención de probabilidades de sucesos referidos a experimentos simples y compuestos.

Dada su naturaleza, la originalidad de la obra no reside en los contenidos de la misma, sino en el modo en que estos son presentados para que su enseñanza resulte lo más atractiva posible al lector. Por nuestra parte, deseamos contribuir a la consecución de este objetivo.

Santander, marzo de 2005

Capítulo 1

Distribuciones de frecuencias unidimensionales

Principales conceptos y resultados

Se denomina **población**[1] a un conjunto de unidades, siendo una **variable** cualquier característica numérica de las unidades de la población.

De la observación de una variable en las unidades de la población se obtienen **datos** u **observaciones** que constituyen una **estadística primaria**. Cada observación *distinta* de una variable es un **valor,** denotándose por $x_1, \ldots, x_h$ los h valores de una variable X, que supondremos ordenados de menor a mayor, siendo x_i el valor genérico.

La **frecuencia absoluta** de un valor de una variable es el número de observaciones iguales a dicho valor o, equivalentemente, el número de unidades de la población que tienen ese valor de la variable. Se denota por n_i la frecuencia absoluta genérica, esto es, la frecuencia absoluta correspondiente al valor x_i. Si N es el número total de datos se tiene:

$$\sum_{i=1}^{h} n_i = N.$$

La **frecuencia relativa** de un valor de una variable es la proporción de observaciones iguales a dicho valor. Se denota por f_i la frecuencia relativa del valor x_i. Teniendo en cuenta que, por definición, que

$$f_i = \frac{n_i}{N},$$

resulta, entonces,

$$\sum_{i=1}^{h} f_i = 1.$$

[1] Esta denominación es debida a que dicho concepto fue estudiado por primera vez en Demografía.

La **frecuencia absoluta acumulada** de un valor de una variable[2] es el número de observaciones menores o iguales a dicho valor. Se denota por N_i la frecuencia absoluta acumulada del valor x_i[3]. Se verifica que

$$N_1 = n_1 \text{ y } N_i = n_1 + \ldots + n_i, \text{ para } i = 2, ..., h.$$

La **frecuencia relativa acumulada** de un valor de una variable es la proporción de observaciones menores o iguales a dicho valor. Denotaremos por F_i la frecuencia relativa acumulada genérica[4]. Se cumple que

$$F_i = \frac{N_i}{N}$$

y, además,

$$F_1 = f_1 \text{ y } F_i = f_1 + \ldots + f_i, \text{ para } i = 2, ..., h.$$

En la siguiente tabla se resumen los conceptos definidos:

	Frecuencias ordinarias		Frecuencias acumuladas	
	Absoluta	Relativa	Absoluta	Relativa
x_1	n_1	$f_1 = n_1/N$	$N_1 = n_1$	$F_1 = f_1$
⋮	⋮	⋮	⋮	⋮
x_i	n_i	$f_i = n_i/N$	$N_i = n_1 + \ldots + n_i$	$F_i = f_1 + \ldots + f_i$
⋮	⋮	⋮	⋮	⋮
x_h	n_h	$f_h = n_h/N$	$N_h = n_1 + \ldots + n_h = N$	$F_h = f_1 + \ldots + f_h = 1$

Una **distribución de frecuencias** elaborada a partir de una estadística primaria es la relación de los valores de una variable junto con sus correspondientes frecuencias. Una distribución de frecuencias se denota mediante el par $(x_i; n_i)$ o bien $(x_i; f_i)$, según se utilicen frecuencias absolutas o relativas[5].

Una distribución de frecuencias es **unitaria**, si todas las frecuencias absolutas son iguales a la unidad.

Llamaremos **valores de la distribución** a todas las observaciones de la variable en las unidades de la población.

Dos variables tienen la misma distribución de frecuencias si coinciden sus valores y sus correspondientes frecuencias relativas.

---

[2] Algunos autores dan una definición más general de este tipo de frecuencias, al considerar la frecuencia absoluta acumulada asociada a cualquier número (no necesariamente a un valor de la variable).

[3] Nótese que N_h es igual a N.

[4] Nótese que F_h es igual a 1.

[5] Pueden considerarse frecuencias *ordinarias* o acumuladas.

Cuando el número de valores de una variable es muy grande puede resultar aconsejable *agrupar* los valores en intervalos o clases. Los *extremos inferior* y *superior* del intervalo genérico se denotan, respectivamente, por L_{i-1} y L_i, siendo c_i la *amplitud* del intervalo, esto es, $c_i = L_i - L_{i-1}$. Cada intervalo está representado por la **marca de clase** o punto medio del mismo; así, $x_i = (L_{i-1} + L_i)/2$ es la marca de clase del intervalo[6] $L_{i-1} - L_i$.

En el caso de variables con valores agrupados, la definición de cada uno de los tipos de frecuencias es análoga a la realizada cuando los valores de la variable no están agrupados, sustituyendo valor por intervalo. Téngase en cuenta que, en el caso de frecuencias acumuladas (absolutas o relativas), hablaremos de observaciones menores o iguales que el extremo superior del intervalo considerado.

Se dispone, entonces, de una distribución de frecuencias[7] **agrupada en intervalos** que denotaremos por $(L_{i-1} - L_i; n_i)$ o bien por $(L_{i-1} - L_i; f_i)$, según el tipo de frecuencia utilizada.

Si sobre la distribución de frecuencias $(x_i; n_i)$ realizamos una *transformación lineal* consistente en multiplicar a todos los valores de la distribución por una constante a y sumar una constante b al resultado (a y b números reales), se tiene la distribución de frecuencias transformada, $(a \cdot x_i + b; n_i)$.

Un caso particular de transformación lineal cuando $a = 1/e$ y $b = -o/e$, (e y o números reales, $e > 0$) es el **cambio de origen y de escala**, con el cual se obtiene la distribución de frecuencias transformada $\left(\dfrac{x_i - o}{e}; n_i\right)$.

Mediante el **diagrama de barras** se representan las distribuciones de frecuencias de variables con valores sin agrupar. La longitud de cada barra sobre el correspondiente valor de la variable es igual a su frecuencia (absoluta o relativa).

Para las distribuciones de frecuencias agrupadas en intervalos, el **histograma de frecuencias** es la representación más adecuada. En él, el área del rectángulo que se eleva sobre el intervalo es igual a su frecuencia (absoluta o relativa)[8]. Se denomina **densidad de frecuencia** de un intervalo, d_i, a la altura del correspondiente rectángulo: $d_i = n_i/c_i$, o bien $d_i = f_i/c_i$ según las frecuencias empleadas sean absolutas o relativas.

Los **polígonos de frecuencias acumuladas** se construyen elevando sobre el extremo superior de cada intervalo una altura igual a su frecuencia acumulada (absoluta o relativa) y uniendo el final de cada altura.

---

[6] Advertimos al lector de la diferencia entre el intervalo $L_{i-1} - L_i$, donde el guión separa el extremo inferior del extremo superior, y la amplitud del intervalo $L_i - L_{i-1}$, donde el guión es el símbolo de la sustracción.

[7] Nótese que la agrupación en clases conlleva una pérdida de información. Consecuentemente, el número de clases debe ser lo suficientemente grande como para no perder demasiada información, pero no excesivo, con el fin de aprovechar las ventajas del agrupamiento.

[8] En el caso de una agrupación en clases de igual amplitud, las alturas de los rectángulos pueden ser iguales a las correspondientes frecuencias, siendo, entonces, cada área *proporcional* a la frecuencia.

Hay una serie de medidas que informan sobre los aspectos fundamentales de las distribuciones de frecuencias de una variable.

En este sentido, las **medidas de posición** *sitúan* la distribución, es decir, indican en torno a qué valor están las observaciones de la variable. Una medida de posición actúa como medida de resumen de la información contenida en los datos.

Una de las medidas de posición más utilizada es la **media aritmética**. Se define como la suma de todas las observaciones de una variable dividida entre el número de ellas. La media aritmética de la variable X, cuya distribución de frecuencias es $(x_i; n_i)$, media aritmética de la distribución de frecuencias $(x_i; n_i)$ o, simplemente, media de X es, por consiguiente,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_i = \sum_{i=1}^{h} x_i \cdot f_i.$$

Dada su definición, la media aritmética es muy sensible a los valores extremos de la variable.

La media aritmética de las desviaciones de los valores de la distribución con respecto a su media aritmética es igual a cero:

$$\sum_{i=1}^{h} (x_i - \bar{x}) f_i = 0.$$

La media aritmética de una distribución se ve afectada por transformaciones lineales y, por tanto, por cambios de origen y de escala en los valores de la distribución. Así, dada la distribución de frecuencias $(x_i; n_i)$, cuya media es $\bar{x}$, la media de la distribución transformada, $(a \cdot x_i + b; n_i)$, (a y b números reales) es $a \cdot \bar{x} + b$. En particular, si $a = 1/e$ y $b = -o/e$ y, (e y o números reales, $e > 0$), es decir, si la transformación lineal es un cambio de origen y de escala, entonces, la media de la distribución transformada es $(\bar{x} - o)/e$.

Para promediar índices y tasas se utiliza la **media geométrica**, raíz N-ésima del producto de las N observaciones de una variable:

$$G = \sqrt[N]{\prod_{i=1}^{h} x_i^{n_i}} = \prod_{i=1}^{h} x_i^{f_i}.$$

La **media armónica** de una distribución de frecuencias $(x_i; n_i)$, que se emplea para promediar magnitudes relativas, se define como el inverso de la media aritmética de la variable inversa, es decir, el inverso de la media aritmética de la distribución $(1/x_i; n_i)$:

$$H = \frac{1}{\frac{1}{N} \sum_{i=1}^{h} \frac{1}{x_i} \cdot n_i} = \frac{N}{\sum_{i=1}^{h} \frac{n_i}{x_i}} = \frac{1}{\sum_{i=1}^{h} \frac{1}{x_i} \cdot f_i}.$$

La **mediana** de una distribución de frecuencias es el número que, supuesta una ordenación creciente de los datos, tiene a su derecha y a su izquierda el mismo número de observaciones. Al no tener en cuenta la magnitud de los valores de la variable, su cálculo resulta adecuado en aquellas distribuciones con valores extremos.

Para calcular la mediana en distribuciones no agrupadas en intervalos, se siguen los siguientes pasos:

- Se obtiene el valor $N/2$.
- Se calcula la frecuencia absoluta acumulada, N_i, de cada valor x_i.
- Si existe un valor x_i tal que $N_i = N/2$ —hecho que sólo puede darse cuando N es un número par—, la mediana es la media aritmética de los dos valores centrales de la distribución:

$$Me = \frac{x_i + x_{i+1}}{2}.$$

- Si no existe un valor x_i tal que $N_i = N/2$, la mediana se define como el mínimo valor x_i tal que N_i es mayor que $N/2$.

En el caso de distribuciones de frecuencias agrupadas en intervalos la mediana responde a la expresión:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i,$$

donde L_{i-1} y c_i son, respectivamente, el extremo inferior y la amplitud del **intervalo mediano**, esto es, del intervalo que ocupa la posición central[9].

La **moda** de una distribución de frecuencias es el valor con mayor frecuencia[10]. En distribuciones agrupadas en intervalos, la moda se calcula como

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i,$$

donde c_i es la amplitud del **intervalo modal** —intervalo con mayor densidad de frecuencia[11]— y d_i es la densidad de frecuencia de dicho intervalo[12].

Los **cuantiles** son medidas de posición que dividen el conjunto de observaciones de una variable en clases, conteniendo cada una de ellas una cierta proporción de observaciones. Denotaremos

---

[9] Para calcular el intervalo mediano se obtiene la frecuencia absoluta acumulada de cada intervalo. Si existe un intervalo cuya frecuencia absoluta acumulada, N_i, es igual a $N/2$, éste es el intervalo mediano, siendo la mediana el extremo superior del intervalo, como puede comprobarse sustituyendo en la fórmula de esta medida de posición. Si no existe un intervalo verificando tal condición, el intervalo mediano es el *primer* intervalo cuya frecuencia absoluta acumulada es estrictamente mayor que $N/2$.

[10] Una distribución de frecuencias puede tener más de una moda cuando haya más de un valor con la máxima frecuencia.

[11] Cuando el intervalo modal es el primero (último), la moda es el extremo superior (inferior) del intervalo.

[12] Si los intervalos son de igual amplitud, puede sustituirse la densidad de frecuencia por la frecuencia correspondiente, tanto en la definición de intervalo modal como en la expresión de la moda.

por x_q el cuantil de orden q, valor al que corresponde una proporción q de observaciones menores o iguales a él. En particular, los **cuartiles**, C_1, C_2 y C_3, dividen la estadística en cuatro *partes* iguales; los **deciles**, D_1, …, D_9, en diez *partes* iguales y los **percentiles**, P_1, …, P_{99}, en cien *partes* iguales[13].

Para distribuciones agrupadas en intervalos, el cuantil de orden q responde a la expresión:

$$x_q = L_{i-1} + \frac{q \cdot N - N_{i-1}}{n_i} \cdot c_i,$$

donde L_{i-1} y c_i son, respectivamente, el extremo inferior y la amplitud del **intervalo cuantílico**[14].

Para medir la *representatividad* de las medidas de posición se emplean las **medidas de dispersión**. Las medidas de dispersión miden el grado de *alejamiento* de las observaciones con respecto a su promedio y, por tanto, el grado de *variabilidad* de los datos.

La **varianza** es una medida de dispersión que acompaña a la media aritmética y, a partir de la distribución de frecuencias $(x_i; n_i)$, se calcula como

$$S^2 = \frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^2 n_i = \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i,$$

con lo cual es la media aritmética de las desviaciones al cuadrado entre las observaciones y su media aritmética. Cuanto mayor sea la varianza, mayor será la dispersión de los datos respecto a la media aritmética, mayor la variabilidad de las observaciones y menor la representatividad del promedio.

Si a todos los valores de la distribución se les suma una constante, la varianza permanece inalterable. Por el contrario, si todas las observaciones se multiplican por una constante la varianza resulta multiplicada por dicha constante al cuadrado.

La **desviación típica**, raíz cuadrada de la varianza, es

$$S = \sqrt{\frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^2 n_i} = \sqrt{\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i}.$$

Dada una distribución de frecuencias $(x_i; n_i)$, se denomina **distribución tipificada** a la que se obtiene restando a cada valor de la distribución su media y dividiendo el resultado por su desviación típica, esto es, a la distribución $\left(\frac{x_i - \bar{x}}{S}; n_i\right)$[15].

---

[13] Nótese que la mediana es un cuantil, pues divide la estadística en dos *partes* iguales.

[14] Para calcular el intervalo cuantílico, se obtiene la frecuencia absoluta acumulada de cada intervalo. Si hay un intervalo tal que su N_i sea igual a $q \cdot N$, tendremos el intervalo cuantílico; en caso contrario, se toma el primer intervalo cuya frecuencia absoluta acumulada sea estrictamente mayor que $q \cdot N$.

[15] Se trata de un cambio de origen y de escala donde $o = \bar{x}$ y $e = S$.

La varianza es un caso particular de la **desviación cuadrática media con respecto a un promedio**, P, que, dada una distribución de frecuencias $(x_i; n_i)$, se define como

$$D_P^2 = \frac{1}{N}\sum_{i=1}^{h}(x_i - P)^2\, n_i = \sum_{i=1}^{h}(x_i - P)^2 f_i.$$

Otra medida de dispersión es la **desviación absoluta media con respecto a un promedio**, P, que, para una distribución de frecuencias $(x_i; n_i)$, es

$$d_P = \frac{1}{N}\sum_{i=1}^{h}|x_i - P| \cdot n_i = \sum_{i=1}^{h}|x_i - P| \cdot f_i.$$

El **coeficiente de variación respecto a un promedio**, P, es una medida de dispersión *relativa* que permite comparar variabilidades de diferentes distribuciones; además, sirve para discriminar entre promedios de una distribución. Dada una distribución de frecuencias $(x_i; n_i)$, se define[16] como

$$\mathrm{V}_P = \frac{\sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{h}(x_i - P)^2\, n_i}}{P} = \frac{\sqrt{\displaystyle\sum_{i=1}^{h}(x_i - P)^2 f_i}}{P}.$$

Este coeficiente se interpreta en valor absoluto: cuanto mayor sea el coeficiente de variación, mayor será la variabilidad de la distribución y, recíprocamente, cuanto menor sea el coeficiente, menor la dispersión.

Cuando el promedio es la media aritmética se obtiene el **coeficiente de variación de Pearson**:

$$\mathrm{V} = \frac{S}{\overline{x}}.$$

El **índice de dispersión respecto a un promedio**, P, es, también, una medida de dispersión relativa. Dada una distribución de frecuencias $(x_i; n_i)$, se define[17] como

$$I_P = \frac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{h}|x_i - P| \cdot n_i}{P} = \frac{\displaystyle\sum_{i=1}^{h}|x_i - P| \cdot f_i}{P}.$$

Las medidas de resumen de la información proporcionada por los datos se basan en ciertas características halladas a partir de los valores de la distribución. Estas características, denominadas **momentos**, son herramientas útiles para muchos cálculos.

---

[16] Este coeficiente solamente está definido cuando P es distinto de cero.
[17] Véase nota anterior.

Dada una distribución de frecuencias $(x_i; n_i)$, el **momento de orden *r* respecto al origen** o **momento no central de orden *r*** de la distribución es

$$a_r = \frac{1}{N} \sum_{i=1}^{h} x_i^r \cdot n_i = \sum_{i=1}^{h} x_i^r \cdot f_i.$$

Obsérvese que

$$a_1 = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_i = \sum_{i=1}^{h} x_i \cdot f_i$$

es la media aritmética de la distribución.

El **momento de orden *r* respecto a la media aritmética** o **momento central de orden *r*** de la distribución de frecuencias $(x_i; n_i)$ es

$$m_r = \frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^r n_i = \sum_{i=1}^{h} (x_i - \bar{x})^r f_i.$$

Nótese que la varianza, S^2, es el momento central de orden dos:

$$m_2 = \frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^2 n_i = \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i.$$

Dos son los aspectos fundamentales en el estudio de la *forma* de una distribución: su **grado de simetría** y su **grado de apuntamiento** o **curtosis**.

El coeficiente de asimetría más utilizado es el **coeficiente de Fisher**, que, para una distribución de frecuencias $(x_i; n_i)$, es

$$g_1 = \frac{m_3}{S^3} = \frac{\frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^3 n_i}{\left[\frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^2 n_i \right]^{3/2}} = \frac{\sum_{i=1}^{h} (x_i - \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i \right]^{3/2}}.$$

Si la distribución es simétrica, esto es, cuando a la derecha y a la izquierda de su media aritmética existe el mismo número de valores de la variable, a la misma distancia de la media y con la misma frecuencia, este coeficiente es nulo, siendo positivo o negativo si la distribución es *asimétrica positiva* o *asimétrica negativa*, respectivamente[18].

[18] Nótese que el numerador de este coeficiente es el promedio de las desviaciones al cubo de las observaciones con respecto a su media aritmética, y que dicho promedio es igual a cero en el caso de que exista simetría, puesto que entonces habrá el mismo número de observaciones a la derecha que a la izquierda de la media. Además, como el denominador de este coeficiente es una potencia de la desviación típica, siempre positiva, el signo del coeficiente de asimetría depende del numerador, positivo en el caso de asimetría positiva (más desviaciones con respecto a la media positivas que negativas) y negativo en caso de asimetría negativa (más desviaciones negativas que positivas).

Para estudiar el grado de curtosis de una distribución de frecuencias $(x_i; n_i)$, se emplea el coeficiente

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{h}(x_i - \bar{x})^4 n_i}{\left[\frac{1}{N}\sum_{i=1}^{h}(x_i - \bar{x})^2 n_i\right]^2} - 3 = \frac{\sum_{i=1}^{h}(x_i - \bar{x})^4 f_i}{\left[\sum_{i=1}^{h}(x_i - \bar{x})^2 f_i\right]^2} - 3.$$

Este coeficiente, que se estudia en distribuciones de frecuencias con aspecto *acampanado*, es nulo cuando la distribución tiene el mismo grado de apuntamiento que la distribución patrón[19] **(mesocúrtica)**; mayor que cero cuando es más apuntada que el perfil de la distribución patrón **(leptocúrtica)**; y, por último, menor que cero cuando es menos apuntada que el perfil de dicha distribución **(platicúrtica)**[20].

Las **medidas de desigualdad** o **concentración** sintetizan el grado de equidad en el reparto de las observaciones de la variable.

Denominando p_i al porcentaje de individuos con renta menor o igual que[21] x_i, esto es,

$$p_i = \frac{N_i}{N} \cdot 100,$$

donde N_i es la frecuencia absoluta acumulada del valor x_i, y q_i al porcentaje de renta percibida por los individuos con renta menor o igual que x_i, es decir,

$$q_i = \frac{x_1 \cdot n_1 + \ldots + x_i \cdot n_i}{x_1 \cdot n_1 + \ldots + x_h \cdot n_h} \cdot 100 = \frac{u_i}{u_h} \cdot 100,$$

donde u_i es la renta percibida por los individuos con renta menor o igual que x_i y u_h es el total de renta, se obtienen los pares de puntos (p_i, q_i) que, representados en un cuadrado de lado 100, determinan una poligonal llamada **curva de Lorenz**.

La curva de Lorenz refleja cómo se reparte el total de recursos entre el total de individuos que perciben dichos recursos.

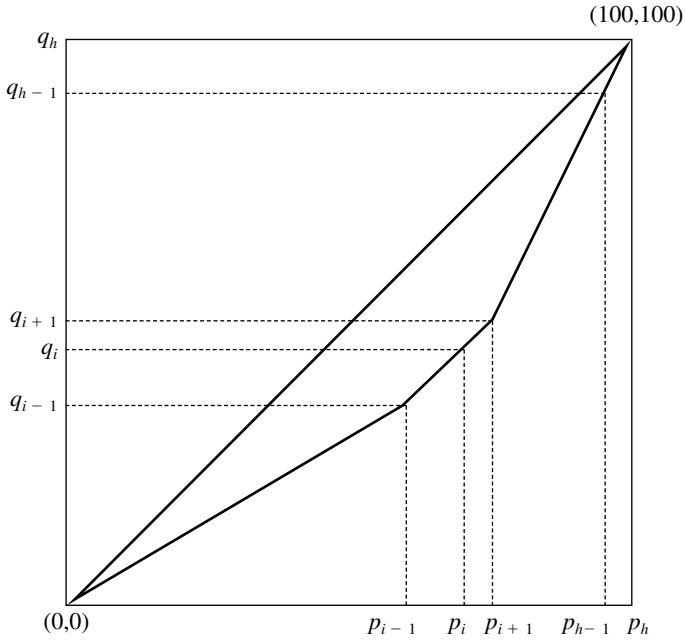
Si la curva coincide con la diagonal del cuadrado, la *concentración es mínima*, es decir, existe máxima equidad en el reparto de los valores de la distribución. Por el contrario, cuando la curva coincide con los lados del cuadrado, la *concentración es máxima* y el grado de equidad en el reparto es, en consecuencia, mínimo.

---

[19] La distribución patrón corresponde a la denominada *distribución normal* cuyo perfil es la llamada *campana de Gauss.*

[20] La definición e interpretación de este coeficiente se basa en que, para la distribución normal, se cumple que el numerador es tres veces el denominador, es decir, para la distribución normal, perfil patrón, g_2 es igual a cero.

[21] Generalmente el estudio de la concentración se realiza sobre variables como la renta o el salario.



La idea ilustrada mediante la curva de Lorenz se concreta con el **índice de Gini**. El índice de Gini se define como el cociente entre el *área de concentración*, esto es, el área entre la diagonal del cuadrado y la curva de Lorenz, y el área del triángulo que hay bajo la diagonal.

Del cálculo geométrico de estas áreas se obtiene la expresión del índice de Gini:

$$I_G = 1 - \sum_{i=0}^{h-1} \frac{(q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{10\,000}$$

o, equivalentemente,

$$I_G = 1 - \sum_{i=0}^{h-1} \frac{q_{i+1} + q_i}{100} \cdot f_{i+1}.$$

El índice de Gini se interpreta, por tanto, como la proporción que el área de concentración representa sobre el área del triángulo[22]. Así, cuando el índice es cero —curva igual a la diagonal del cuadrado—, la concentración es mínima; cuando el índice es uno —curva coincidente con los lados del triángulo—, la concentración es máxima.

[22] Expresiones alternativas del índice de Gini, que el lector puede encontrar en otros textos, se obtienen, en realidad, como aproximaciones al cálculo de las áreas de la curva de concentración y del triángulo.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

1.1 Debido a la falta de personal, los trabajadores de la empresa Superporte, dedicada al servicio de mensajería, realizaron horas extraordinarias durante el pasado ejercicio. Las horas extras realizadas por los 100 trabajadores de la empresa fueron:

35	40	20	55	20	35	55	60	50	50
35	50	35	60	30	40	60	40	65	65
35	55	35	65	35	50	65	65	35	50
40	60	40	40	40	35	55	65	30	50
50	55	50	55	50	20	55	75	40	60
55	60	55	55	40	40	65	65	35	50
60	55	60	60	20	55	20	70	55	65
65	60	55	30	50	30	75	20	55	20
70	55	70	55	60	30	50	65	30	50
75	60	70	55	30	50	30	65	40	60

a) ¿Qué población se ha considerado? ¿Por cuántas unidades está constituida? ¿A qué variable corresponden estos datos?

b) Obténgase la distribución de frecuencias de la variable.

c) Represéntese gráficamente mediante un diagrama de barras, la distribución obtenida en el apartado anterior.

SOLUCIÓN

a) La población está constituida por los 100 trabajadores de la empresa Superporte, sobre la que se ha observado la variable *número de horas extraordinarias*.

b) La distribución de frecuencias $(x_i; n_i)$, donde x_i es el valor genérico de la variable *número de horas extraordinarias*, X, y n_i la frecuencia absoluta genérica, esto es, el número de observaciones que tienen un valor de la variable igual a x_i, se recoge en la tabla siguiente:

Horas x_i	N.º trabajadores n_i
20	7
30	8
35	10
40	11
50	14
55	18
60	13
65	12
70	4
75	3
	$N = 100$

Como puede observarse, en la primera columna de la tabla aparecen los 10 *valores* de la variable X, ordenados de menor a mayor, y, en la segunda columna, las frecuencias absolutas que se obtienen contando el número de observaciones que son iguales a cada valor de la variable. Así, por ejemplo, dado el valor $x_8 = 65$, su frecuencia absoluta es $n_8 = 12$, lo cual significa que hay 12 trabajadores que han realizado 65 horas extras durante el pasado ejercicio.

La tabla anterior puede completarse con las frecuencias relativas y con las frecuencias acumuladas, tanto absolutas como relativas:

Valor x_i	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa acumulada F_i
20	7	0,07	7	0,07
30	8	0,08	15	0,15
35	10	0,10	25	0,25
40	11	0,11	36	0,36
50	14	0,14	50	0,50
55	18	0,18	68	0,68
60	13	0,13	81	0,81
65	12	0,12	93	0,93
70	4	0,04	97	0,97
75	3	0,03	100	1
	$N = 100$	1		

La frecuencia relativa genérica, proporción de trabajadores que realizó un número de horas extraordinarias igual a x_i, responde a la expresión:

$$f_i = \frac{n_i}{N},$$

con lo cual, por ejemplo,

$$f_2 = \frac{8}{100} = 0{,}08,$$

frecuencia relativa del valor $x_2 = 30$, indica que el 8 por ciento de los trabajadores realizó 30 horas extras.

Las relaciones de las frecuencias absolutas acumuladas en función de las frecuencias absolutas,

$$N_1 = n_1 \text{ y } N_i = n_1 + \ldots + n_i, \text{ para } i = 2, ..., h,$$

expresan que la frecuencia genérica es el número de trabajadores que realizó a lo sumo x_i horas extraordinarias. De este modo, por ejemplo, la frecuencia absoluta acumulada del

valor $x_3 = 35$, esto es, $N_3 = 25$, indica que 25 trabajadores realizaron como máximo 35 horas extras.

También pueden calcularse las frecuencias absolutas acumuladas de modo sucesivo, cada una a partir de la anterior:

$$N_1 = n_1$$

y, para el resto de valores de la variable,

$$N_i = N_{i-1} + n_i.$$

Mediante este tipo de frecuencias podemos hallar, por ejemplo, el número de trabajadores que realizaron más de 60 horas extras:

$$N - N_7 = 100 - 81 = 19.$$

Obsérvese que podríamos haber llegado a idéntico resultado empleando frecuencias absolutas ordinarias:

$$n_8 + n_9 + n_{10} = 12 + 4 + 3 = 19.$$

Por último, las frecuencias relativas acumuladas recogidas en la última columna de la tabla anterior se obtienen como

$$F_i = \frac{N_i}{N},$$

siendo esta frecuencia genérica la proporción de trabajadores que realizó como máximo x_i horas extraordinarias.

Otra posibilidad de cálculo de este tipo de frecuencias es, al igual que en el caso de las frecuencias absolutas acumuladas, de modo encadenado, obteniendo cada una a partir de la anterior según las relaciones:

$$F_1 = f_1$$

y, para los siguientes valores de la variable,

$$F_i = F_{i-1} + f_i,$$

ya que, por un lado,

$$F_1 = \frac{N_1}{N} = \frac{n_1}{N} = f_1$$

y, por otro lado,

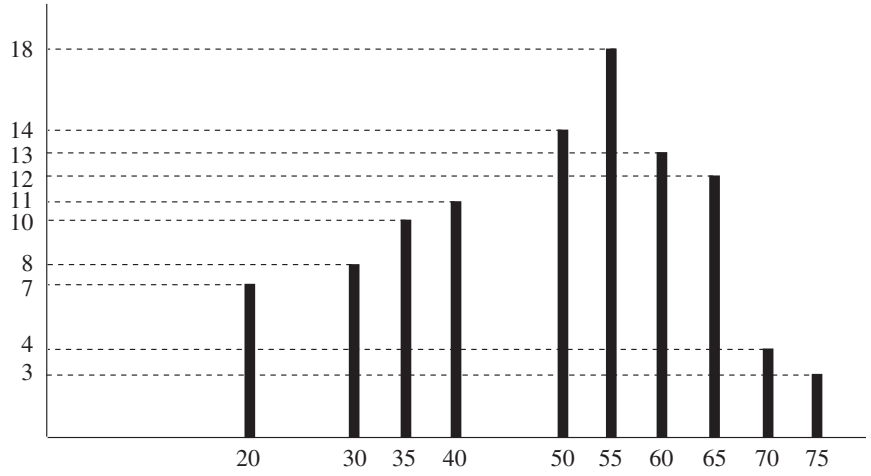
$$F_i = \frac{N_i}{N} = \frac{N_{i-1}}{N} + \frac{n_i}{N} = F_{i-1} + f_i.$$

Las frecuencias relativas acumuladas permiten responder a preguntas del tipo: ¿qué porcentaje de trabajadores realizó menos de 50 horas extras? Tal porcentaje se corresponde con la frecuencia relativa acumulada del valor $x_4 = 40$:

$$F_4 = 0{,}36,$$

es decir, el 36 por ciento.

c) Colocando en el eje horizontal, o eje de abscisas, los valores de la variable X y en el eje vertical, o eje de ordenadas, las respectivas frecuencias absolutas, basta con elevar sobre cada x_i una altura igual a la frecuencia n_i para obtener la siguiente gráfica, correspondiente al diagrama de barras de la distribución de frecuencias $(x_i; n_i)$.



Proponemos al lector la representación del diagrama de barras, considerando frecuencias relativas en lugar de absolutas.

1.2 Durante el pasado verano el Club del Lector, empresa dedicada a la venta de libros a domicilio, contrató a 200 estudiantes en todo el territorio nacional con objeto de captar nuevos socios. El número de suscripciones que realizaron estos 200 estudiantes fueron:

0	2	3	4	5	10	6	7	8	9	32	31	34	40	40	31	32	33	40	34
11	12	13	14	15	16	17	18	19	20	32	32	35	40	32	35	37	39	40	32
11	12	13	14	15	16	17	18	19	20	31	31	34	31	32	34	31	35	32	31
11	12	13	14	15	21	22	23	24	25	35	30	32	31	35	32	35	33	35	35
26	28	27	29	23	24	25	26	27	26	37	36	33	36	35	36	37	36	34	38
27	28	29	21	22	26	28	29	28	28	40	34	33	40	33	32	38	39	31	40
41	50	50	50	42	50	42	45	42	42	60	60	60	60	59	70	61	61	65	70
23	24	21	22	23	26	27	28	29	30	43	43	42	42	41	52	52	55	55	55
21	22	23	24	31	36	36	38	31	39	1	2	3	4	5	10	6	7	8	9
32	33	34	35	36	33	33	33	32	31	21	22	23	24	25	26	28	27	21	22

a) ¿Cuál es la población objeto de estudio? ¿Cuántas unidades tiene dicha población? ¿A qué variable corresponden las observaciones de esta estadística primaria?

b) Hállese una distribución de frecuencias con valores agrupados de la variable considerada en intervalos de igual amplitud.

c) Represéntese un histograma de frecuencias de la distribución obtenida en el apartado anterior.

SOLUCIÓN

a) La población está formada por los estudiantes contratados por el Club del Lector. Sobre las 200 unidades de esta población se ha observado la variable *número de socios captados*.

b) Es importante tener en cuenta que en este apartado no se pide *la* distribución de frecuencias, sino *una* distribución de frecuencias, ya que podríamos hallar tantas distribuciones de frecuencias como clases de igual amplitud podamos hacer a partir de la estadística primaria, considerando como extremo inferior del primer intervalo el mínimo valor de la variable, que en este caso es 0, y como extremo superior del último intervalo el máximo valor que, como puede comprobarse, es 70.

De este modo, si, por ejemplo, tomamos intervalos de longitud 10, una posible distribución de frecuencias con datos agrupados en clases de igual amplitud es la que se recoge en la siguiente tabla, donde $L_{i-1} - L_i$ no contiene los datos iguales a L_{i-1}.

N.º socios $L_{i-1} - L_i$	N.º trabajadores n_i
0-10	20
10-20	25
20-30	50
30-40	75
40-50	15
50-60	10
60-70	5
	$N = 200$

La segunda columna corresponde a las frecuencias absolutas que se obtienen contando el número de observaciones que pertenecen a cada uno de los siete intervalos, de amplitud 10, en los que hemos agrupado los valores de la distribución.

En la siguiente tabla, que completa la anterior, aparecen las frecuencias relativas, así como las frecuencias acumuladas, absolutas y relativas.

$L_{i-1}-L_i$	Frecuencia absoluta n_i	Frecuencia relativa f_i	Frecuencia absoluta acumulada N_i	Frecuencia relativa acumulada F_i
0-10	20	0,100	20	0,100
10-20	25	0,125	45	0,225
20-30	50	0,250	95	0,475
30-40	75	0,375	170	0,850
40-50	15	0,075	185	0,925
50-60	10	0,050	195	0,975
60-70	5	0,025	200	1
	$N = 200$			

Las frecuencias relativas, de expresión genérica $f_i = n_i/N$, proporción de estudiantes que realizaron un número de suscripciones comprendido entre L_{i-1} y L_i, aparecen en la tercera columna de la tabla. Por ejemplo,

$$f_3 = \frac{50}{200} = 0{,}250$$

es la frecuencia relativa del intervalo 20-30, lo cual supone que el 25 por ciento de los estudiantes captaron un número de socios comprendido entre 20 y 30.

En la cuarta columna de la tabla se recogen las frecuencias absolutas acumuladas, siendo la frecuencia genérica, $N_i = n_1 + \ldots + n_i$, el número de estudiantes que realizaron un número de suscripciones menor o igual que L_i.

Así, por ejemplo, la frecuencia absoluta acumulada del intervalo 40-50 es

$$N_5 = n_1 + n_2 + n_3 + n_4 + n_5 = 20 + 25 + 50 + 75 + 15 = 185,$$

y representa el número de estudiantes que captaron una cantidad de socios menor o igual que 50.

Las frecuencias absolutas acumuladas permiten plantearnos, por ejemplo, cuántos estudiantes realizaron más de 40 suscripciones, cantidad que podemos hallar como

$$N - N_4 = 200 - 170 = 30,$$

o bien como

$$n_5 + n_6 + n_7 = 15 + 10 + 5 = 30.$$

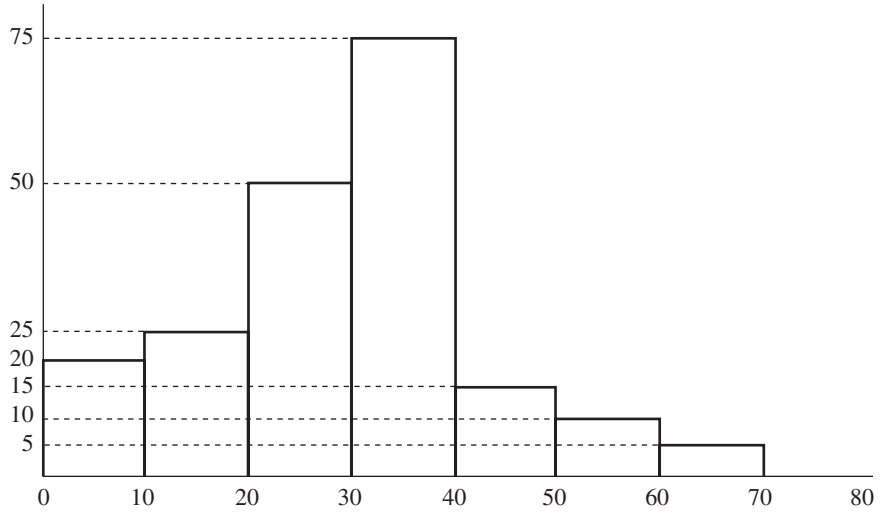
La última columna de la tabla contiene las frecuencias relativas acumuladas, $F_i = N_i/N$, expresión genérica de la proporción de estudiantes que captó un número de clientes como máximo igual a L_i.

La frecuencia relativa acumulada del intervalo 10-20,

$$F_2 = \frac{45}{200} = 0{,}225,$$

que también puede hallarse como $f_1 + f_2$, indica que el 22,5 por ciento de los estudiantes han realizado a lo sumo 20 suscripciones.

c) En el eje de abscisas, o eje horizontal, colocamos los intervalos en los que hemos agrupado los valores de la variable *X*. En el eje de ordenadas, o eje vertical, las frecuencias absolutas, ya que, al tener los intervalos igual amplitud, podemos prescindir de las densidades de frecuencia. De este modo, dibujamos rectángulos cuyas áreas, *proporcionales* a las frecuencias, conforman el histograma que aparece en la gráfica siguiente.



1.3 Represéntese gráficamente la distribución de los salarios mensuales, en miles de euros, de los trabajadores de una empresa dedicada a la construcción de viviendas.

Salarios	N.º trabajadores
0,6 - 1,0	10
1,0 - 1,2	15
1,2 - 2,0	40
2,0 - 3,0	30
3,0 - 3,2	5

SOLUCIÓN

Puesto que los rectángulos que se elevan sobre cada intervalo componiendo el histograma de frecuencias han de tener un área igual a la correspondiente frecuencia, hay que calcular, para cada uno de ellos, la altura, conocida la longitud de la base o amplitud del intervalo. Así, si

$$n_i = d_i \cdot c_i,$$

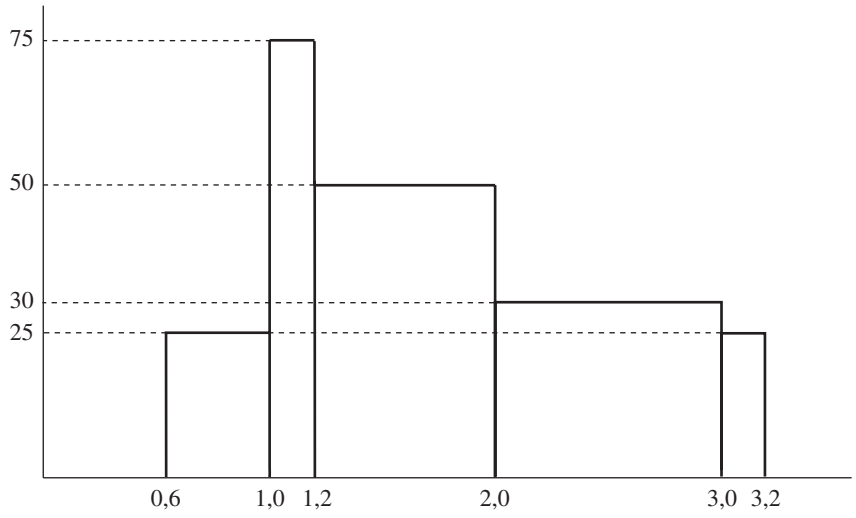
es el área (frecuencia) del intervalo genérico expresada como producto de la longitud del intervalo, c_i, y de la altura o densidad de frecuencia, d_i, entonces,

$$d_i = \frac{n_i}{c_i}.$$

En la siguiente tabla se recoge la información correspondiente a cada intervalo, esto es, la distribución de frecuencias de la variable, junto con las amplitudes y las densidades de frecuencia.

$L_{i-1}-L_i$	n_i	c_i	d_i
0,6 - 1,0	10	0,4	25
1,0 - 1,2	15	0,2	75
1,2 - 2,0	40	0,8	50
2,0 - 3,0	30	1,0	30
3,0 - 3,2	5	0,2	25
	$N = 100$		

Partiendo de la tabla anterior, construimos el siguiente histograma donde cada rectángulo tiene como base el intervalo y como altura la densidad de frecuencia. Observemos, por ejemplo, que el intervalo 1,2-2,0 tiene una altura igual a 50, lo cual significa que su área es $0{,}8 \cdot 50 = 40$, cantidad que, evidentemente, coincide con su frecuencia absoluta.



1.4 Se considera la distribución de frecuencias con datos agrupados $(L_{i-1} - L_i; f_i)$.

a) Se realiza una transformación, obteniéndose la nueva distribución $(k \cdot L_{i-1} - k \cdot L_i; f_i)$ (k número real, $k \neq 0$). ¿Qué efecto produce esta transformación sobre las amplitudes y las densidades de frecuencias de los intervalos?

b) Dada la distribución transformada $((L_{i-1} + k) - (L_i + k); f_i)$ (k número real), relaciónense sus densidades de frecuencia con las correspondientes en la distribución inicial.

SOLUCIÓN

a) La amplitud del intervalo genérico de la distribución transformada es

$$k \cdot L_i - k \cdot L_{i-1} = k(L_i - L_{i-1}) = k \cdot c_i,$$

es decir, la amplitud del intervalo genérico inicial, c_i, queda multiplicada por la misma constante, k.

Por otro lado, la densidad de frecuencia del nuevo intervalo genérico, esto es, el cociente entre la frecuencia absoluta y la amplitud del intervalo,

$$\frac{n_i}{k \cdot c_i} = \frac{1}{k} \cdot \frac{n_i}{c_i} = \frac{1}{k} \cdot d_i,$$

resulta ser igual a la densidad de frecuencia del intervalo original, d_i, dividida por la constante, k.

b) La densidad de frecuencia del intervalo genérico en la distribución transformada es

$$\frac{n_i}{(L_i + k) - (L_{i-1} + k)} = \frac{n_i}{L_i - L_{i-1}} = \frac{n_i}{c_i} = d_i,$$

que coincide, por tanto, con la densidad de frecuencia del intervalo genérico en la distribución de partida.

1.5 El número de contratos formalizados por los 20 trabajadores del departamento de ventas de una promotora inmobiliaria durante el pasado año han sido:

10	10	30	18	32
21	32	32	29	28
21	21	30	15	28
22	24	28	18	21

a) Calcúlese el número medio de contratos formalizados por trabajador.

b) Obténgase el número total de ventas del departamento.

SOLUCIÓN

a) A partir de la estadística primaria se obtiene la distribución de frecuencias recogida en la siguiente tabla:

Contratos de ventas	N.º trabajadores
10	2
15	1
18	2
21	4
22	1
24	1
28	3
29	1
30	2
32	3

El cálculo de la media aritmética,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_i,$$

conduce al valor

$$\bar{x} = \frac{10 \cdot 2 + 15 \cdot 1 + 18 \cdot 2 + 21 \cdot 4 + 22 \cdot 1 + 24 \cdot 1 + 28 \cdot 3 + 29 \cdot 1 + 30 \cdot 2 + 32 \cdot 3}{20} = 23{,}5,$$

esto es, el número medio de contratos formalizados por trabajador es 23,5.

Se podría haber llegado a la misma solución a partir de la estadística primaria sin necesidad de obtener la distribución de frecuencias. Para ello, bastaría con haber sumado todas las observaciones y dividido el resultado por 20, número de ellas. En realidad, es lo que hemos hecho con nuestros cálculos, apoyándonos en una presentación simplificada de la estadística primaria como es la distribución de frecuencias.

b) Partiendo del valor medio calculado en el apartado anterior, se obtiene que el total de ventas es $N \cdot \bar{x} = 20 \cdot 23{,}5 = 470$, cantidad a la que, evidentemente, también llegaríamos sumando los datos de la estadística primaria y que, por supuesto, coincide con el numerador de la expresión de la media aritmética.

1.6 Dada una distribución de frecuencias $(x_i; f_i)$, demuéstrese que

$$\sum_{i=1}^{h} (x_i - \bar{x})\, f_i = 0.$$

SOLUCIÓN

Operando en el sumatorio,

$$\sum_{i=1}^{h} (x_i - \bar{x})\, f_i = \sum_{i=1}^{h} (x_i \cdot f_i - \bar{x} \cdot f_i) = \sum_{i=1}^{h} x_i \cdot f_i - \sum_{i=1}^{h} \bar{x} \cdot f_i,$$

y teniendo en cuenta que

$$\sum_{i=1}^{h} x_i \cdot f_i = \bar{x},$$

y que, además, $\bar{x}$ no depende de i y, por tanto, puede escribirse fuera del sumatorio, la expresión anterior resulta ser igual a

$$\bar{x} - \bar{x} \cdot \sum_{i=1}^{h} f_i.$$

Ahora bien, la suma de las frecuencias relativas de una distribución,

$$\sum_{i=1}^{h} f_i = \sum_{i=1}^{h} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{h} n_i = \frac{N}{N},$$

es, en consecuencia, igual a la unidad, con lo cual,

$$\sum_{i=1}^{h} (x_i - \bar{x})\, f_i = \bar{x} - \bar{x} = 0.$$

De la propiedad demostrada se deduce que, también, $\sum_{i=1}^{h} (x_i - \bar{x})\, n_i$ es igual a cero.

1.7 Demuéstrese que la media aritmética de las desviaciones al cuadrado de los valores de una distribución $(x_i; f_i)$, respecto a un valor constante, se hace mínima cuando dicha constante es la media aritmética de la distribución.

SOLUCIÓN

La media aritmética de las desviaciones al cuadrado de las observaciones respecto de una constante, k, es una función de dicho valor constante. Denotemos por $d\,(k)$ a esa función:

$$d(k) = \sum_{i=1}^{h} (x_i - k)^2 f_i.$$

Sumando y restando la media aritmética de la distribución, $\bar{x}$, y agrupando términos, se tiene que

$$d(k) = \sum_{i=1}^{h} (x_i - \bar{x} + \bar{x} - k)^2 f_i = \sum_{i=1}^{h} [(x_i - \bar{x}) + (\bar{x} - k)]^2 f_i.$$

Desarrollando el binomio y descomponiendo el sumatorio en tres sumandos:

$$\begin{aligned} d(k) &= \sum_{i=1}^{h} [(x_i - \bar{x})^2 + (\bar{x} - k)^2 + 2(x_i - \bar{x}) \cdot (\bar{x} - k)]\, f_i = \\ &= \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i + \sum_{i=1}^{h} (\bar{x} - k)^2 f_i + 2 \sum_{i=1}^{h} (x_i - \bar{x}) \cdot (\bar{x} - k)\, f_i. \end{aligned}$$

Como $(\bar{x} - k)$ es un valor constante, esto es, no depende de i, puede escribirse fuera de los correspondientes sumatorios:

$$d(k) = \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i + (\bar{x} - k)^2 \sum_{i=1}^{h} f_i + 2(\bar{x} - k) \sum_{i=1}^{h} (x_i - \bar{x})\, f_i.$$

Dado que $\sum_{i=1}^{h} f_i = 1$ y que el último sumando es cero pues, según se demostró en **1.6**, $\sum_{i=1}^{h} (x_i - \bar{x})\, f_i = 0$, entonces,

$$d(k) = \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i + (\bar{x} - k)^2.$$

Puesto que estos dos sumandos son cantidades positivas y el primero no depende de k, el mínimo valor de la función $d\,(k)$ se alcanza cuando $(\bar{x} - k)^2$ es igual a cero, hecho que se produce cuando la constante k coincide con la media aritmética, $\bar{x}$.

1.8 Dada una distribución de frecuencias $(x_i; f_i)$, cuya media es $\bar{x}$, obténgase la media de la distribución de frecuencias $(a \cdot x_i + b; f_i)$, donde a y b son números reales cualesquiera. En particular, calcúlese la media aritmética de la distribución transformada por un cambio de origen y de escala.

SOLUCIÓN

Aplicando la definición de media aritmética a la distribución $(a \cdot x_i + b; f_i)$ y operando en el sumatorio, resulta que la media aritmética de la distribución transformada es igual a

$$\sum_{i=1}^{h} (a \cdot x_i + b) f_i = \sum_{i=1}^{h} (a \cdot x_i \cdot f_i + b \cdot f_i) = a \sum_{i=1}^{h} x_i \cdot f_i + b \sum_{i=1}^{h} f_i.$$

Ahora bien, $\sum_{i=1}^{h} x_i \cdot f_i = \bar{x}$ y $\sum_{i=1}^{h} f_i = 1$, con lo cual, la media de la distribución $(a \cdot x_i + b; f_i)$ es $a \cdot \bar{x} + b$.

En particular, si $a = 1/e$ y $b = -o/e$, es decir, si realizamos un cambio de origen y de escala, la media aritmética de la distribución resultante es

$$\frac{1}{e} \cdot \bar{x} - \frac{o}{e} = \frac{\bar{x} - o}{e},$$

con e y o números reales $(e > 0)$.

1.9 Dada una distribución de frecuencias $(x_i; f_i)$, compruébese que el inverso de su media armónica, H, es igual a la media aritmética de los inversos de los valores de la distribución.

SOLUCIÓN

El inverso de la media armónica,

$$\mathrm{H} = \frac{1}{\sum_{i=1}^{h} \frac{1}{x_i} \cdot f_i},$$

es, sin más que invertir los dos miembros de la igualdad anterior,

$$\frac{1}{\mathrm{H}} = \sum_{i=1}^{h} \frac{1}{x_i} \cdot f_i,$$

valor que coincide con la media aritmética de los inversos de los valores de la distribución, esto es, con la media aritmética de la distribución de frecuencias $(1/x_i; f_i)$.

1.10 En una nueva zona de expansión de la ciudad, la promotora Miraluna está construyendo apartamentos, pisos de dos habitaciones y dúplex.

El precio por metro cuadrado de las baldosas de las cocinas en los apartamentos es de 24 euros, en los pisos de 30 euros y en los dúplex de 42, y el coste total de los suelos de cocina en cada tipo de viviendas de 21 600, 36 000 y 10 080 euros, respectivamente.

Calcúlese el precio medio por metro cuadrado de azulejado del suelo de las cocinas en toda la obra.

SOLUCIÓN

La distribución de frecuencias del precio por metro cuadrado se recoge en la tabla siguiente:

Precio por metro cuadrado	Coste
24	21 600
30	36 000
42	10 080

Para calcular el precio medio por metro cuadrado, promedio de una magnitud relativa, hay que obtener la media armónica de la distribución anterior:

$$H = \frac{N}{\sum_{i=1}^{h} \frac{1}{x_i} \cdot n_i} = \frac{21\,600 + 36\,000 + 10\,080}{\frac{1}{24} \cdot 21\,600 + \frac{1}{30} \cdot 36\,000 + \frac{1}{42} \cdot 10\,080} = 28{,}92 \text{ euros.}$$

Téngase en cuenta que esta media armónica es, en realidad,

$$H = \frac{\text{coste total}}{\text{superficie total}},$$

donde la superficie total es el resultado de sumar la superficie del suelo para cada tipo de vivienda obtenida, a su vez, dividiendo el correspondiente coste entre el respectivo precio por metro cuadrado de las baldosas.

1.11 La siguiente tabla recoge la distribución de ayudas para estudios, en miles de euros, que prestan las empresas de un determinado sector, así como el número de trabajadores por empresa que reciben dichas ayudas.

Importe	N.º empresas	N.º trabajadores por empresa
0-10	600	0-50
10-100	500	110-150
100-500	50	150-200
500-2 500	8	50-100
2 500-5 000	1	100-120

a) ¿Cuál es el importe medio de la ayuda por empresa?

b) ¿Qué número medio de trabajadores por empresa es receptor de la ayuda?

SOLUCIÓN

a) Para calcular el valor medio de las ayudas por empresa se considera la siguiente distribución de frecuencias:

Importe	N.º empresas
0-10	600
10-100	500
100-500	50
500-2 500	8
2 500-5 000	1

Utilizando las marcas de clase de los intervalos anteriores, que son, respectivamente, 5, 55, 300, 1 500 y 3 750, se obtiene la media de la distribución, esto es, el importe medio de las ayudas por empresa, en miles de euros,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_i = \frac{5 \cdot 600 + 55 \cdot 500 + 300 \cdot 50 + 1\,500 \cdot 8 + 3\,750 \cdot 1}{1\,159} = 52{,}85.$$

b) El número medio de trabajadores receptor de la ayuda por empresa es

$$\frac{\text{número total de trabajadores}}{\text{número total de empresas}},$$

con

$$\text{número total de empresas} = 600 + 500 + 50 + 8 + 1 = 1\,159$$

y

$$\text{número total de trabajadores} = 600 \cdot 25 + 500 \cdot 130 + 50 \cdot 175 + 8 \cdot 75 + 1 \cdot 110 = 89\,460.$$

Dado que no conocemos el número exacto de trabajadores receptores de ayuda en cada empresa, el número total de trabajadores se ha calculado de modo aproximado, tomando las marcas de clase de los intervalos de la última columna de la tabla proporcionada por el enunciado.

En definitiva, la media pedida es

$$\frac{89\,460}{1\,159} = 77{,}19 \text{ trabajadores por empresa.}$$

Esta media, tal y como la hemos calculado, se corresponde con la *media aritmética* de una distribución con valores —sin ordenar— 25, 130, 175, 75 y 110 y con frecuencias 600, 500, 50, 8 y 1, respectivamente. Ahora bien, también podría interpretarse como la *media armónica* de la siguiente distribución de frecuencias, donde cada elemento de la segunda columna es el producto entre el número de trabajadores por empresa y el correspondiente número de empresas:

N.º trabajadores por empresa	N.º trabajadores
25	15 000
130	65 000
175	8 750
75	600
110	110

En efecto,

$$H = \frac{15\,000 + 65\,000 + 8\,750 + 600 + 110}{\frac{1}{25}\cdot 15\,000 + \frac{1}{130}\cdot 65\,000 + \frac{1}{175}\cdot 8\,750 + \frac{1}{75}\cdot 600 + \frac{1}{110}\cdot 110} = 77{,}19$$

es el promedio de la magnitud relativa número de trabajadores por empresa.

1.12 Dada la distribución de frecuencias $(x_i; n_i)$, demuéstrese que

$$G = \prod_{i=1}^{h} x_i^{f_i}.$$

SOLUCIÓN

La demostración es inmediata, sin más que aplicar propiedades aritméticas elementales:

$$G = \sqrt[N]{\prod_{i=1}^{h} x_i^{n_i}} = \left(\prod_{i=1}^{h} x_i^{n_i}\right)^{1/N} = \prod_{i=1}^{h} x_i^{n_i/N},$$

Ahora bien, puesto que $n_i/N = f_i$, se tiene que la media geométrica puede expresarse también en función de las frecuencias relativas de la distribución:

$$G = \prod_{i=1}^{h} x_i^{f_i}.$$

1.13 Dada una distribución de frecuencias $(x_i; f_i)$, demuéstrese que el logaritmo de la media geométrica, G, es la media aritmética de los logaritmos de los valores de la distribución.

SOLUCIÓN

Partiendo del resultado probado en **1.12**,

$$G = \prod_{i=1}^{h} x_i^{f_i},$$

tomamos logaritmos y aplicamos las propiedades de los mismos, obteniéndose:

$$\log G = \log \left(\prod_{i=1}^{h} x_i^{f_i} \right) = \sum_{i=1}^{h} \log x_i^{f_i} = \sum_{i=1}^{h} \log x_i \cdot f_i,$$

con lo cual, el logaritmo de la media geométrica es la media aritmética de los logaritmos de los valores de la distribución, esto es, la media aritmética de la distribución $(\log x_i; f_i)$.

1.14 En un grupo de empresas dedicadas a conservas de pescado se conocen los porcentajes de empleadas que trabajan en ellas:

Empresa	% mujeres
A	20
B	20
C	30
D	50
E	40
F	30

Calcúlese la media geométrica del porcentaje de mujeres trabajadoras.

SOLUCIÓN

La transformación de la estadística primaria en la correspondiente distribución de frecuencias,

x_i	n_i
20	2
30	2
40	1
50	1

permite calcular

$$G = \sqrt[N]{\prod_{i=1}^{h} x_i^{n_i}} = \sqrt[6]{20^2 \cdot 30^2 \cdot 40 \cdot 50} = 29{,}94,$$

porcentaje medio de mujeres trabajadoras por empresa.

1.15. El señor Pérez, al llegar a su vejez, decide adaptarse a los tiempos modernos, adquiriendo un teléfono móvil. Transcurrido un mes, la compañía telefónica le remite «el detalle» de las llamadas efectuadas durante ese periodo:

Duración (en minutos)	Llamadas a móviles	Llamadas a fijos	Llamadas al extranjero
0-10	3	2	1
10-30	10	25	0
30-60	25	10	0

El precio por minuto de las llamadas realizadas a móviles es de 0,12 euros, siendo éste de 0,15 y 0,8 euros, respectivamente, para las llamadas a fijos y al extranjero. Se sabe, además, que el coste de establecimiento es de 0,2 euros por llamada.

Calcúlese:

a) El gasto total del mes en llamadas de duración no superior a treinta minutos.

b) El coste medio por llamada efectuada por el señor Pérez a teléfonos móviles.

SOLUCIÓN

a) Para calcular el gasto total en llamadas de duración inferior a 30 minutos es necesario hallar, en primer lugar, el coste de las llamadas de duración entre 0 y 10 minutos, utilizando la marca de clase de este intervalo,

$$0{,}2\ (3 + 2 + 1) + 5\ (0{,}12 \cdot 3 + 0{,}15 \cdot 2 + 0{,}8 \cdot 1) = 8{,}5 \text{ euros},$$

y sumar a esta cantidad el coste en llamadas de duración entre 10 y 30 minutos,

$$0{,}2\ (10 + 25 + 0) + 20\ (0{,}12 \cdot 10 + 0{,}15 \cdot 25 + 0{,}8 \cdot 0) = 106 \text{ euros},$$

obteniéndose, así, el coste total pedido:

$$8{,}5 + 106 = 114{,}5 \text{ euros}.$$

b) Si X es la duración, en minutos, de las llamadas a móviles, variable cuya distribución proporciona el enunciado con las dos primeras columnas de la tabla anterior, y C el coste de este tipo de llamadas, se tiene la relación lineal:

$$C = 0{,}2 + 0{,}12 \cdot X,$$

es decir, la distribución de la variable C es una distribución transformada de la distribución de la variable X, donde, siguiendo la notación de **1.8**, $a = 0{,}12$ y $b = 0{,}2$.

Por tanto, calculada la duración media de estas llamadas,

$$\overline{x} = \frac{5 \cdot 3 + 20 \cdot 10 + 45 \cdot 25}{3 + 10 + 25} = 35{,}26 \text{ minutos},$$

donde 5, 20 y 45 son las marcas de clase de los intervalos, y teniendo en cuenta el resultado ya citado donde se relacionan las medias de una distribución de frecuencias y de una distribución obtenida a partir de ella mediante transformación lineal, se concluye que el coste medio por llamada a teléfonos móviles del señor Pérez es

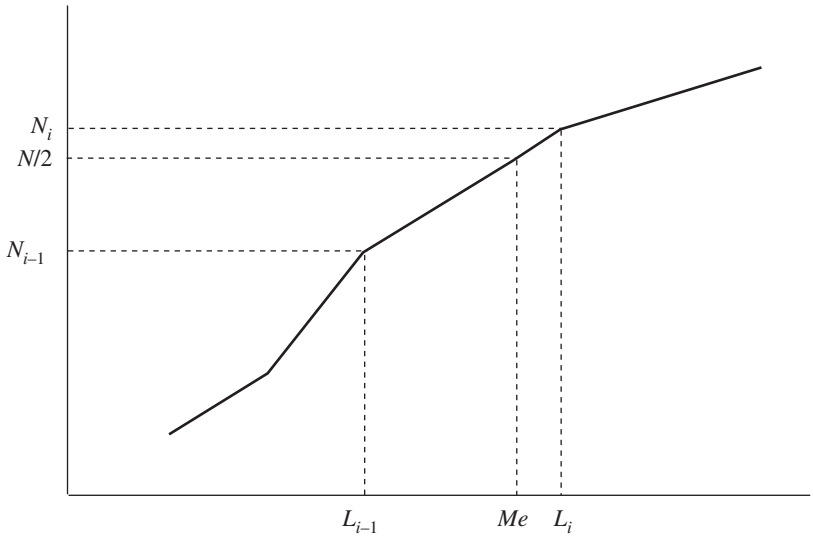
$$\overline{c} = 0{,}2 + 0{,}12 \cdot \overline{x} = 0{,}2 + 0{,}12 \cdot 35{,}26 = 4{,}43 \text{ euros}.$$

Invitamos al lector a que calcule con el mismo procedimiento el coste medio por llamada del señor Pérez a teléfonos fijos y al extranjero.

1.16 Dada una distribución de frecuencias agrupada en intervalos $(L_{i-1} - L_i; f_i)$, obténgase la expresión de la mediana.

SOLUCIÓN

Representemos la parte del polígono de frecuencias acumuladas correspondiente al intervalo mediano, $L_{i-1} - L_i$.



Según se ilustra en esta gráfica, la mediana es un valor cuya frecuencia absoluta acumulada es igual a $N/2$. Podemos observar, también, que el punto de coordenadas $(Me, N/2)$ pertenece a la recta que une los puntos (L_{i-1}, N_{i-1}) y (L_i, N_i), con lo cual, para hallar la expresión de la mediana basta con sustituir el valor de la abscisa, Me, y el de la ordenada, $N/2$, en la ecuación de la recta que une dichos puntos[1]:

$$\frac{x - L_{i-1}}{L_i - L_{i-1}} = \frac{y - N_{i-1}}{N_i - N_{i-1}},$$

o, lo que es lo mismo, en

$$\frac{x - L_{i-1}}{c_i} = \frac{y - N_{i-1}}{n_i}.$$

[1] Recuerde el lector que dados los puntos (a, b) y (c, d), la expresión de la recta que pasa por ellos es

$$\frac{x - a}{c - a} = \frac{y - b}{d - b}.$$

Sustituyendo, entonces, en esta ecuación el punto (Me, $N/2$), se tiene:

$$\frac{Me - L_{i-1}}{c_i} = \frac{\frac{N}{2} - N_{i-1}}{n_i},$$

con lo que, despejando, resulta el valor de la mediana:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i.$$

1.17 Una empresa dedicada al transporte de viajeros cuenta con cien vehículos para largos recorridos.

El pasado año la distribución del número de kilómetros recorridos, en miles, por los vehículos se recoge en la siguiente tabla.

Kilómetros recorridos	N.º vehículos
100	20
120	10
160	60
230	5
250	5

a) ¿Qué número de kilómetros recorre la mayoría de los vehículos?

b) Hállese el número mínimo de kilómetros que tiene que recorrer un vehículo para estar dentro del 50 por ciento de los que más kilómetros recorren.

SOLUCIÓN

a) Se trata de obtener el valor de la variable con mayor frecuencia, esto es, la moda de la distribución de frecuencias proporcionada por el enunciado. En este caso, la mayor frecuencia, 60, corresponde al valor $x_3 = 160$, concluyéndose que la moda, es decir, el número de kilómetros que recorre la mayoría de los vehículos, es 160 mil kilómetros.

b) En la siguiente tabla, que completa la anterior, se recogen las frecuencias absolutas acumuladas que permitirán la obtención de la mediana, medida de posición que hay que calcular en este apartado.

Recuerde el lector que

$$N_1 = n_1 \text{ y } N_i = n_1 + \ldots + n_i, \text{ para } i = 2, ..., h,$$

con lo cual, como ya es sabido, cada frecuencia absoluta acumulada puede calcularse a partir de la anterior:

$$N_1 = n_1$$

y

$$N_i = N_{i-1} + n_i.$$

Así,

x_i	n_i	N_i
100	20	20
120	10	30
160	60	90
230	5	95
250	5	100

Puesto que no existe ninguna frecuencia absoluta acumulada que coincida con

$$\frac{N}{2} = 50,$$

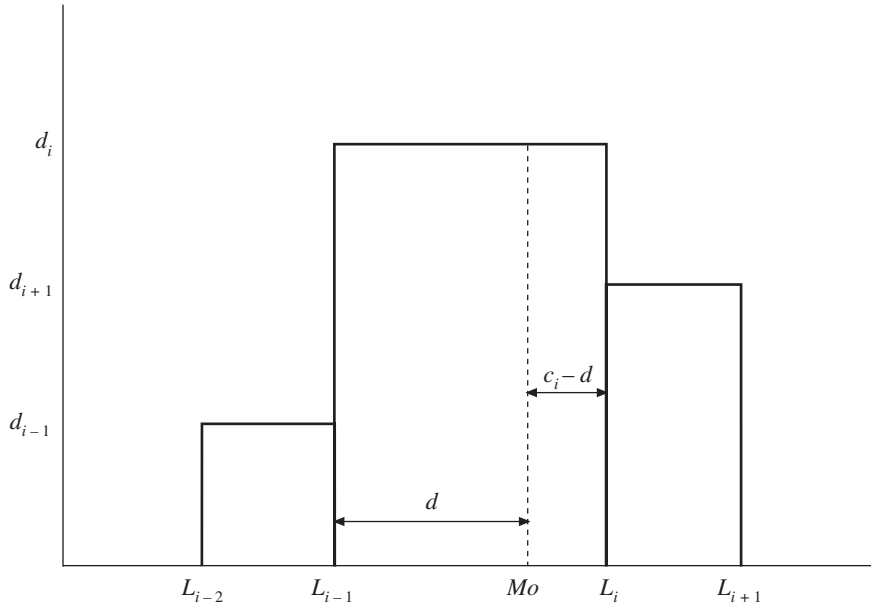
la mediana es el mínimo valor de la variable cuya frecuencia absoluta acumulada es estrictamente mayor que 50: la mediana es, en este caso, $x_3 = 160$, ya que a este valor le corresponde una frecuencia $N_3 = 90 > 50$, siendo el valor más pequeño que cumple tal condición.

Obsérvese que, en esta situación, coinciden la moda y la mediana de la distribución.

1.18 Dada una distribución de frecuencias agrupada en intervalos $(L_{i-1} - L_i; f_i)$, obténgase la expresión de la moda.

SOLUCIÓN

Dentro del histograma de frecuencias, fijémonos en el intervalo modal, $L_{i-1} - L_i$, y en sus dos intervalos contiguos, $L_{i-2} - L_{i-1}$ y $L_i - L_{i+1}$:



Suponiendo que la moda está más cerca del intervalo con mayor densidad de frecuencia —hipótesis que parece sostenible por el concepto de moda—, se cumple, entonces, que la distancia de la moda a cada uno de los intervalos contiguos, d y $c_i - d$, es *inversamente proporcional* a la correspondiente densidad de frecuencia. Esto es lo mismo que decir que el cociente entre las distancias, d y $c_i - d$, es igual al inverso del cociente entre las densidades de frecuencias, d_{i-1} y d_{i+1}:

$$\frac{d}{c_i - d} = \frac{d_{i+1}}{d_{i-1}}.$$

Despejando, se tiene que

$$d = \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i$$

y, en consecuencia, la moda que, como puede verse en la gráfica, es igual a $L_{i-1} + d$, responde a la expresión:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i,$$

sólo con sustituir d por su valor.

Si los intervalos tienen la misma amplitud, pueden utilizarse las frecuencias en lugar de las densidades de frecuencia:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c,$$

donde c es la amplitud de los intervalos.

1.19 La distribución de salarios mensuales, en miles de euros, de una empresa constructora es la siguiente:

Salarios	N.º trabajadores
0,6-0,9	30
0,9-1,2	60
1,2-1,5	5
1,5-1,8	3
1,8-2,1	2

a) ¿Cuál es el salario medio mensual?

b) Hállese el valor del salario tal que la mitad de los trabajadores perciba un salario superior a dicho valor y la otra mitad un salario inferior.

c) El salario más frecuente.

SOLUCIÓN

De los datos del enunciado se obtiene la siguiente tabla en la que aparecen las marcas de clase y las frecuencias absolutas y absolutas acumuladas de cada intervalo.

Salarios	x_i	n_i	N_i
0,6-0,9	0,75	30	30
0,9-1,2	1,05	60	90
1,2-1,5	1,35	5	95
1,5-1,8	1,65	3	98
1,8-2,1	1,95	2	100

a) El salario medio mensual por trabajador, esto es, la media aritmética de la distribución de frecuencias es, en miles de euros,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_i = \frac{0{,}75 \cdot 30 + 1{,}05 \cdot 60 + 1{,}35 \cdot 5 + 1{,}65 \cdot 3 + 1{,}95 \cdot 2}{100} = 1{,}01.$$

b) La medida de posición pedida se corresponde con la mediana de la distribución. Para hallarla hay que considerar, en primer lugar, que el intervalo mediano es 0,9-1,2, ya que es el primer intervalo cuya frecuencia absoluta acumulada, $N_2 = 90$, es estrictamente mayor que $N/2$, que, en este caso, es igual a 50.

Del intervalo mediano se obtiene la mediana, aplicando la expresión:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i.$$

Así, con los datos del enunciado resulta que

$$Me = 0{,}9 + \frac{50 - 30}{60} \cdot 0{,}3 = 1,$$

es decir, la mediana de los salarios es igual a mil euros.

c) El salario más frecuente, es decir, la moda de la distribución de los salarios, se encuentra dentro del intervalo modal, o intervalo con mayor frecuencia —pues todos los intervalos tienen la misma amplitud—, que, en esta ocasión, es el segundo intervalo, 0,9-1.2.

Al ser, como hemos dicho, todos los intervalos de igual amplitud, pueden utilizarse las frecuencias, en lugar de las densidades de frecuencias, en la expresión que permite el cálculo de la moda:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c,$$

resultando, por tanto,

$$Mo = 0{,}9 + \frac{5}{30 + 5} \cdot 0{,}3 = 0{,}94 \text{ miles de euros.}$$

1.20 Estúdiese el efecto de una transformación lineal sobre la moda de una distribución $(L_{i-1} - L_i; f_i)$.

SOLUCIÓN

Si $L_{i-1} - L_i$ es el intervalo modal de la distribución $(L_{i-1} - L_i; f_i)$, o intervalo con mayor densidad de frecuencia, la moda de la distribución es

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i.$$

Puesto que la transformación lineal no afecta a las frecuencias de la distribución, el intervalo modal de la distribución transformada $((a \cdot L_{i-1} + b) - (a \cdot L_i + b); f_i)$, con a y b constantes cualesquiera, será $(a \cdot L_{i-1} + b) - (a \cdot L_i + b)$, donde

$$a \cdot L_i + b - (a \cdot L_{i-1} + b) = a(L_i - L_{i-1}) = a \cdot c_i$$

es su amplitud, y

$$\frac{n_{i-1}}{a(L_{i-1} - L_{i-2})} = \frac{n_{i-1}}{a \cdot c_{i-1}} = \frac{1}{a} \cdot d_{i-1}$$

y

$$\frac{n_{i+1}}{a(L_{i+1} - L_i)} = \frac{n_{i+1}}{a \cdot c_{i+1}} = \frac{1}{a} \cdot d_{i+1}$$

son las densidades de frecuencia de los intervalos contiguos.

En consecuencia, la moda de la distribución transformada es

$$Mo' = (a \cdot L_{i-1} + b) + \frac{\frac{1}{a} \cdot d_{i+1}}{\frac{1}{a} \cdot d_{i-1} + \frac{1}{a} \cdot d_{i+1}} \cdot a \cdot c_i = (a \cdot L_{i-1} + b) + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a \cdot c_i,$$

expresión que, tras sencillas operaciones, se convierte en

$$Mo' = a\left(L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i\right) + b = a \cdot Mo + b.$$

Por ello, si se realiza una transformación en los valores de la distribución, pasando del valor genérico x_i al valor $x_i + b$, la moda, Mo, como valor de la distribución, se verá afectada también por la transformación, pasando a ser $Mo + b$. Ahora bien, este valor de la distribución transformada será también la moda de la nueva distribución, ya que, al no modificarse las frecuencias, seguirá teniendo la mayor de todas ellas.

1.21 Dada una distribución de frecuencias $(x_i; f_i)$, demuéstrese que

$$S^2 = \sum_{i=1}^{h} x_i^2 \cdot f_i - \bar{x}^2.$$

SOLUCIÓN

Operando en la expresión de la varianza, esto es, desarrollando el binomio, descomponiendo en tres sumandos y poniendo fuera de los sumatorios los términos constantes, resulta que

$$S^2 = \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i = \sum_{i=1}^{h} (x_i^2 + \bar{x}^2 - 2 \cdot x_i \cdot \bar{x}) f_i = \sum_{i=1}^{h} x_i^2 \cdot f_i + \bar{x}^2 \sum_{i=1}^{h} f_i - 2 \cdot \bar{x} \sum_{i=1}^{h} x_i \cdot f_i.$$

Como $\sum_{i=1}^{h} f_i = 1$ y $\sum_{i=1}^{h} x_i \cdot f_i = \bar{x}$, se tiene que la varianza de la distribución de frecuencias es

$$S^2 = \sum_{i=1}^{h} x_i^2 \cdot f_i + \bar{x}^2 - 2 \cdot \bar{x} \cdot \bar{x} = \sum_{i=1}^{h} x_i^2 \cdot f_i - \bar{x}^2,$$

según queríamos probar.

1.22 Dada la distribución de frecuencias $(x_i; f_i)$, cuya varianza es S^2, determínese la varianza de la distribución de frecuencias $(a \cdot x_i + b; f_i)$, donde a y b son constantes cualesquiera. ¿Cuál es la desviación típica? Aplíquense los resultados obtenidos al caso particular de un cambio de origen y de escala.

SOLUCIÓN

Aplicando la definición de varianza a la distribución transformada cuya media, según hemos demostrado anteriormente, es $a \cdot \bar{x} + b$, se tiene que la varianza de la nueva distribución es

$$\sum_{i=1}^{h} [a \cdot x_i + b - (a \cdot \bar{x} + b)]^2 f_i = \sum_{i=1}^{h} (a \cdot x_i + b - a \cdot \bar{x} - b)^2 f_i = \sum_{i=1}^{h} (a \cdot x_i - a \cdot \bar{x})^2 f_i,$$

sin más que simplificar.

Sacando factor común a la constante a^2, resulta que la varianza pedida es

$$a^2 \sum_{i=1}^{h} (x_i - \bar{x})^2 f_i = a^2 \cdot S^2.$$

Por tanto, la desviación típica de la nueva distribución es $|a| \cdot S$, esto es, la raíz cuadrada positiva de la varianza.

En particular, si $a = 1/e$ y $b = -o/e$, la varianza de la distribución transformada por un cambio de origen y de escala es

$$\left(\frac{1}{e}\right)^2 S^2 = \frac{S^2}{e^2}.$$

Además, como $e > 0$, la desviación típica de la distribución transformada por un cambio de origen y de escala es S/e.

En consecuencia, tanto la varianza como la desviación típica se ven afectadas únicamente por cambios de escala.

1.23 La Administración Autonómica de cierta región cuenta con 1 620 empleados públicos cuya distribución de salarios, en miles de euros, se refleja en la siguiente tabla.

Salarios	N.º empleados
0,6	20
1,0	200
1,5	500
1,8	300
2,0	400
2,3	200

a) Hállese la media, la mediana y la moda de la distribución de los salarios.

b) ¿Cuál de los tres promedios es más representativo?

SOLUCIÓN

a) La primera columna de la tabla anterior corresponde a los valores de la variable y la segunda a las frecuencias absolutas, con lo cual, el salario medio mensual por empleado público, es decir, la media aritmética, en miles de euros, de la distribución de los salarios es

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_i = \frac{0{,}6 \cdot 20 + 1 \cdot 200 + 1{,}5 \cdot 500 + 1{,}8 \cdot 300 + 2 \cdot 400 + 2{,}3 \cdot 200}{1\,620} = 1{,}7.$$

El salario más frecuente, la moda de la distribución de los salarios, es el valor de la variable con mayor frecuencia:

$$Mo = 1{,}5 \text{ miles de euros.}$$

Por último, la mediana de los salarios es el mínimo valor cuya frecuencia absoluta acumulada es estrictamente mayor que 1 620/2 = 810. En esta distribución, $x_4 = 1{,}8$ tiene una frecuencia absoluta acumulada $N_4 = 1\,020$, por lo que

$$Me = 1{,}8 \text{ miles de euros.}$$

b) Para estudiar la representatividad de los promedios, utilizaremos el índice de dispersión calculado respecto a cada uno de ellos.

Así, por lo que respecta a la media aritmética, hallaremos

$$I_{\bar{x}} = \frac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{h} |x_i - \bar{x}| \cdot n_i}{\bar{x}},$$

donde

$$\frac{1}{N}\sum_{i=1}^{h} |x_i - \bar{x}| \cdot n_i = \frac{1}{1\,620}\big(|0{,}6 - 1{,}7| \cdot 20 + |1 - 1{,}7| \cdot 200 + |1{,}5 - 1{,}7| \cdot 500 +$$

$$+ |1{,}8 - 1{,}7| \cdot 300 + |2 - 1{,}7| \cdot 400 + |2{,}3 - 1{,}7| \cdot 200\big) = 0{,}33,$$

con lo cual, el índice de dispersión de la media es

$$I_{\bar{x}} = \frac{0{,}33}{1{,}7} = 0{,}19.$$

El índice de dispersión de la moda es

$$I_{Mo} = \frac{\dfrac{1}{N}\displaystyle\sum_{i=1}^{h} |x_i - Mo| \cdot n_i}{Mo}.$$

Por ello, calculamos:

$$\frac{1}{N}\sum_{i=1}^{h} |x_i - Mo| \cdot n_i = \frac{1}{1\,620}\big(|0{,}6 - 1{,}5| \cdot 20 + |1 - 1{,}5| \cdot 200 + |1{,}5 - 1{,}5| \cdot 500 +$$

$$+ |1{,}8 - 1{,}5| \cdot 300 + |2 - 1{,}5| \cdot 400 + |2{,}3 - 1{,}5| \cdot 200\big) = 0{,}35.$$

En definitiva,

$$I_{Mo} = \frac{0{,}35}{1{,}5} = 0{,}23.$$

Por último, el índice de dispersión de la mediana es

$$I_{Me} = \frac{\frac{1}{N}\sum_{i=1}^{h} |x_i - Me| \cdot n_i}{Me},$$

para cuyo cálculo obtendremos:

$$\frac{1}{N}\sum_{i=1}^{h} |x_i - Me| \cdot n_i = \frac{1}{1\,620}\big(|0{,}6 - 1{,}8| \cdot 20 + |1 - 1{,}8| \cdot 200 + |1{,}5 - 1{,}8| \cdot 500 +$$

$$+ |1{,}8 - 1{,}8| \cdot 300 + |2 - 1{,}8| \cdot 400 + |2{,}3 - 1{,}8| \cdot 200\big) = 0{,}32.$$

De este modo,

$$I_{Me} = \frac{0{,}32}{1{,}8} = 0{,}18.$$

La comparación de los tres índices de dispersión permite afirmar que la mediana, con un índice de dispersión ligeramente más pequeño, es la medida de posición más representativa para esta distribución de frecuencias, seguida de la media aritmética y después de la moda.

1.24 Dada la distribución de frecuencias $(x_i; f_i)$, cuya media y desviación típica son $\overline{x}$ y S, respectivamente, obténganse la media y la desviación típica de la distribución tipificada $\left(\frac{x_i - \overline{x}}{S}; f_i\right)$.

SOLUCIÓN

Denotemos por

$$z_i = \frac{x_i - \overline{x}}{S}$$

a la observación genérica de la variable transformada.

Aplicando el resultado de **1.8** a las constantes $a = \frac{1}{S}$ y $b = \frac{-\overline{x}}{S}$, la media de la variable tipificada es

$$\overline{z} = \frac{1}{S} \cdot \overline{x} + \left(-\frac{\overline{x}}{S}\right) = 0.$$

Análogamente, mediante aplicación de **1.22**, la varianza de la nueva distribución es

$$S_Z^2 = \left(\frac{1}{S}\right)^2 S^2 = \frac{S^2}{S^2} = 1.$$

En conclusión, la distribución tipificada tiene media cero y varianza uno.

Podríamos haber llegado al mismo resultado considerando que la transformación anterior es un cambio de origen y de escala donde $e = S$ y $o = \bar{x}$.

1.25 Un almacén farmacéutico se compone de dos secciones: perfumería y farmacia. En 2003, la distribución de ingresos mensuales, en miles de euros, de la sección de perfumería tuvo una media de 150 y una desviación típica de 9, siendo estas medidas 450 y 20 para la distribución de ingresos mensuales en la sección de farmacia.

En el mes de agosto de dicho año se obtuvieron unos ingresos de 500 mil euros en la sección de farmacia y de 160 mil euros en la de perfumería. ¿Cuál de estos valores es relativamente mayor en comparación con el resto del año?

SOLUCIÓN

La información proporcionada por el enunciado aparece resumida en la siguiente tabla:

	Ingresos medios	Desviación típica ingresos	Ingresos agosto
Farmacia	450	20	500
Perfumería	150	9	160

Estos dos valores, 160 y 500, pertenecen a distribuciones distintas, con lo cual, para poder compararlos hemos de homogeneizarlos, eliminando la influencia de sus correspondiente unidades de medida. A partir de una distribución de frecuencias, el proceso denominado *tipificación* permite, mediante un cambio de origen y de escala, obtener una distribución transformada desprovista de tal influencia; los valores de esta distribución transformada, que, según hemos demostrado en **1.24**, tiene media 0 y desviación típica 1, podrán ser comparados con los valores de otras distribuciones *tipificadas*.

Procedamos, entonces, a *normalizar* estos dos valores correspondientes al mes de agosto, obteniendo, así, los siguientes valores *homogéneos* denominados valores tipificados:

$$z_1 = \frac{500 - 450}{20} = 2{,}5,$$

para la sección de farmacia, y

$$z_2 = \frac{160 - 150}{9} = 1{,}11,$$

para la sección de perfumería.

Se concluye, de este modo, que los mayores ingresos en el mes de agosto, en términos relativos, han correspondido a la sección de farmacia.

1.26 Analícese el efecto que produce una transformación lineal sobre el coeficiente de variación de Pearson de una distribución de frecuencias, $(x_i; f_i)$. Aplíquese el resultado al caso particular de un cambio de origen y de escala.

SOLUCIÓN

Sea V el coeficiente de variación de Pearson de la distribución de frecuencias $(x_i; f_i)$,

$$\mathrm{V} = \frac{S}{\bar{x}},$$

donde $\bar{x}$ y S son, respectivamente, la media y la desviación típica de la distribución.

Como el lector sabe por resultados anteriores, la distribución resultante de una transformación lineal, $(a \cdot x_i + b; f_i)$, tiene por media y por desviación típica:

$$a \cdot \bar{x} + b$$

y

$$|a| \cdot S,$$

respectivamente.

Por tanto, el coeficiente de variación de Pearson de la nueva distribución es

$$\frac{|a| \cdot S}{a \cdot \bar{x} + b}.$$

Téngase en cuenta que, si b es igual a cero y a es un número positivo, la distribución transformada tiene un coeficiente de variación igual al de la distribución de partida, pues

$$\frac{|a| \cdot S}{a \cdot \bar{x}} = \frac{a \cdot S}{a \cdot \bar{x}} = \frac{S}{\bar{x}}.$$

Además, si b es cero y a es menor que cero, el coeficiente de variación de la nueva distribución será

$$V' = \frac{|a| \cdot S}{a \cdot \bar{x}} = \frac{-a \cdot S}{a \cdot \bar{x}} = -\frac{S}{\bar{x}} = -V.$$

Ahora bien, puesto que la interpretación del coeficiente de variación es en valor absoluto, al ser $|V'| = |-V| = |V|$, puede afirmarse que la distribución inicial y la distribución transformada tienen idéntica dispersión relativa.

Particularmente, si a es igual a -1, es decir, si comparamos las dispersiones con respecto a sus correspondientes medias en $(x_i; f_i)$ y $(-x_i; f_i)$, concluiremos que ambas tienen la misma, hecho que, por otro lado, es intuitivamente claro: si todos los valores de la distribución cambian de signo, siendo iguales las frecuencias, seguirán manteniéndose las posiciones relativas de cada valor con respecto a su media aritmética, o, equivalentemente, el grado de homogeneidad de ambas distribuciones será el mismo. Obsérvese, además, que en este caso también coinciden las varianzas de las dos distribuciones.

Un caso particular del planteado en este ejercicio se obtiene si $a = 1/e$ y $b = -o/e$, $(e > 0)$, esto es, si la transformación lineal es un cambio de origen y de escala. En tal caso, el coeficiente de variación de la distribución transformada resultará ser

$$\frac{\left|\frac{1}{e}\right| \cdot S}{\frac{1}{e} \cdot \bar{x} - \frac{o}{e}} = \frac{\frac{1}{e} \cdot S}{\frac{1}{e}(\bar{x} - o)} = \frac{S}{\bar{x} - o},$$

ya que $e > 0$ implica que $|1/e| = 1/e$.

En consecuencia, el coeficiente de variación de Pearson de una distribución se ve afectado únicamente por cambios de origen.

1.27 Una empresa desea contratar un peón para pintar la raya divisoria entre carriles de una nueva carretera. Para ello pone a prueba a dos personas durante 5 días, con los siguientes resultados correspondientes al número de kilómetros pintados por cada uno de ellos:

Días	1	2	3	4	5
Peón A	6	2	11	0	6
Peón B	5	3	5	4	3

Aunque el número medio de kilómetros pintados por día es 5 para el peón A y 4 para el peón B, la empresa, aplicando el criterio de constancia y homogeneidad en el trabajo, contrata al operario B. Justifíquese estadísticamente dicha decisión.

SOLUCIÓN

Desde el punto de vista estadístico, hablar de mayor homogeneidad significa hablar de menor dispersión. Por ello, utilizando el coeficiente de variación de Pearson, medida de dispersión relativa, se podrán comparar las dispersiones de ambas distribuciones, tomando como referencia la media aritmética.

Teniendo en cuenta que $\bar{x}_A$ y $\bar{x}_B$, valores medios de cada distribución, son iguales a 5 y 4, respectivamente, la varianza de la distribución del peón A es

$$S_A^2 = \frac{1}{5}(6^2 + 2^2 + 11^2 + 0^2 + 6^2) - 5^2 = 14{,}4$$

y la del peón B,

$$S_B^2 = \frac{1}{5}(5^2 + 3^2 + 5^2 + 4^2 + 3^2) - 4^2 = 0{,}8,$$

con lo cual, las desviaciones típicas, raíces cuadradas de las cantidades anteriores, son

$$S_A = 3{,}79$$

y

$$S_B = 0{,}89.$$

Por consiguiente, los coeficientes de variación son, respectivamente,

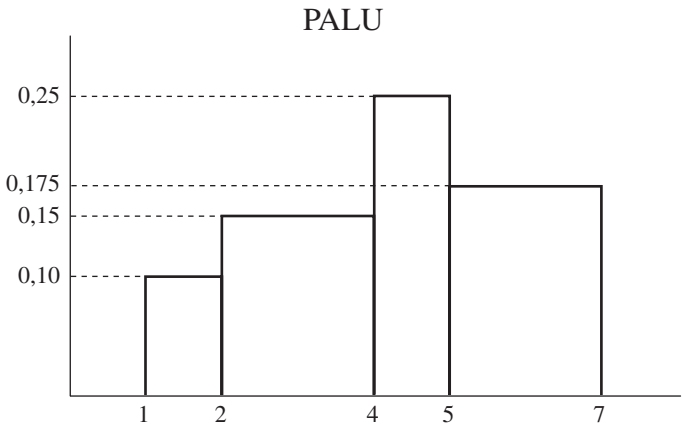
$$\mathrm{V}_A = \frac{S_A}{\bar{x}_A} = \frac{3{,}79}{5} = 0{,}76$$

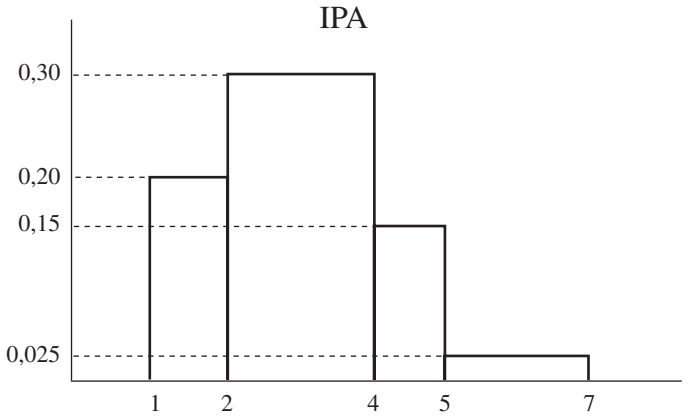
y

$$\mathrm{V}_B = \frac{S_B}{\bar{x}_B} = \frac{0{,}89}{4} = 0{,}22.$$

El resultado obtenido justifica estadísticamente la decisión de la empresa, puesto que la distribución correspondiente al peón B es más homogénea, como indica el menor valor de su coeficiente de variación. En concreto, el coeficiente de variación de la primera distribución nos dice que la dispersión de los datos en torno a su media representa un 76 por ciento de ésta, siendo el porcentaje de dispersión de las observaciones de la segunda distribución con respecto a su media tan solo del 22 por ciento del valor de dicha media.

1.28 Dos cadenas de alimentación, PALU e IPA, tienen instalados supermercados a lo largo de todo el territorio nacional. Los siguientes histogramas representan las distribuciones de frecuencias de los beneficios mensuales, en miles de euros, correspondientes a los supermercados de ambas cadenas.





a) ¿Cuál de los dos grupos de supermercados presenta unos beneficios mensuales más homogéneos?

b) ¿Qué cadena tiene un mayor porcentaje de supermercados con beneficios entre 4 y 5 mil euros?

SOLUCIÓN

a) Del histograma correspondiente a la distribución de beneficios mensuales de la cadena PALU se obtienen los datos que aparecen en la tabla siguiente:

Ingresos PALU	x_i	d_i	f_i
1-2	1,5	0,10	0,10
2-4	3,0	0,15	0,30
4-5	4,5	0,25	0,25
5-7	6,0	0,175	0,35

Las tres primeras columnas contienen los intervalos, las marcas de clase y las densidades de frecuencia, esto es, las alturas, de los histogramas.

Obsérvese que cada cantidad de la última columna, f_i, se calcula a partir la densidad de frecuencia de cada intervalo, d_i, y de la longitud de cada rectángulo, c_i, según la relación:

$$f_i = d_i \cdot c_i.$$

Los datos de la tabla permiten hallar la media aritmética de la distribución de beneficios mensuales de la cadena PALU,

$$\bar{x}_P = \sum_{i=1}^{h} x_i \cdot f_i = 1{,}5 \cdot 0{,}10 + 3 \cdot 0{,}30 + 4{,}5 \cdot 0{,}25 + 6 \cdot 0{,}35 = 4{,}275 \text{ miles de euros,}$$

la varianza,

$$S_P^2 = \sum_{i=1}^{h} x_i^2 \cdot f_i \; - \bar{x}_P^2 = 1{,}5^2 \cdot 0{,}10 + 3^2 \cdot 0{,}30 + 4{,}5^2 \cdot 0{,}25 + 6^2 \cdot 0{,}35 \; - 4{,}275^2 = 2{,}31,$$

y, en consecuencia, la desviación típica, raíz cuadrada de la varianza,

$$S_P = 1{,}52.$$

Con estas características se halla el coeficiente de variación de Pearson de la distribución de beneficios de la cadena PALU, que utilizaremos para comparar el grado de homogeneidad de las distribuciones de beneficios de ambas cadenas:

$$\mathrm{V}_P = \frac{S_P}{\bar{x}_P} = \frac{1{,}52}{4{,}275} = 0{,}3555,$$

es decir, la dispersión de los valores de la distribución con respecto a su media representa un 35,55 por ciento de dicha media.

Por los que respecta a la distribución de beneficios mensuales de la cadena IPA, en la siguiente tabla se recoge la información que proporciona el histograma de frecuencias:

Ingresos IPA	y_j	d_j	f_j
1-2	1,5	0,20	0,20
2-4	3,0	0,30	0,60
4-5	4,5	0,15	0,15
5-7	6,0	0,025	0,05

La media de beneficios, en miles de euros, de la segunda distribución es

$$\bar{y}_I = \sum_{j=1}^{k} y_j \cdot f_j = 1{,}5 \cdot 0{,}20 + 3 \cdot 0{,}60 + 4{,}5 \cdot 0{,}15 + 6 \cdot 0{,}05 = 3{,}075,$$

siendo la varianza

$$S_I^2 = \sum_{j=1}^{k} y_j^2 \cdot f_j - \overline{y}_I^2 = 1{,}5^2 \cdot 0{,}20 + 3^2 \cdot 0{,}60 + 4{,}5^2 \cdot 0{,}15 + 6^2 \cdot 0{,}05 - 3{,}075^2 = 1{,}23,$$

y la desviación típica

$$S_I = 1{,}11.$$

En consecuencia, el coeficiente de variación de Pearson de la distribución de beneficios de la cadena IPA es

$$V_I = \frac{S_I}{\overline{y}_I} = \frac{1{,}11}{3{,}075} = 0{,}3609.$$

Se concluye, así, que la distribución de beneficios de la cadena de supermercados PALU es ligeramente más homogénea por ser algo menor su coeficiente de variación.

b) Puesto que la frecuencia relativa del intervalo 4-5 es 0,15, ello significa que el 15 por ciento de los supermercados de la cadena IPA tienen beneficios entre 4 y 5 mil euros, mientras que un 25 por ciento de los supermercados de la cadena PALU tienen beneficios en dicho intervalo, ya que la frecuencia relativa del intervalo 4-5 en la distribución de frecuencias de los beneficios de esta cadena es 0,25.

1.29 La siguiente tabla recoge los ingresos medios durante 2003, en miles de euros, y la desviación típica de las doscientas empresas que una multinacional posee en América, Asia y Europa.

	N.º empresas	Ingresos medios	Desviación típica
América	20	330	70
Asia	50	165	22,5
Europa	130	256	42

En 2004 cada una de las veinte empresas de América incrementó sus ingresos en un 5 por ciento, siendo este incremento de 15 mil euros en las empresas ubicadas en Asia y manteniéndose constantes los ingresos de las que están en Europa.

a) ¿En qué continente fue más homogénea la distribución de los ingresos en 2004?

b) Los mayores ingresos en 2004 han correspondido en América a una empresa con 361,2 miles de euros, en Asia a una empresa con 191,25 miles de euros y en Europa a una empresa con 293,8 miles de euros. ¿Qué empresa ha obtenido mayores ingresos en términos relativos?

SOLUCIÓN

a) Puesto que tenemos información sobre medias y desviaciones típicas de 2003, parece razonable realizar la comparación de la dispersión mediante el coeficiente de variación de Pearson, aunque para ello será necesario disponer de la media y la desviación típica de las distribuciones de ingresos de cada continente en 2004.

Ahora bien, si $(x_i; f_i)$, $(y_j; f_j)$ y $(u_l; f_l)$ son las distribuciones de ingresos de cada continente en 2003, la información del enunciado permite conocer que, en 2004, la distribución de ingresos en América, con un incremento de los valores del 5 por ciento, se transforma en la distribución $(1{,}05 \cdot x_i; f_i)$; la distribución de ingresos en Asia, con un aumento lineal de 15 mil euros, se convierte en $(y_j + 15; f_j)$; manteniéndose constante, es decir, igual a $(u_l; f_l)$ la distribución de ingresos de las empresas europeas.

Las conocidas propiedades de la media y la desviación típica de una distribución permiten la obtención de estas características en cada una de las distribuciones de ingresos en 2004, a partir de las correspondientes al año anterior, según se recoge en la tabla siguiente:

	Ingresos medios	Desviación típica
América	$1{,}05 \cdot 330 = 346{,}5$	$1{,}05 \cdot 70 = 73{,}5$
Asia	$165+15 = 180$	22,5
Europa	256	42

En consecuencia, el coeficiente de variación de Pearson de cada distribución es

$$V_{AMÉRICA} = \frac{73{,}5}{346{,}5} = 0{,}212,$$

$$V_{ASIA} = \frac{22{,}5}{180} = 0{,}125$$

y

$$V_{EUROPA} = \frac{42}{256} = 0{,}164,$$

concluyéndose que la distribución de los ingresos en Asia es más homogénea, al ser menor su coeficiente de variación: la dispersión de los valores de la distribución en torno a su media representa un 12,5 por ciento del valor de esta en dicha distribución, siendo los correspondientes porcentajes en América y Europa del 21.2 y del 16.4, respectivamente.

b) Los valores tipificados de los ingresos de las *mejores* empresas de cada continente,

$$z_{AMÉRICA} = \frac{361{,}2 - 346{,}5}{73{,}5} = 0{,}2,$$

$$z_{ASIA} = \frac{191{,}25 - 180}{22{,}5} = 0{,}5$$

y

$$z_{EUROPA} = \frac{293{,}8 - 256}{42} = 0{,}9,$$

muestran que los mayores ingresos, en términos relativos, corresponden a la empresa europea, con un valor tipificado mayor que el resto.

1.30 Una empresa dedicada a la producción de piezas para coches desea adquirir una máquina para la fabricación de cubiertas. El proveedor le ofrece la posibilidad de elegir entre dos tipos de máquinas.

De una muestra seleccionada para cada uno de los tipos de máquinas se sabe que la distribución del número de unidades producidas diariamente tiene una media de 120 y una desviación típica de 7 para las máquinas de tipo A, mientras que estos valores son 100 y 5 para las máquinas del tipo B.

Además, las unidades fabricadas al día por una máquina del tipo A tienen el siguiente coste, en euros,

$$C_A = 60 \cdot X,$$

siendo el coste diario de producción en una máquina del tipo B:

$$C_B = 50 \cdot Y + 10,$$

donde X e Y representan, respectivamente, el número de unidades producidas al día por una máquina del tipo A y por una del tipo B.

Si el criterio de decisión de la empresa se basa en la mayor homogeneidad en el coste diario de producción, ¿qué tipo de máquina deberá comprar?

SOLUCIÓN

La media de producción diaria, de la máquina A, $\bar{x}$, es de 120 unidades, siendo el número medio diario de unidades producidas por la máquina B, $\bar{y}$, igual a 100. Se sabe, también, que la desviación típica de la distribución de unidades producidas por la máquina A, S_X, es de 7 unidades y la que corresponde a la máquina B, S_Y, es de 5 unidades.

A partir de aquí, y como el coste de la producción de la máquina A es

$$C_A = 60 \cdot X,$$

es posible calcular el coste medio y la desviación típica del coste de producción de la máquina, aplicando las propiedades de la media y la desviación típica:

$$\overline{c}_A = 60 \cdot \overline{x} = 60 \cdot 120 = 7\,200 \text{ euros}$$

y

$$S_{C_A} = 60 \cdot S_X = 60 \cdot 7 = 420.$$

De igual modo, el coste de producción de la máquina B es

$$C_B = 50 \cdot Y + 10,$$

con lo que, la media y la desviación típica de C_B son, respectivamente,

$$\overline{c}_B = 50 \cdot \overline{y} + 10 = 50 \cdot 100 + 10 = 5\,010 \text{ euros}$$

y

$$S_{C_B} = 50 \cdot S_Y = 50 \cdot 5 = 250.$$

Con los datos obtenidos se halla el coeficiente de variación de Pearson de la distribución del coste de producción para cada una de las máquinas:

$$\mathrm{V}_{C_A} = \frac{S_{C_A}}{\overline{c}_A} = \frac{420}{7\,200} = 0{,}0583$$

y

$$\mathrm{V}_{C_B} = \frac{S_{C_B}}{\overline{c}_B} = \frac{250}{5\,010} = 0{,}0499,$$

pudiendo afirmarse que, según el criterio de mayor homogeneidad en el coste diario de producción, debería comprarse la máquina B, a la cual corresponde un coeficiente de variación algo menor.

Proponemos al lector que resuelva este ejercicio aplicando el resultado **1.26**.

1.31 Las distribuciones mensuales de los salarios, en miles de euros, de dos empresa, Micovusa y Nossan, dedicadas a la fabricación de piezas para coches, son $(x_i; f_i)$ e $(y_j; f_j)$, respectivamente, con S desviación típica común a ambas.

El salario mensual del señor Pórrez, gerente de Micovusa, es de 30 mil euros y el de la señora Fuji, gerente de Nossan, de 36 mil. Para conocer cuál de los dos salarios es mayor en relación con el resto de los trabajadores de sus respectivas empresas, se han tipificado dichos salarios, obteniéndose el mismo valor en ambos casos.

a) Interprétese el resultado.

b) Sabiendo que, en términos relativos, la dispersión de los salarios de la empresa Micovusa con respecto a su media es el doble que la dispersión de los salarios en Nossan en relación a la suya, ¿cuáles son los salarios medios de cada una de las empresas?

SOLUCIÓN

a) Al tipificar los salarios de ambos gerentes el resultado es idéntico, por lo que podemos afirmar que ambos tienen el mismo salario *en relación* con el resto de los trabajadores de su empresa.

b) Como el resultado de tipificar los salarios de los dos gerentes ha sido el mismo, y puesto que $S_X = S_Y = S$, se tiene que

$$\frac{x_i - \overline{x}}{S} = \frac{y_j - \overline{y}}{S},$$

donde x_i e y_j son, respectivamente, los salarios del señor Pórrez y de la señora Fuji.

En consecuencia, simplificando,

$$x_i - \overline{x} = y_j - \overline{y},$$

esto es,

$$30 - \overline{x} = 36 - \overline{y},$$

y, por tanto, se obtiene la siguiente relación entre las medias de ambas distribuciones:

$$\overline{x} = \overline{y} - 6.$$

Como, además, la dispersión relativa de los salarios en Micovusa es doble que la dispersión de los salarios en Nossan, resulta, igualmente, una relación entre los respectivos coeficientes de variación:

$$\mathrm{V}_X = 2 \cdot \mathrm{V}_Y.$$

Por consiguiente, sustituyendo las expresiones de estos coeficientes, se tiene que

$$\frac{S}{\overline{x}} = 2 \cdot \frac{S}{\overline{y}},$$

con lo cual

$$\overline{y} = 2 \cdot \overline{x}.$$

En definitiva, por un lado,

$$\overline{x} = \overline{y} - 6,$$

y, por otro,

$$\overline{y} = 2 \cdot \overline{x},$$

de lo que se concluye, despejando, que

$$\bar{x} = 6 \text{ mil euros}$$

y

$$\bar{y} = 12 \text{ mil euros}$$

son los valores medios de cada distribución, es decir, los salarios medios de cada empresa.

1.32 Un restaurante ofrece a sus clientes tres tipos diferentes de platos combinados, enumerados del I al III. Los precios de cada menú, en euros, así como los ingresos obtenidos el domingo pasado por la venta de cada uno de ellos, se reflejan en la siguiente tabla:

Tipo de menú	Precio	Ingresos
I	6,5	520
II	8,0	280
III	9,0	324

a) Hállese el precio medio por menú.

b) Calcúlese la dispersión relativa de la distribución del precio por menú.

SOLUCIÓN

a) Para calcular el promedio de la magnitud relativa *precio por menú* hay que hallar la media armónica de su distribución:

$$\mathrm{H} = \frac{N}{\sum_{i=1}^{h} \frac{n_i}{x_i}} = \frac{520 + 280 + 324}{\frac{520}{6,5} + \frac{280}{8} + \frac{324}{9}} = 7,44 \text{ euros.}$$

b) Calculemos, en primer lugar, la *desviación cuadrática media con respecto a la media armónica*:

$$D_{\mathrm{H}}^2 = \frac{1}{N} \sum_{i=1}^{h} (x_i - \mathrm{H})^2 \, n_i,$$

que, para el caso que nos ocupa, es

$$D_{\mathrm{H}}^2 = \frac{(6,5 - 7,44)^2 \, 520 + (8 - 7,44)^2 \, 280 + (9 - 7,44)^2 \, 324}{1\,124} = 1,19.$$

En definitiva, la dispersión relativa, esto es, el coeficiente de variación respecto a la media armónica, es

$$V_H = \frac{\sqrt{D_H^2}}{H} = \frac{1{,}09}{7{,}44} = 0{,}15,$$

con lo cual, la dispersión de los valores de la distribución del precio por menú con relación a su media representa un 15 por ciento del valor de dicha media.

1.33 Indíquese si las siguientes afirmaciones son verdaderas o falsas:

a) El coeficiente de asimetría de Fisher no varía si todos los valores de la distribución se multiplican por una constante.

b) Si a cada valor de una distribución asimétrica negativa se le suma una constante, k, siendo k un valor mayor que la media aritmética de dicha distribución, ésta pasará a ser una distribución asimétrica positiva.

SOLUCIÓN

a) El coeficiente de asimetría de Fisher de la distribución $(x_i; f_i)$ es

$$g_1 = \frac{\sum_{i=1}^{h} (x_i - \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i\right]^{3/2}}.$$

El coeficiente de asimetría de la distribución transformada, $(k \cdot x_i; f_i)$, con k valor constante, es igual a

$$g_1' = \frac{\sum_{i=1}^{h} (k \cdot x_i - k \cdot \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} (k \cdot x_i - k \cdot \bar{x})^2 f_i\right]^{3/2}},$$

puesto que, como es sabido, la media de la distribución transformada es $k \cdot \bar{x}$.

Sacando factor común a k^3 en el numerador y a k en el denominador, y teniendo en cuenta que se trata de un valor constante que, por tanto, puede escribirse fuera del sumatorio, resulta:

$$g_1' = \frac{\sum_{i=1}^{h} k^3 (x_i - \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} k^2 (x_i - \bar{x})^2 f_i\right]^{3/2}} = \frac{k^3 \sum_{i=1}^{h} (x_i - \bar{x})^3 f_i}{k^3 \left[\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i\right]^{3/2}} = \frac{\sum_{i=1}^{h} (x_i - \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i\right]^{3/2}} = g_1.$$

En consecuencia, la afirmación del apartado ***a)*** es *verdadera*.

b) El coeficiente de asimetría de la distribución $(x_i + k; f_i)$, con k una constante cualquiera, es

$$g_1' = \frac{\sum_{i=1}^{h} [(x_i + k) - (\bar{x} + k)]^3 f_i}{\left[\sum_{i=1}^{h} [(x_i + k) - (\bar{x} + k)]^2 f_i\right]^{3/2}},$$

ya que la media de la distribución transformada es igual a $\bar{x} + k$. Por tanto, operando, resulta que

$$g_1' = \frac{\sum_{i=1}^{h} (x_i - \bar{x})^3 f_i}{\left[\sum_{i=1}^{h} (x_i - \bar{x})^2 f_i\right]^{3/2}} = g_1.$$

En definitiva, el coeficiente de asimetría es invariante ante este tipo de transformaciones, sea cual sea el valor de la constante k, siendo, consecuentemente, *falsa* la afirmación de este apartado.

1.34 En la siguiente tabla se recoge la distribución de frecuencias del número de unidades diarias de un producto vendidas durante el pasado mes:

x_i	0	1	2	3	4
n_i	7	8	9	4	2

a) Hállese la media, la mediana y la moda de esta distribución.

b) Obténgase la varianza, la desviación típica, el coeficiente de variación de Pearson, la desviación cuadrática media con respecto a la mediana y el coeficiente de variación con respecto a la mediana.

c) Calcúlese los coeficientes de asimetría y de curtosis.

SOLUCIÓN

a) Sirva este sencillo ejercicio para fijar ideas sobre el cálculo de las principales características de una distribución de frecuencias.

Por definición de media aritmética de una distribución, se tiene que

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_i = \frac{0 \cdot 7 + 1 \cdot 8 + 2 \cdot 9 + 3 \cdot 4 + 4 \cdot 2}{30} = 1{,}53 \text{ unidades.}$$

Para el cálculo de la mediana completamos la tabla del enunciado con la fila correspondiente a las frecuencias absolutas acumuladas:

x_i	0	1	2	3	4
n_i	7	8	9	4	2
N_i	7	15	24	28	30

Obsérvese que, en este caso, existe un valor de la variable, $x_2 = 1$, tal que su frecuencia absoluta acumulada coincide con $N/2 = 15$. La mediana es, por tanto, igual al punto medio entre ese valor y el siguiente:

$$\frac{x_2 + x_3}{2} = \frac{1 + 2}{2} = 1{,}5 \text{ unidades.}$$

Por último, la moda, valor de la variable con mayor frecuencia, es, en este caso, el valor $x_3 = 2$.

b) A partir de la varianza,

$$S^2 = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_i - \bar{x}^2 = \frac{0^2 \cdot 7 + 1^2 \cdot 8 + 2^2 \cdot 9 + 3^2 \cdot 4 + 4^2 \cdot 2}{30} - 1{,}53^2 = 1{,}39,$$

se calculan, tanto la desviación típica,

$$S = \sqrt{S^2} = 1{,}18,$$

como el coeficiente de variación de Pearson,

$$\mathrm{V} = \frac{S}{\bar{x}} = \frac{1{,}18}{1{,}53} = 0{,}77.$$

La desviación cuadrática respecto a la mediana es

$$D_{Me}^2 = \frac{1}{N}\sum_{i=1}^{h} (x_i - Me)^2 n_i,$$

con lo cual, para los datos del enunciado, se tiene que

$$D^2_{Me} = \frac{1}{30}\left[(0-1{,}5)^2 \cdot 7 + (1-1{,}5)^2 \cdot 8 + (2-1{,}5)^2 \cdot 9 + (3-1{,}5)^2 \cdot 4 + (4-1{,}5)^2 \cdot 2\right] = 1{,}38.$$

En definitiva, el coeficiente de variación de la mediana es

$$V_{Me} = \frac{\sqrt{D^2_{Me}}}{Me} = 0{,}78.$$

c) El coeficiente de asimetría responde a la expresión:

$$g_1 = \frac{m_3}{S^3},$$

donde

$$m_3 = \frac{1}{N}\sum_{i=1}^{h}(x_i - \bar{x})^3 n_i$$

es el momento de orden 3 respecto a la media.

En cuanto al coeficiente de curtosis,

$$g_2 = \frac{m_4}{S^4} - 3,$$

en su expresión aparece el momento respecto a la media de orden 4,

$$m_4 = \frac{1}{N}\sum_{i=1}^{h}(x_i - \bar{x})^4 n_i.$$

Los datos de la siguiente tabla servirán de apoyo en el cálculo de los coeficientes pedidos:

x_i	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
0	– 3,58	5,48
1	– 0,15	0,08
2	0,10	0,05
3	3,18	4,67
4	15,07	37,22

Así,

$$m_3 = \frac{1}{N}\sum_{i=1}^{h}(x_i - \bar{x})^3 n_i = \frac{(-3{,}58)\cdot 7 + (-0{,}15)\cdot 8 + 0{,}1 \cdot 9 + 3{,}18 \cdot 4 + 15{,}07 \cdot 2}{30} = 0{,}58$$

y, puesto que

$$S^3 = 1{,}18^3 = 1{,}64,$$

se tiene que el coeficiente de asimetría es

$$g_1 = \frac{m_3}{S^3} = \frac{0{,}58}{1{,}64} = 0{,}35.$$

Análogamente,

$$m_4 = \frac{1}{N} \sum_{i=1}^{h} (x_i - \bar{x})^4 \, n_i = \frac{5{,}48 \cdot 7 + 0{,}08 \cdot 8 + 0{,}05 \cdot 9 + 4{,}67 \cdot 4 + 37{,}22 \cdot 2}{30} = 4{,}42$$

y

$$S^4 = 1{,}18^4 = 1{,}94,$$

por lo que el coeficiente de curtosis resulta ser

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{4{,}42}{1{,}94} - 3 = -0{,}72.$$

1.35 De la distribución $(x_i; f_i)$, se obtiene la distribución $(y_i; f_i)$, mediante cambio de variable en los valores de la primera distribución. Obténgase la expresión de y_i en función de x_i, sabiendo que la media de la distribución transformada es igual al momento respecto al origen de orden 2 de la distribución inicial.

SOLUCIÓN

Como, según el enunciado,

$$\bar{y} = a_2 (x),$$

sustituyendo, se tiene que

$$\sum_{i=1}^{h} y_i \cdot f_i = \sum_{i=1}^{h} x_i^2 \cdot f_i,$$

con lo cual, identificando términos, resulta la relación:

$$y_i = x_i^2.$$

1.36 Se considera la variable X en las unidades de una población dividida en L partes o *estratos*. Hállese la expresión de la media de X en función de los valores medios de la variable en cada estrato.

SOLUCIÓN

Denotemos por x_{ih} el valor de la observación i-ésima de la variable en el estrato h-ésimo y sea N_h el tamaño del estrato h-ésimo. Por definición, la media aritmética de la variable es igual a la suma de todas las observaciones de la variable dividida por el número de ellas, N. Así, ordenando éstas según el estrato al que pertenecen y agrupando en sumatorios las observaciones que están en el mismo estrato, se obtiene que

$$\bar{x} = \frac{(x_{11} + \ldots + x_{N_1 1}) + \ldots + (x_{1L} + \ldots + x_{N_L L})}{N} = \frac{\sum_{i=1}^{N_1} x_{i1} + \ldots + \sum_{i=1}^{N_L} x_{iL}}{N} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{x_{ih}}{N}.$$

Nótese que el valor medio se calcula con dos sumatorios, ya que cada observación consta de dos subíndices. Ahora bien, este valor medio puede también expresarse, según se verá a continuación, a partir de la media de la variable en cada estrato. Así, puesto que la media correspondiente al estrato h-ésimo es igual a la suma de las observaciones de dicho estrato entre el número de ellas,

$$\bar{x}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{ih},$$

multiplicando y dividiendo por N_h la media $\bar{x}$, se tiene que

$$\bar{x} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{N_h}{N_h} \cdot \frac{x_{ih}}{N} = \sum_{h=1}^{L} \frac{N_h}{N} \sum_{i=1}^{N_h} \frac{x_{ih}}{N_h} = \sum_{h=1}^{L} \frac{N_h}{N} \cdot \bar{x}_h = \sum_{h=1}^{L} W_h \cdot \bar{x}_h,$$

donde, para $h = 1, \ldots, L$,

$$W_h = \frac{N_h}{N}$$

es el *peso* o *ponderación* del estrato h-ésimo.

1.37 Un cierto producto ha estado a la venta en tres establecimientos al mismo precio por unidad en todos ellos durante un año.

En el primer cuatrimestre estuvo a la venta en el establecimiento A; el trimestre siguiente se vendió en el establecimiento B y el resto del año en el C. El número medio mensual de unidades vendidas en cada uno de ellos ha sido: 100, 200 y 125, respectivamente.

a) Obténgase el número medio mensual de unidades vendidas en el total de los establecimientos.

b) Sabiendo que el ingreso medio mensual por las ventas del producto ha sido de 7 500 euros, hállese el precio por unidad del citado producto.

SOLUCIÓN

a) El número medio mensual de unidades vendidas para el total de los establecimientos, $\bar{x}$, se halla a partir de las ventas medias mensuales en cada establecimiento, $\bar{x}_A$, $\bar{x}_B$ y $\bar{x}_C$, mediante la expresión:

$$\bar{x} = \frac{\bar{x}_A \cdot N_A + \bar{x}_B \cdot N_B + \bar{x}_C \cdot N_C}{N_A + N_B + N_C},$$

correspondiente, según vimos en **1.36**, a la media de una población dividida en estratos para el caso en el que el número de ellos sea 3.

Observe el lector que N_A, N_B y N_C son, en esta ocasión, el número de meses que el artículo ha estado a la venta en los establecimientos A, B y C, esto es, 4, 3 y 5, respectivamente.

En definitiva, utilizando los datos del enunciado, resulta que

$$\bar{x} = \frac{100 \cdot 4 + 200 \cdot 3 + 125 \cdot 5}{12} = 135{,}42 \text{ unidades.}$$

b) El ingreso mensual, Y, se relaciona con X, número de unidades vendidas al mes, mediante la expresión:

$$Y = p \cdot X,$$

donde p es el precio por unidad.

Aplicando la propiedad de la media aritmética demostrada en **1.8**, se cumple, asimismo, la relación entre las respectivas medias:

$$\bar{y} = p \cdot \bar{x}.$$

Por tanto, despejando, el precio por unidad resulta ser

$$p = \frac{\bar{y}}{\bar{x}} = \frac{7\,500}{135{,}42} = 55{,}38 \text{ euros.}$$

1.38 Dada una población dividida en L estratos, obténgase la varianza de la variable X a partir de las varianzas de la variable en cada uno de los estratos.

SOLUCIÓN

La varianza de la distribución de la variable X es la media aritmética de las desviaciones al cuadrado de cada observación con respecto a su media, $\overline{x}$. Así, si ordenamos las observaciones según el estrato al que pertenecen, tendremos que la varianza es

$$S^2 = \frac{(x_{11} - \overline{x})^2 + \ldots + (x_{N11} - \overline{x})^2 + \ldots + (x_{1L} - \overline{x})^2 + \ldots + (x_{N_LL} - \overline{x})^2}{N}.$$

Agrupando términos semejantes en sumatorios, se obtiene la expresión equivalente:

$$S^2 = \frac{\sum_{i=1}^{N_1}(x_{i1} - \overline{x})^2 + \ldots + \sum_{i=1}^{N_L}(x_{iL} - \overline{x})^2}{N} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(x_{ih} - \overline{x})^2}{N}.$$

Sumando y restando en la expresión anterior la media del estrato h-ésimo, $\overline{x}_h$, desarrollando el binomio y separando en tres sumandos, se tiene que

$$S^2 = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}[(x_{ih} - \overline{x}_h) + (\overline{x}_h - \overline{x})]^2}{N} =$$

$$= \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(x_{ih} - \overline{x}_h)^2}{N} + \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(\overline{x}_h - \overline{x})^2}{N} + 2 \cdot \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(x_{ih} - \overline{x}_h) \cdot (\overline{x}_h - \overline{x})}{N}.$$

Ahora bien, puesto que $(\overline{x}_h - \overline{x})$ no depende de i, el segundo sumando es igual a

$$\frac{\sum_{h=1}^{L} N_h (\overline{x}_h - \overline{x})^2}{N}$$

y el tercer sumando es

$$2 \cdot \frac{\sum_{h=1}^{L}(\overline{x}_h - \overline{x})\sum_{i=1}^{N_h}(x_{ih} - \overline{x}_h)}{N} = 0,$$

pues la suma de las desviaciones de las observaciones del estrato h-ésimo con respecto a su media, $\sum_{i=1}^{N_h} (x_{ih} - \overline{x}_h)$, es igual a cero.

Por tanto,

$$S^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} (x_{ih} - \overline{x}_h)^2}{N} + \frac{\sum_{h=1}^{L} N_h(\overline{x}_h - \overline{x})^2}{N}.$$

Ahora bien, como la varianza del estrato h-ésimo es

$$S_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (x_{ih} - \overline{x}_h)^2,$$

entonces, despejando, se tiene que

$$\sum_{i=1}^{N_h} (x_{ih} - \overline{x}_h)^2 = N_h \cdot S_h^2$$

y, por tanto, sustituyendo en el primer sumando:

$$S^2 = \frac{\sum_{h=1}^{L} N_h \cdot S_h^2}{N} + \frac{\sum_{h=1}^{L} N_h(\overline{x}_h - \overline{x})^2}{N}.$$

En definitiva, la varianza de la variable X es

$$S^2 = \sum_{h=1}^{L} W_h \cdot S_h^2 + \sum_{h=1}^{L} W_h \ (\overline{x}_h - \overline{x})^2,$$

donde $W_h = N_h/N$, es, según vimos en **1.36**, la ponderación del estrato h-ésimo.

1.39 El Tour Operador de circuitos por Europa Eurovacaciones organizó, durante el pasado año, viajes con tres destinos diferentes: París, Roma y Londres. A París hicieron un total de 100 viajes, a Roma 150 y a Londres 250. Se da la circunstancia de que la media de ingresos por viaje coincide en los tres itinerarios, siendo sus desviaciones típicas 20, 30 y 40 mil euros, respectivamente.

a) ¿En cuál de los tres destinos la media de ingresos por viaje es más representativa?

b) Calcúlese la varianza de la distribución de ingresos obtenidos en el total de los viajes (París, Roma y Londres) durante dicho año.

SOLUCIÓN

a) Los datos proporcionados por el enunciado se presentan, para más claridad, en la siguiente tabla:

	N.º viajes	Media ingreso	Desviación típica ingreso
París	100	$\bar{x}_P$	20
Roma	150	$\bar{x}_R$	30
Londres	250	$\bar{x}_L$	40

Puesto que la media de las tres distribuciones es la misma,

$$\bar{x}_P = \bar{x}_R = \bar{x}_L,$$

y las variables están expresadas en las mismas unidades de medida, a la hora de elegir la más representativa no es necesario hallar la dispersión en términos relativos de cada distribución con respecto a su media mediante el coeficiente de variación: basta con comparar las desviaciones típicas.

Consecuentemente, la menor dispersión corresponde a la distribución de ingresos por viajes a París, puesto que su desviación típica es la más pequeña.

b) La expresión de la varianza cuando la población está divida en estratos es, según se demostró en **1.38**,

$$S^2 = \frac{1}{N} \sum_{h=1}^{L} N_h \cdot S_h^2 + \frac{1}{N} \sum_{h=1}^{L} N_h \ (\bar{x}_h - \bar{x})^2.$$

El segundo sumando de esta expresión es, en la situación que nos ocupa, igual a cero, ya que las medias de cada estrato son idénticas y, por tanto, iguales a la media de la distribución de ingresos obtenidos en el total de viajes, $\bar{x}$, como puede comprobar el lector con la expresión de dicha media. Así,

$$S^2 = \frac{1}{N} \sum_{h=1}^{L} N_h \cdot S_h^2.$$

Con los datos del enunciado, la varianza pedida es

$$S^2 = \frac{100}{500} \cdot 20^2 + \frac{150}{500} \cdot 30^2 + \frac{250}{500} \cdot 40^2 = 1\,150,$$

donde 100/500, 150/500 y 250/500 son los pesos de cada uno de los estratos.

1.40 La directiva del club deportivo Cantabric, nuevo en la ciudad, contrata a tres trabajadores para la captación de socios durante un periodo de prueba de 10 días. La media diaria de clientes conseguidos es igual a 10, 20 y 50, respectivamente, para cada trabajador, siendo las correspondientes desviaciones típicas 2, 5 y 2.

a) Hállese el número medio diario de socios que se ha inscrito en el club.

b) Estúdiese la representatividad del promedio obtenido en el apartado anterior.

SOLUCIÓN

a) El número medio diario de socios inscritos en el club, $\bar{x}$, es decir, la media diaria de clientes captados por los tres trabajadores, se obtiene de la expresión:

$$\bar{x} = \frac{\bar{x}_1 \cdot N_1 + \bar{x}_2 \cdot N_2 + \bar{x}_3 \cdot N_3}{N_1 + N_2 + N_3},$$

donde N_1, N_2 y N_3, tamaños de los estratos en los que se clasifica la población, se corresponden, en este caso, con el número de días empleados por cada trabajador.

Por tanto,

$$\bar{x} = \frac{10 \cdot 10 + 20 \cdot 10 + 50 \cdot 10}{10 + 10 + 10} = 26{,}67 \text{ socios.}$$

Observe el lector que, al coincidir el tamaño de todos los estratos, es decir, el número de días empleado por cada uno de los trabajadores, la media podría haberse hallado, *en esta ocasión*, como

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3}{3},$$

porque el peso de cada estrato es el mismo.

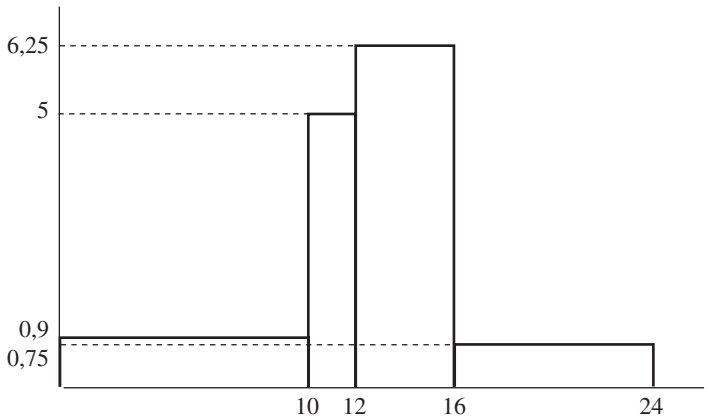
b) La representatividad del promedio se analiza, según es habitual, con una medida de dispersión como puede ser la varianza; en este caso, la varianza de una población dividida en estratos,

$$S^2 = \frac{1}{N}\sum_{h=1}^{L} N_h \cdot S_h^2 + \frac{1}{N}\sum_{h=1}^{L} N_h \ (\overline{x}_h - \overline{x})^2,$$

que, con los datos del enunciado, es igual a

$$S^2 = \frac{10}{30}(2^2 + 5^2 + 2^2) + \frac{10}{30}\left[(10 - 26{,}67)^2 + (20 - 26{,}67)^2 + (50 - 26{,}67)^2\right] = 299{,}89.$$

1.41 Una empresa dedicada a transformados metálicos cuenta con 50 trabajadores en su cadena de producción. En 2004, la distribución de la cantidad de alambre, en miles de toneladas, producida por trabajador se representa en la siguiente gráfica:



a) ¿Cuántas toneladas obtiene el 12 por ciento de los trabajadores que más producen?

b) ¿Cuál es la cantidad máxima obtenida por el 25 por ciento de los trabajadores que menos producen?

c) Calcúlese la producción media por trabajador durante dicho año.

d) La gráfica anterior se ha obtenido a partir de la siguiente estadística primaria:

x_i	4	9	10	12	13	14	15	17	19	22	24
n_i	2	3	4	10	12	10	3	1	2	2	1

Hállese a partir de esta estadística la cantidad media producida por cada trabajador y compárese el resultado con el obtenido en el apartado anterior.

SOLUCIÓN

A partir del histograma de frecuencias resulta la siguiente tabla de la distribución de la cantidad de alambre producida por trabajador.

Producción	x_i	d_i	c_i	n_i	f_i
0-10	5	0,90	10	9	0,18
10-12	11	5,00	2	10	0,20
12-16	14	6,25	4	25	0,50
16-24	20	0,75	8	6	0,12

Observe el lector que cada dato de la penúltima columna, es decir, la frecuencia absoluta o número de trabajadores de cada intervalo de producción, se ha obtenido utilizando las dos columnas anteriores:

$$n_i = d_i \cdot c_i.$$

a) El 12 por ciento de los trabajadores que más producen tienen una producción entre 16 y 24 mil toneladas, pues 0,12 es la frecuencia relativa del intervalo 16-24. En consecuencia, el número *aproximado* de toneladas que estos trabajadores producen es

$$x_4 \cdot n_4 = 20 \cdot 6 = \ 120 \text{ mil toneladas},$$

donde x_4 es la marca de clase y n_4 la frecuencia absoluta del intervalo.

b) La cantidad máxima obtenida por el 25 por ciento de los trabajadores que menos producen es el primer cuartil, C_1.

El intervalo cuartílico, $L_{i-1} - L_i$, es 10-12, pues su frecuencia absoluta acumulada, $N_2 = 9 + 10 = 19$, es la más pequeña que supera a $N/4 = 12{,}5$.

Una vez identificado el intervalo cuartílico, para obtener el cuartil C_1 aplicamos la expresión:

$$C_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \cdot c_i,$$

con lo cual, en este caso, y teniendo en cuenta que $N_{i-1} = N_1 = 9$,

$$C_1 = 10 + \frac{12{,}5 - 9}{10} \cdot 2 = 10{,}7 \text{ miles de toneladas.}$$

c) Para obtener la producción media por trabajador, esto es, la media aritmética de esta distribución de frecuencias agrupada hay que utilizar las marcas de clase de cada intervalo. Así,

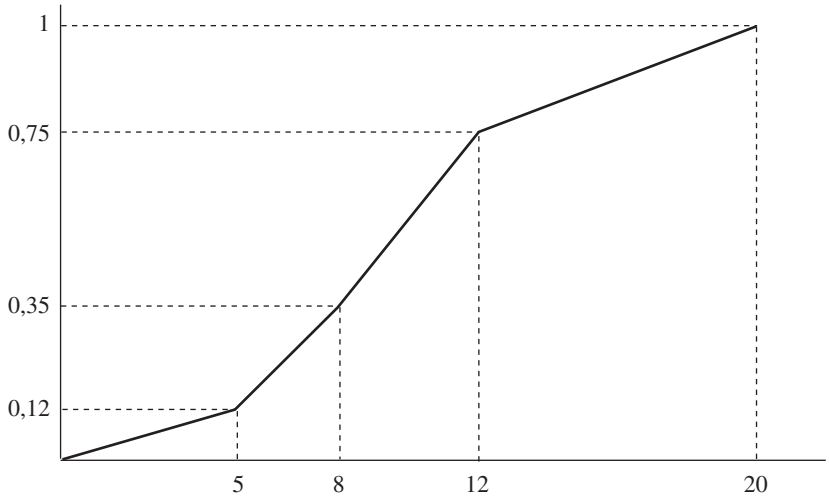
$$\bar{x} = \frac{1}{50}(5 \cdot 9 + 11 \cdot 10 + 14 \cdot 25 + 20 \cdot 6) = 12{,}5 \text{ miles de toneladas.}$$

d) Si se halla la media de producción por trabajador a partir de la estadística primaria, se obtiene:

$$\bar{x} = \frac{8 + 27 + 40 + 120 + 156 + 140 + 45 + 17 + 38 + 44 + 24}{50} = 13{,}18 \text{ miles de toneladas.}$$

El resultado no coincide con el del apartado anterior, como consecuencia de la pérdida de información originada por la agrupación de los datos en clases.

1.42 El siguiente polígono de frecuencias representa la distribución de la cantidad, en kilogramos, de carne picada que se ha vendido diariamente en una carnicería en un cierto periodo.



a) Hállese la cantidad media vendida diariamente.

b) Calcúlese la cantidad máxima de carne que se ha vendido el 42,5 por ciento de los días que menos se ha vendido.

SOLUCIÓN

El polígono de frecuencias *relativas* acumuladas aporta los datos que se detallan en la tabla siguiente correspondiente a la distribución de frecuencias de la variable:

Cantidad	x_i	F_i
0-5	2,5	0,12
5-8	6,5	0,35
8-12	10,0	0,75
12-20	16,0	1

a) A partir de las frecuencias relativas acumuladas obtenemos las frecuencias relativas ordinarias, según las relaciones conocidas:

$$f_1 = F_1$$

y, para $i = 2, 3, 4$,

$$f_i = F_i - F_{i-1},$$

con lo cual,

f_i	0,12	0,23	0,4	0,25

Estas frecuencias, junto con las marcas de clase de los intervalos, permiten hallar la media aritmética de la distribución:

$$\bar{x} = \sum_{i=1}^{h} x_i \cdot f_i = 2{,}5 \cdot 0{,}12 + 6{,}5 \cdot 0{,}23 + 10 \cdot 0{,}4 + 16 \cdot 0{,}25 = 9{,}79 \text{ kilogramos.}$$

b) Como sabemos, la información que proporciona el polígono de frecuencias está expresada en términos *relativos*. Por ello, a la hora de calcular el cuantil de orden 42,5/100 = 0,425,

$$x_{0{,}425} = L_{i-1} + \frac{0{,}425 \cdot N - N_{i-1}}{n_i} \cdot c_i,$$

lo más adecuado es transformar las frecuencias absolutas que aparecen en la expresión anterior en frecuencias relativas, dividiendo numerador y denominador del segundo sumando por N. De este modo, la expresión del cuantil de orden 0,425 se convierte en esta otra equivalente,

$$x_{0{,}425} = L_{i-1} + \frac{0{,}425 - F_{i-1}}{f_i} \cdot c_i,$$

en función, exclusivamente, de frecuencias relativas.

Teniendo en cuenta que el intervalo cuantílico es 8-12, ya que su frecuencia relativa acumulada, 0,75, es la primera estrictamente mayor que 0,425, entonces, $F_{i-1} = F_2 = 0{,}35$ y $c_i = c_3 = 4$, con lo cual,

$$x_{0{,}425} = 8 + \frac{0{,}425 - 0{,}35}{0{,}40} \cdot 4 = 8{,}75 \text{ kilogramos.}$$

1.43 El alcalde de una localidad andaluza ha decidido abonar la cantidad de 1 200 euros en concepto de ayuda al 25 por ciento de los jubilados del municipio con pensión más baja. La siguiente tabla refleja la pensión mensual, en euros, de los ancianos de la localidad:

Pensión mensual	f_i
200-400	0,10
400-600	0,15
600-1 000	0,60
1 000-1 200	0,15

a) Hállese el importe máximo mensual que deberá cobrar un pensionista para poder recibir la ayuda.

b) Calcúlese el importe que deberá consignar anualmente el Ayuntamiento para hacer frente a la deuda comprometida, suponiendo que en la localidad hay 100 ancianos.

SOLUCIÓN

a) Por definición, el primer cuartil, C_1, es el importe máximo que cobra el 25 por ciento de los pensionistas con pensión más baja.

La información que proporciona el enunciado, en términos relativos, permite obtener las frecuencias relativas acumuladas según las relaciones:

$$F_1 = f_1$$

y, para $i = 2, 3, 4$,

$$F_i = F_{i-1} + f_i,$$

con lo que

F_i	0,10	0,25	0,85	1

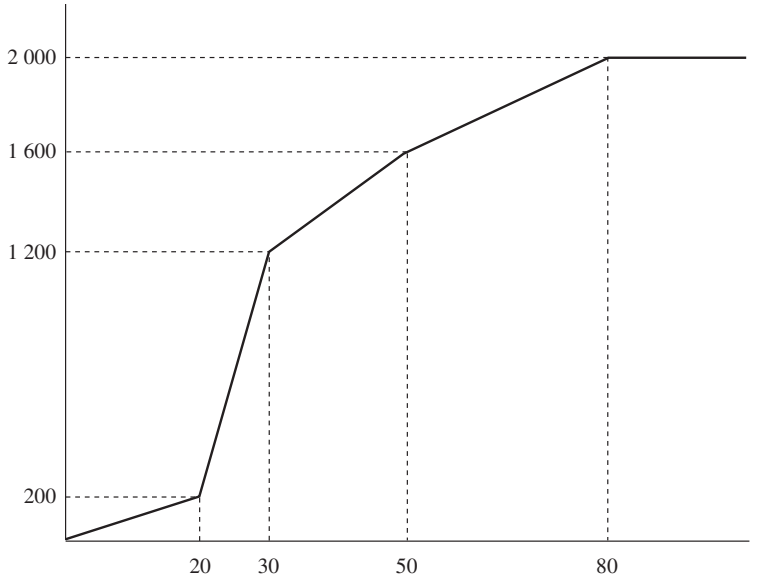
Puesto que el intervalo 400-600 tiene una frecuencia relativa acumulada $F_2 = 0{,}25$, el primer cuartil es el extremo superior de este intervalo: 600 euros.

b) El 25 por ciento de 100 es igual a 25, con lo cual, el Ayuntamiento tendrá que abonar 1 200 euros a 25 ancianos, siendo, por tanto,

$$25 \cdot 1\,200 = 30\,000 \text{ euros},$$

el importe que deberá consignar a tal efecto.

1.44 El servicio municipal de aguas de una ciudad está realizando un estudio con objeto de una posible privatización. Entre otros datos se ha obtenido que el consumo de agua, en metros cúbicos, de las 2 000 familias de dicha ciudad durante el último trimestre del año 2004 es el que se refleja en el siguiente gráfico:



a) Calcúlese la cantidad media trimestral consumida por familia.

b) Sabiendo que el precio por metro cúbico de agua es de 0,5 euros y que, además, cada trimestre se paga una cantidad fija de 2 euros por alquiler de contador y 6 euros en concepto de aguas residuales, ¿cuál ha sido el importe medio por familia abonado dicho trimestre?

c) ¿Cuál es el máximo consumo del 35 por ciento de las familias que menos consumen?

SOLUCIÓN

A partir de la representación gráfica, polígono de frecuencias acumuladas, se obtiene la siguiente tabla correspondiente a la distribución de frecuencias de la variable consumo:

Consumo	x_i	N_i	n_i
0-20	10	200	200
20-30	25	1 200	1 000
30-50	40	1 600	400
50-80	65	2 000	400

Obsérvese que la última columna de frecuencias absolutas ordinarias se ha obtenido a partir de las frecuencias acumuladas, según las relaciones:

$$n_1 = N_1$$

y

$$n_i = N_i - N_{i-1},$$

para $i = 2, 3, 4$.

a) La cantidad media consumida por familia, esto es, la media aritmética de la distribución de frecuencias, es

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_i = \frac{10 \cdot 200 + 25 \cdot 1\,000 + 40 \cdot 400 + 65 \cdot 400}{2\,000} = 34{,}5 \text{ metros cúbicos.}$$

b) La relación entre la variable consumo, X, y la variable precio, Y,

$$Y = 0{,}5 \cdot X + 2 + 6,$$

es decir,

$$Y = 0{,}5 \cdot X + 8,$$

permite conocer, también, la relación entre las medias de estas dos variables:

$$\overline{y} = 0{,}5 \cdot \overline{x} + 8.$$

Por tanto, utilizando la media calculada en el apartado anterior, el importe medio trimestral pagado por familia es

$$\overline{y} = 0{,}5 \cdot 34{,}5 + 8 = 25{,}25 \text{ euros.}$$

c) El consumo máximo del 35 por ciento de las familias que menos consumen es el percentil 35:

$$P_{35} = L_{i-1} + \frac{\frac{35 \cdot N}{100} - N_{i-1}}{n_i} \cdot c_i.$$

Teniendo en cuenta que el intervalo percentílico es 20-30, primer intervalo cuya frecuencia absoluta acumulada, $N_2 = 1\,200$, es estrictamente mayor que $\frac{35 \cdot N}{100} = 700$, se tiene que

$$P_{35} = 20 + \frac{700 - 200}{1\,000} \cdot 10 = 25 \text{ metros cúbicos.}$$

1.45 Los empleados de una empresa conservera trabajan a destajo, cobrando mensualmente una cantidad fija de 800 euros y 1,5 euros por cada mil unidades producidas. Los trabajadores del turno de noche representan el 25 por ciento de los empleados que menor producción tienen debido a la falta de luz natural. En el año 2004, de la distribución del número de unidades producidas mensualmente por trabajador, ha resultado que la cantidad máxima obtenida por los trabajadores del turno de noche ha sido 50 mil unidades.

a) Analícese el efecto que produciría una transformación lineal sobre los cuantiles de la distribución $(L_{i-1} - L_i\,; f_i)$.

b) Calcúlese el salario máximo que perciben los trabajadores del turno de noche.

SOLUCIÓN

a) Si $L_{i-1} - L_i$ es el intervalo cuantílico de orden q de la distribución $(L_{i-1} - L_i\,; f_i)$, el cuantil de orden q de esta distribución es

$$x_q = L_{i-1} + \frac{q \cdot N - N_{i-1}}{n_i} \cdot c_i.$$

El intervalo cuantílico de la distribución transformada es $(a \cdot L_{i-1} + b) - (a \cdot L_i + b)$, donde a y b son constantes cualesquiera, ya que una transformación lineal no produce ningún cambio sobre las frecuencias de los intervalos.

Por tanto, el cuantil de orden q de la distribución transformada es

$$y_q = (a \cdot L_{i-1} + b) + \frac{q \cdot N - N_{i-1}}{n_i}\,[(a \cdot L_i + b) - (a \cdot L_{i-1} + b)],$$

donde $(a \cdot L_i + b) - (a \cdot L_{i-1} + b) = a\,(L_i - L_{i-1}) = a \cdot c_i$ es la longitud de intervalo cuantílico de la distribución transformada.

Operando en esta expresión, se tiene la siguiente relación entre los cuantiles de orden q de ambas distribuciones:

$$y_q = (a \cdot L_{i-1} + b) + \frac{q \cdot N - N_{i-1}}{n_i} \cdot a \cdot c_i = a\left(L_{i-1} + \frac{q \cdot N - N_{i-1}}{n_i} \cdot c_i\right) + b = a \cdot x_q + b.$$

Este resultado puede aplicarse a la mediana de la distribución, pues, como es conocido, esta medida de posición es un cuantil; en consecuencia, si Me es la mediana de la distribución $(L_{i-1} - L_i\,; f_i)$, entonces, $a \cdot Me + b$ es la mediana de la distribución transformada linealmente, siendo a y b constantes cualesquiera.

b) Según se lee en el enunciado, los trabajadores del turno de noche representan el 25 por ciento de los empleados que menor producción tienen, con una cantidad máxima obtenida por ellos igual a 50 mil unidades; esto significa que el primer cuartil de la distribución del número de unidades producidas mensualmente es igual a 50 mil.

Puesto que, además, entre el número de unidades producidas mensualmente X, en miles, y el salario mensual, Y, en euros, de los trabajadores existe la relación lineal:

$$Y = 800 + 1{,}5 \cdot X,$$

aplicando el resultado demostrado en el apartado anterior al primer cuartil, se tiene que

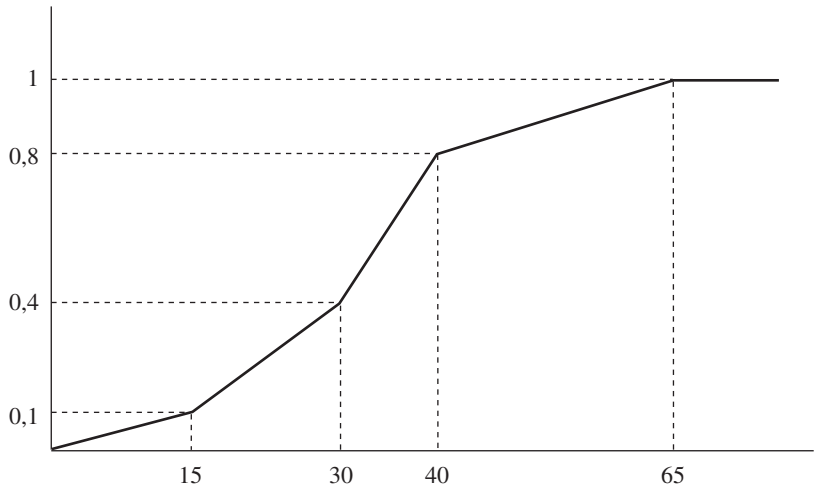
$$C_1(y) = 800 + 1{,}5 \cdot C_1(x) = 800 + 1{,}5 \cdot 50 = 875 \text{ euros}$$

es el primer cuartil de la distribución del salario que perciben los trabajadores, esto es, el salario máximo que perciben los trabajadores del turno de noche.

1.46 Para una cantidad fija asegurada, el precio de un cierto seguro de vida, P, en euros, depende de la edad del individuo, X, en años:

$$P = 3 \cdot X + 7.$$

Analizada una población de 200 individuos se obtiene el siguiente polígono de frecuencias acumuladas para la distribución de la variable edad:



a) Hállese el precio medio de los seguros de vida.

b) ¿Cuál el es precio más frecuente?

c) ¿Cuál es el precio máximo del 50 por ciento de las pólizas más baratas?

d) ¿Cuántos individuos tienen suscritas pólizas cuyos precios están comprendidos entre 127 y 202 euros?

SOLUCIÓN

A partir del polígono de frecuencias acumuladas resulta la siguiente distribución de frecuencias de la variable X:

Edad	x_i	F_i	f_i
0-15	7,5	0,1	0,1
15-30	22,5	0,4	0,3
30-40	35,0	0,8	0,4
40-65	52,5	1,0	0,2

La última columna de esta tabla, frecuencias relativas de la distribución de edades, se obtiene con los datos de la columna anterior, es decir, a partir de las frecuencias relativas acumuladas, según las relaciones:

$$f_1 = F_1,$$

para el primer intervalo, y

$$f_i = F_i - F_{i-1},$$

para el resto de los intervalos.

Para hallar los tres promedios media, moda y mediana de la variable P, de la cual se desconoce su distribución de frecuencias, hay que tener en cuenta la relación lineal existente entre P y X,

$$P = 3 \cdot X + 7,$$

así como las propiedades de los tres promedios estudiadas en problemas anteriores.

a) Hallemos, en primer lugar, la edad media de los individuos, media de X, variable cuya distribución es conocida:

$$\overline{x} = \sum_{i=1}^{h} x_i \cdot f_i = 7{,}5 \cdot 0{,}1 + 22{,}5 \cdot 0{,}3 + 35 \cdot 0{,}4 + 52{,}5 \cdot 0{,}2 = 32 \text{ años.}$$

Entonces, por las propiedades de la media aritmética, se tiene que la media de la distribución de los precios es

$$\overline{p} = 3 \cdot \overline{x} + 7,$$

con lo cual,

$$\overline{p} = 3 \cdot 32 + 7 = 103 \text{ euros.}$$

b) El cálculo del precio más frecuente, esto es, de la moda de la distribución de la variable P, requiere la obtención previa de la moda de la variable X. Para ello, hallamos el intervalo modal y aplicamos la expresión:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i.$$

Puesto que, en esta ocasión, disponemos de frecuencias relativas, calcularemos las densidades de frecuencia de los intervalos como

$$d_i = \frac{f_i}{c_i},$$

según aparece en la tabla siguiente, de la que se deduce que 30-40 es el intervalo modal:

d_i	0,006	0,02	0,04	0,008

Sustituyendo por los datos del problema, se tiene que

$$Mo_X = 30 + \frac{0{,}008}{0{,}02 + 0{,}008} \cdot 10 = 32{,}857 \text{ años.}$$

Por aplicación de las propiedades de esta medida de posición, resulta el valor de la moda de la distribución transformada, es decir, el precio más frecuente:

$$Mo_P = 3 \cdot Mo_X + 7,$$

es decir,

$$Mo_P = 3 \cdot 32{,}857 + 7 = 105{,}571 \text{ euros.}$$

c) De modo semejante a lo realizado en el apartado anterior, calculamos la mediana de la distribución de la variable edad cuyo intervalo mediano es, también, 30-40, primer intervalo cuya frecuencia relativa acumulada, 0,8, es estrictamente mayor que 0,5. La expresión de la mediana a partir de frecuencias relativas es, basándonos en **1.42**,

$$Me = L_{i-1} + \frac{0{,}5 - F_{i-1}}{f_i} \cdot c_i.$$

Con los datos del problema resulta que el valor mediano de la variable edad, X, es

$$Me_X = 30 + \frac{0{,}5 - 0{,}4}{0{,}4} \cdot 10 = 32{,}5 \text{ años,}$$

por lo que, aplicando las propiedades de la mediana,

$$Me_P = 3 \cdot Me_X + 7,$$

esto es,

$$Me_P = 3 \cdot 32{,}5 + 7 = 104{,}5 \text{ euros.}$$

b) De la relación entre el precio del seguro, P, y la edad del individuo, X, resulta, despejando, que

$$X = \frac{P - 7}{3}.$$

En consecuencia, una póliza con un precio de 202 euros corresponde, sólo con sustituir y operar después en la expresión anterior, a individuos con 65 años; y una póliza de 127 euros corresponde, de igual manera, a individuos con 40 años.

Por tanto, responder a la pregunta sobre cuántos individuos tienen suscritas pólizas cuyas primas están entre 127 y 202 euros es equivalente a responder sobre cuántos tienen edades comprendidas entre 40 y 65 años, que, dada la distribución de frecuencias de la variable X, supone el 20 por ciento de los 200 individuos, es decir, 40 individuos.

1.47 Se han calculado los percentiles, en miles de euros, de la distribución de ingresos recaudados en concepto de impuesto sobre bienes inmuebles en el Ayuntamiento de Santiuste de Camarreal en 2003, arrojando los siguientes resultados: $P_{20} = 10$, $P_{40} = 40$ y $P_{70} = 60$. Se sabe, también, que la recaudación máxima se obtuvo el último día del plazo establecido para el pago y fue de 100 mil euros.

a) ¿Qué cantidad máxima se recaudó el 40 por ciento de los días en que hubo menor recaudación?

b) Calcúlese la recaudación media diaria.

c) ¿Cuál ha sido la cantidad recaudada un mayor número de días?

SOLUCIÓN

a) Este apartado se responde con el percentil 40 que es un dato del enunciado: 40 mil euros.

b) La información en forma de percentiles permite considerar una posible agrupación en intervalos de la distribución de los ingresos recaudados.

Esta agrupación resulta de suponer que la frecuencia relativa acumulada de cada clase coincide con la proporción de observaciones de la distribución que son menores o iguales que el or-

den del correspondiente percentil. Como consecuencia de esta hipótesis, el percentil es el extremo superior del intervalo, con lo cual, conocido el percentil, el extremo queda determinado.

El resultado de aplicar esta suposición a los datos del enunciado permite considerar como posible distribución de los ingresos la que se recoge en la siguiente tabla:

$L_{i-1} - L_i$	F_i
0-10	0,20
10-40	0,40
40-60	0,70
60-100	1

Así, por ejemplo, puesto que $P_{70} = 60$, 60 es el extremo superior del intervalo cuya frecuencia relativa acumulada es 70/100 = 0,7; de la misma forma, se obtendría el resto de las casillas de la tabla anterior.

Completamos la tabla con dos columnas más, consecuencia inmediata de las anteriores:

$L_{i-1} - L_i$	x_i	F_i	f_i
0-10	5	0,20	0,20
10-40	25	0,40	0,20
40-60	50	0,70	0,30
60-100	80	1	0,30

La recaudación media se halla, como es habitual, utilizando las marcas de clase:

$$\bar{x} = \sum_{i=1}^{h} x_i \cdot f_i = 5 \cdot 0{,}20 + 25 \cdot 0{,}20 + 50 \cdot 0{,}30 + 80 \cdot 0{,}30 = 45 \text{ mil euros.}$$

c) La cantidad recaudada un mayor número de días, es decir, la moda de la distribución, se obtiene aplicando la expresión siguiente, en la que se utilizan las densidades de frecuencia puesto que los intervalos son de distinta amplitud:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i+1} + d_{i-1}} \cdot c_i,$$

siendo c_i la amplitud del intervalo modal y d_{i-1} y d_{i+1} las densidades de frecuencia de sus intervalos contiguos.

Las amplitudes, c_i, y las densidades de frecuencia de los intervalos, $d_i = f_i/c_i$, aparecen en la tabla siguiente:

$L_{i-1} - L_i$	0-10	10-40	40-60	60-100
c_i	10	30	20	40
d_i	0,02	0,006	0,015	0,0075

En definitiva, y puesto que el intervalo modal, o intervalo con mayor densidad de frecuencia, es el intervalo 0-10, la moda es, sin más que sustituir en la expresión genérica,

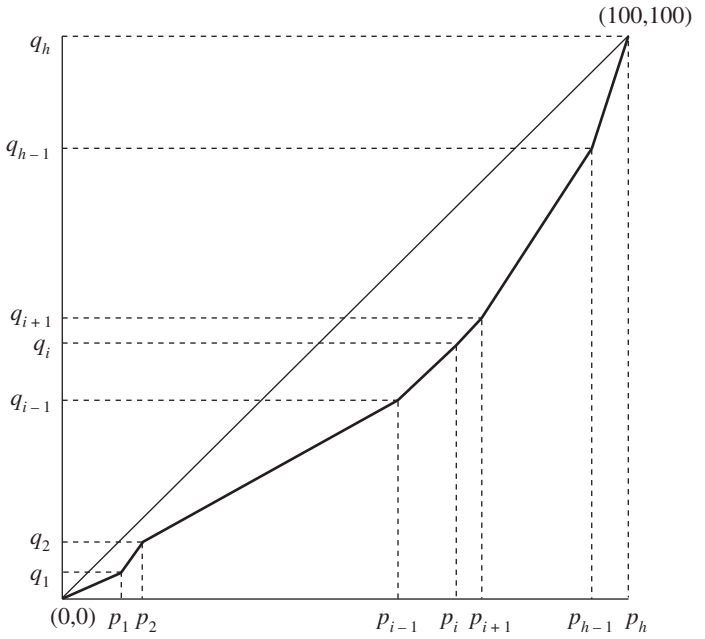
$$Mo = 0 + \frac{0{,}006}{0 + 0{,}006} \cdot 10 = 10 \text{ mil euros.}$$

Obsérvese que, al ser el intervalo modal el primer intervalo de la distribución, la moda es el extremo superior del intervalo.

1.48 Dados los pares de puntos (p_i, q_i), $i = 1, \ldots, h$, que conforman la curva de Lorenz de una distribución de frecuencias, obténgase la expresión del índice de Gini.

SOLUCIÓN

A partir de la representación de la curva de Lorenz en el cuadrado de lado 100, se consideran tres áreas: el área del triángulo, A_t, el área de concentración, A_c, y el área por debajo de ella, A_d.



Así, teniendo en cuenta que entre estas áreas se verifica la relación:

$$A_t = A_c + A_d,$$

el índice de Gini que, por definición, es igual al cociente entre el área de concentración y el área del triángulo, puede escribirse como

$$I_G = \frac{A_c}{A_t} = \frac{A_t - A_d}{A_t} = 1 - \frac{A_d}{A_t}.$$

Para calcular el área A_d basta considerar que ésta puede descomponerse, a su vez, en áreas de trapecios, siendo el área del trapecio genérico, que aparece en la figura, igual a

$$\frac{(q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{2},$$

donde q_i es su base menor, q_{i+1} es su base mayor y $(p_{i+1} - p_i)$ es la longitud de su altura.

La suma de las áreas de los trapecios en que puede dividirse la figura bajo la curva de concentración es

$$\frac{(q_2 + q_1) \cdot (p_2 - p_1)}{2} + \ldots + \frac{(q_h + q_{h-1}) \cdot (p_h - p_{h-1})}{2}.$$

Para completar el área A_d hay que añadir el área del triángulo rectángulo situado a la izquierda del primer trapecio:

$$\frac{q_1 \cdot p_1}{2},$$

que puede escribirse como

$$\frac{(q_1 + q_0) \cdot (p_1 - p_0)}{2},$$

ya que, tanto p_0 como q_0 son iguales a cero.

En definitiva, el área A_d resulta ser

$$A_d = \sum_{i=0}^{h-1} \frac{(q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{2}.$$

En consecuencia, y puesto que el área A_t, área del triángulo, es, evidentemente, igual a 10 000/2, el índice de Gini es

$$I_G = 1 - \frac{\sum_{i=0}^{h-1} \frac{(q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{2}}{\frac{10\,000}{2}} = 1 - \frac{\sum_{i=0}^{h-1} (q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{10\,000}.$$

1.49 Los ingresos anuales, en miles de euros, de 10 empleados de una empresa son los siguientes:

Ingresos	N.º empleados
155	4
15	1
60	2
25	1
40	2

Calcúlese el índice de Gini de esta distribución.

En la expresión del índice de Gini,

$$I_G = 1 - \frac{\sum_{i=0}^{h-1} (q_{i+1} + q_i) \cdot (p_{i+1} - p_i)}{10\,000},$$

p_i es el porcentaje de empleados con ingresos anuales menores o iguales que x_i, y q_i el porcentaje de ingresos anuales percibidos por los individuos con renta menor o igual que x_i; el cálculo de estos porcentajes, p_i y q_i, requiere, por tanto, la ordenación previa de los valores de la variable ingreso anual.

Los valores

$$p_i = \frac{N_i}{N} \cdot 100$$

y

$$q_i = \frac{u_i}{u_h} \cdot 100,$$

con $u_i = \sum_{j=1}^{i} x_j \cdot n_j$, ingresos anuales percibidos por los individuos con renta menor o igual que x_i, y

$u_h = \sum_{j=1}^{h} x_j \cdot n_j$, ingresos anuales totales, aparecen en la siguiente tabla:

x_i	n_i	$x_i \cdot n_i$	N_i	u_i	p_i	q_i
15	1	15	1	15	10	1,74
25	1	25	2	40	20	4,65
40	2	80	4	120	40	13,95
60	2	120	6	240	60	27,91
155	4	620	10	860	100	100

Por ejemplo, a un ingreso de 25 mil euros le corresponde un valor

$$p_2 = \frac{N_2}{N} \cdot 100 = \frac{2}{10} \cdot 100 = 20$$

y un valor

$$q_2 = \frac{u_2}{u_h} \cdot 100 = \frac{40}{860} \cdot 100 = 4{,}65;$$

por tanto, el 20 por ciento de los empleados perciben unos ingresos anuales menores o iguales a 25 mil euros, que suponen el 4,65 por ciento del total de ingresos anuales que reciben los empleados de esta empresa. De igual modo se halla el resto de las cantidades p_i y q_i.

En definitiva, el índice de Gini es

$$I_G = 1 - \frac{(1{,}74 + 0) \cdot (10 - 0) + (4{,}65 + 1{,}8) \cdot (20 - 10) + (13{,}95 + 4{,}65) \cdot (40 - 20)}{10\,000} -$$

$$- \frac{(27{,}91 + 13{,}95) \cdot (60 - 40) + (100 + 27{,}91) \cdot (100 - 60)}{10\,000},$$

esto es, $I_G = 0{,}36$.

1.50 Demuéstrese que el índice de Gini de la distribución $(x_i; f_i)$ coincide con el de la distribución $(a \cdot x_i; f_i)$, donde a es una constante cualquiera.

SOLUCIÓN

Teniendo en cuenta que

$$\frac{p_{i+1} - p_i}{100} = \frac{1}{100}\left(\frac{N_{i+1}}{N} \cdot 100 - \frac{N_i}{N} \cdot 100\right) = \frac{N_{i+1}}{N} - \frac{N_i}{N} = F_{i+1} - F_i = f_{i+1},$$

el índice de Gini de la distribución $(a \cdot x_i; f_i)$ puede expresarse como

$$I'_G = 1 - \frac{\sum_{i=0}^{h-1} (q'_{i+1} + q'_i)}{100} \cdot f_{i+1},$$

donde

$$q'_i = \frac{u'_i}{u'_h} \cdot 100 = \frac{\sum_{j=1}^{i} a \cdot x_j \cdot n_j}{\sum_{j=1}^{h} a \cdot x_j \cdot n_j} \cdot 100$$

y

$$q'_{i+1} = \frac{u'_{i+1}}{u'_h} \cdot 100 = \frac{\sum_{j=1}^{i+1} a \cdot x_j \cdot n_j}{\sum_{j=1}^{h} a \cdot x_j \cdot n_j} \cdot 100,$$

sin más que sustituir los valores de la variable de la distribución transformada.

Sacando factor común a la constante a en los numeradores y denominadores de las expresiones anteriores, se tiene que

$$q_i' = \frac{a\sum_{j=1}^{i} x_j \cdot n_j}{a\sum_{j=1}^{h} x_j \cdot n_j} \cdot 100 = \frac{\sum_{j=1}^{i} x_j \cdot n_j}{\sum_{j=1}^{h} x_j \cdot n_j} \cdot 100 = q_i,$$

y, de modo análogo, se comprueba que $q_{i+1}' = q_{i+1}$.

En definitiva,

$$I_G' = 1 - \frac{\sum_{i=0}^{h-1} (q_{i+1}' + q_i')}{100} \cdot f_{i+1} = 1 - \frac{\sum_{i=0}^{h-1} (q_{i+1} + q_i)}{100} \cdot f_{i+1} = I_G,$$

según queríamos demostrar.

1.51 Las distribuciones de las acciones de dos sociedades A y B, agrupadas en intervalos, se representan en la siguiente tabla:

N.º acciones	N.º accionistas (A)	N.º accionistas (B)
0-20	10	60
20-30	30	12
30-50	40	7
50-150	20	1

a) Calcúlese el promedio de acciones por accionista para cada una de las sociedades. ¿Cuál de los dos promedios es más representativo?

b) ¿En qué sociedad está más concentrado el reparto de acciones?

SOLUCIÓN

a) A partir del enunciado se tiene, para la sociedad A, la siguiente distribución de frecuencias:

N.º acciones (A)	x_i	n_i
0-20	10	10
20-30	25	30
30-50	40	40
50-150	100	20

Utilizando las correspondientes marcas de clase, se halla la media aritmética de esta distribución,

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_i = \frac{10 \cdot 10 + 25 \cdot 30 + 40 \cdot 40 + 100 \cdot 20}{10 + 30 + 40 + 20} = 44{,}5 \text{ acciones},$$

así como la varianza,

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_i - \bar{x}^2 = \frac{10^2 \cdot 10 + 25^2 \cdot 30 + 40^2 \cdot 40 + 100^2 \cdot 20}{100} - 44{,}5^2 = 857{,}25,$$

y la desviación típica, raíz cuadrada positiva de la varianza,

$$S_X = 29{,}28.$$

Con estos resultados se calcula el coeficiente de variación,

$$V_X = \frac{S_X}{\bar{x}} = \frac{29{,}28}{44{,}5} = 0{,}66,$$

medida de dispersión relativa que emplearemos para comparar la representatividad del promedio de esta distribución con la del promedio de la segunda distribución, que se analiza a continuación.

Así, por lo que se refiere a la distribución de las acciones en la sociedad B, se tiene la siguiente tabla:

N.º acciones (B)	y_j	n_j
0-20	10	60
20-30	25	12
30-50	40	7
50-150	100	1

El valor medio de esta distribución es

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{k} y_j \cdot n_j = \frac{10 \cdot 60 + 25 \cdot 12 + 40 \cdot 7 + 100 \cdot 1}{60 + 12 + 7 + 1} = 16 \text{ acciones},$$

siendo la varianza

$$S_Y^2 = \frac{1}{N}\sum_{j=1}^{k} y_j^2 \cdot n_j - \bar{y}^2 = \frac{10^2 \cdot 60 + 25^2 \cdot 12 + 40^2 \cdot 7 + 100^2 \cdot 1}{80} - 16^2 = 177{,}75$$

y su raíz cuadrada, esto es, la desviación típica,

$$S_Y = 13{,}33.$$

Por consiguiente, el coeficiente de variación de la segunda distribución es

$$V_Y = \frac{S_Y}{\bar{y}} = \frac{13{,}33}{16} = 0{,}83.$$

A la vista de los coeficientes de variación obtenidos se concluye que la media de acciones por accionistas es más representativa en la sociedad A, puesto que es menor el coeficiente de variación de su distribución de acciones.

b) Para calcular el grado de concentración de las acciones en cada una de las sociedades, se construyen las siguientes tablas que servirán de apoyo en la obtención de los respectivos índices de Gini.

Así, por lo que respecta a la sociedad A se tiene:

x_i	n_i	$x_i \cdot n_i$	N_i	u_i	q_i	p_i	f_i
10	10	100	10	100	2,25	10	0,10
25	30	750	40	850	19,10	40	0,30
40	40	1 600	80	2 450	55,06	80	0,40
100	20	2 000	100	4 450	100	100	0,20

Y, en cuanto a la sociedad B,

y_j	n_j	$y_j \cdot n_j$	N_j	u_j	q_j	p_j	f_j
10	60	600	60	600	46,88	75	0,75
25	12	300	72	900	70,31	90	0,15
40	7	280	79	1 180	92,19	99	0,09
100	1	100	80	1 280	100	100	0,01

Nótese que la columna de las frecuencias absolutas acumuladas de la distribución de acciones de la sociedad A coincide con la columna de los porcentajes p_i, ya que, para dicha distribución, N es igual a 100.

Sustituyendo en la expresión del índice de Gini,

$$I_G = 1 - \frac{\sum_{i=0}^{h-1} (q_{i+1} + q_i)}{100} \cdot f_{i+1},$$

los datos de ambas distribuciones que aparecen en las tablas anteriores, se obtienen, respectivamente, el índice de concentración de la sociedad A,

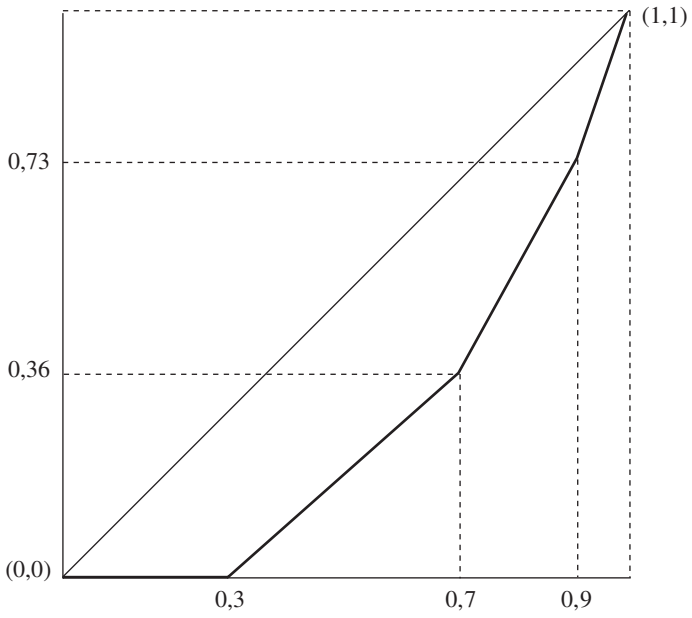
$$I_{G_A} = 1 - \frac{1}{100}[(2{,}25 + 0)0{,}10 + (19{,}10 + 2{,}25)0{,}30 + (55{,}06 + 19{,}10)0{,}40 + \\ + (100 + 55{,}06)0{,}20] = 0{,}327,$$

y el índice de concentración de la sociedad B,

$$I_{G_B} = 1 - \frac{1}{100}[(46{,}88 + 0)0{,}75 + (70{,}31 + 46{,}88)0{,}15 + (92{,}19 + 70{,}31)0{,}09 + \\ + (100 + 92{,}19)0{,}01] = 0{,}307.$$

Puesto que el índice de Gini de la distribución de acciones es ligeramente mayor en la sociedad A, en ella existe una menor igualdad en el reparto de las acciones, esto es, existe una concentración mayor en esa sociedad, aunque, ciertamente, la diferencia es escasa.

1.52 Se ha estudiado el número de hijos (0, 1, 2 ó 3) de una población de 100 familias, y se ha obtenido la siguiente curva de concentración:



a) Hállese una medida del grado de concentración de esta distribución.

b) Obténgase la media aritmética, la mediana y la desviación típica de la distribución correspondiente a la variable número de hijos por familia.

SOLUCIÓN

a) Una primera observación que sugiere la lectura del enunciado es que, si bien el análisis de la concentración se realiza habitualmente con variables económicas tales como la renta, es posible llevarlo a cabo en cualquier otra variable, como es el caso que nos ocupa.

Por lo demás, a partir de la curva de Lorenz se obtienen los pares de valores que figuran en la siguiente tabla:

q_i	0	0,36	0,73	1
p_i	0,3	0,70	0,90	1

Los datos, en tanto por uno, permiten utilizar la siguiente expresión del índice de Gini:

$$I_G = 1 - \sum_{i=0}^{h-1}(q_{i+1} + q_i) \cdot (p_{i+1} - p_i),$$

que proponemos al lector que compruebe, utilizando la demostración realizada en **1.48** y teniendo en cuenta que, en este caso, el área del triángulo bajo la curva de concentración es 1/2.

En definitiva, el grado de concentración de esta distribución es

$$I_G = 1 - [0 \cdot 0{,}30 + (0{,}36 + 0)\ 0{,}40 + (0{,}73 + 0{,}36)\ 0{,}2 + (1 + 0{,}73)\ 0{,}1] = 0{,}465.$$

b) Teniendo en cuenta que, en esta situación,

$$p_{i+1} - p_i = f_{i+1},$$

ya que, como sabemos, p_i y p_{i+1} son proporciones, se obtiene la siguiente distribución de frecuencias de la variable número de hijos:

x_i	0	1	2	3
f_i	0,3	0,4	0,2	0,1

La media aritmética de esta distribución de frecuencias es, por tanto,

$$\bar{x} = \sum_{i=1}^{h} x_i \cdot f_i = 0 \cdot 0{,}3 + 1 \cdot 0{,}4 + 2 \cdot 0{,}2 + 3 \cdot 0{,}1 = 1{,}1 \text{ hijos.}$$

A partir de cada frecuencia relativa, f_i, se obtiene la frecuencia absoluta, n_i, y la frecuencia absoluta acumulada, N_i:

n_i	30	40	20	10
N_i	30	70	90	100

Obsérvese, igualmente, que podríamos haber calculado las frecuencias absolutas acumuladas utilizando los valores p_i, puesto que, en este problema, en el que dichos valores son proporciones, se cumple que $N_i = N \cdot p_i$.

Como $N/2$ es igual a 50, para calcular la mediana se toma el menor valor de la variable, x_i, tal que la frecuencia absoluta acumulada, N_i, sea estrictamente mayor que 50, resultando que la mediana es 1 hijo.

En cuanto a la varianza de la variable, se halla aplicando la expresión:

$$S^2 = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_i - \bar{x}^2 = \frac{0^2 \cdot 30 + 1^2 \cdot 40 + 2^2 \cdot 20 + 3^2 \cdot 10}{100} - 1{,}1^2 = 0{,}89,$$

por lo que la desviación típica, S, raíz cuadrada de la varianza, es igual a 0,94.

1.53 En dos pueblos limítrofes de la Comunidad de Castilla y León, los terrenos dedicados a la agricultura son propiedad de los vecinos. El porcentaje de familias propietarias de, a lo sumo, x_i hectáreas, p_i, y el porcentaje de terrenos que poseen dichas familias, q_i, se recogen en las siguientes tablas, correspondientes a cada una de las localidades:

Localidad A	
p_i	q_i
10	10
20	20
80	75
90	85
100	100

Localidad B	
p_i	q_i
10	5
20	15
80	80
90	90
100	100

a) Calcúlese el índice de Gini para cada una de las distribuciones de terreno.

b) A la vista de los resultados obtenidos en el apartado anterior, ¿podría afirmarse que el total de terrenos «se reparte» entre el total de familias de igual modo en ambas distribuciones?

SOLUCIÓN

a) Con los datos que aparecen en la tabla siguiente referidos a la localidad A, se halla la columna de las frecuencias relativas, según la relación $(p_{i+1} - p_i)/100 = f_{i+1}$.

p_i	q_i	f_i
10	10	0,10
20	20	0,10
80	75	0,60
90	85	0,10
100	100	0,10

Por tanto, sustituyendo en la expresión de índice de Gini,

$$I_G = 1 - \frac{\sum_{i=0}^{h-1}(q_{i+1} + q_i)}{100} \cdot f_{i+1},$$

los datos de la distribución de hectáreas en la primera localidad, se tiene que

$$I_{G_A} = 1 - \frac{1}{100}(10 \cdot 0{,}10 + 30 \cdot 0{,}10 + 95 \cdot 0{,}60 + 160 \cdot 0{,}10 + 185 \cdot 0{,}10) = 0{,}045.$$

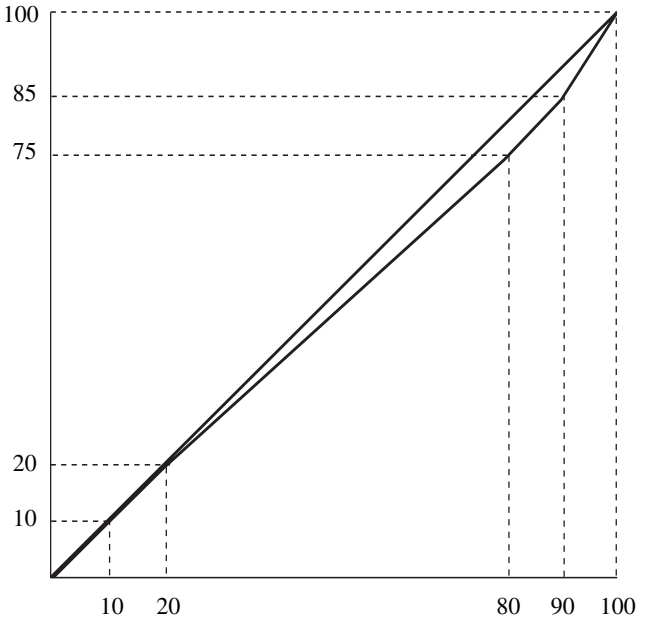
De igual forma, los datos de la localidad B permiten elaborar la siguiente tabla:

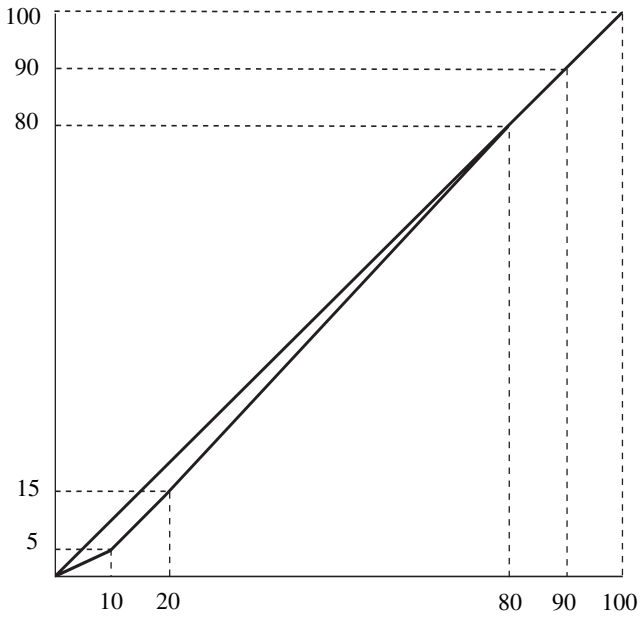
p_i	q_i	f_i
10	5	0,10
20	15	0,10
80	80	0,60
90	90	0,10
100	100	0,10

A partir de ella se calcula el índice de Gini de la distribución de hectáreas de la localidad B:

$$I_{G_B} = 1 - \frac{1}{100}(5 \cdot 0{,}10 + 20 \cdot 0{,}10 + 95 \cdot 0{,}60 + 170 \cdot 0{,}10 + 190 \cdot 0{,}10) = 0{,}045.$$

b) Aunque el valor del índice de Gini es el mismo en las dos distribuciones, no puede afirmarse que el reparto sea igual en ambas, como reflejan las columnas de los valores p_i y q_i. Esta idea se confirma dibujando las respectivas curvas de Lorenz.





A la vista de las dos curvas, comprobamos que, efectivamente, el índice de Gini tiene idéntico valor en las dos distribuciones, puesto que las áreas de concentración son iguales, aunque las curvas de concentración sean diferentes.

Este ejemplo pone de manifiesto que dos distribuciones con igual índice de Gini pueden tener distinta concentración y que, por tanto, para comparar concentraciones es necesario completar la información que los índices proporcionan con las respectivas curvas de Lorenz, ya que éstas ponen de manifiesto las posibles diferencias que puedan existir entre los pares de puntos (p_i, q_i) de cada distribución.

1.54 Demuéstrese que, si el salario de cada uno de los trabajadores de una empresa se duplica, la concentración de salarios sigue siendo la misma.

SOLUCIÓN

Si $(x_i\,;f_i)$ es la distribución inicial de los salarios de los trabajadores de la empresa y su grado de concentración viene dado por el índice de Gini, I_G, el índice de Gini de la distribución de los salarios una vez duplicado estos, $(2 \cdot x_i\,;f_i)$, sigue siendo el índice de Gini el de la distribución inicial, I_G, utilizando el resultado de **1.50** para a igual a 2.

El lector aventajado podría encontrar, a primera vista, contradicción entre este razonamiento, con el cual comparamos el grado de concentración de dos distribuciones utilizando únicamente los respectivos índices de Gini, y los comentarios realizados en el apartado ***b)*** del problema anterior, sobre la necesidad de acompañar el valor del índice con la representación de la curva de Lorenz. Sin embargo, a estas alturas ya habrá caído en la cuenta de que en esta situación no es necesario completar la información, porque, aunque nuestras deducciones han sido sobre los valores de los índices de Gini de las dos distribuciones, previamente hemos comprobado —en **1.50**— que, tanto los porcentajes p_i como los porcentajes q_i, son *idénticos* en ambas, siendo en definitiva, idénticos, también, los grados de concentración.

Capítulo 2

Distribuciones de frecuencias bidimensionales

Principales conceptos y resultados

La observación conjunta de dos variables X e Y en las N unidades de una población conduce a la obtención de pares de datos. Si $x_1, ..., x_h$ son los valores de X e $y_1, ..., y_k$ son los valores de Y, los pares de valores (x_i, y_j), $(i = 1, ..., h, j = 1, ..., k)$, son los valores de la **variable bidimensional** (X, Y).

La frecuencia absoluta de un valor (x_i, y_j), o **frecuencia absoluta conjunta** es el número de veces que aparecen simultáneamente los valores x_i e y_j en las unidades de la población y se denota por n_{ij}. Se cumple que

$$\sum_{i=1}^{h} \sum_{j=1}^{k} n_{ij} = N.$$

La frecuencia relativa de un valor (x_i, y_j) o **frecuencia relativa conjunta**, f_{ij}, es la proporción de observaciones iguales a dicho valor. Por definición,

$$f_{ij} = \frac{n_{ij}}{N},$$

con lo cual,

$$\sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = 1.$$

Una **distribución de frecuencias bidimensional** es el conjunto de valores de la variable (X, Y), junto con sus correspondientes frecuencias. Se denota por $(x_i, y_j; n_{ij})$ o bien $(x_i, y_j; f_{ij})$, según se utilicen las frecuencias conjuntas absolutas o relativas.

La forma más cómoda y sencilla de disponer la información proporcionada por una distribución de frecuencias bidimensional es una tabla de doble entrada denominada **tabla de correlación**. Así, si suponemos que, tanto los valores de la variable X, como de la variable Y, están ordenados de menor a mayor, tendremos[1]:

X \ Y	y_1	…	y_j	…	y_k
x_1	n_{11}	…	n_{1j}	…	n_{1k}
⋮	⋮		⋮		⋮
x_i	n_{i1}	…	n_{ij}	…	n_{ik}
⋮	⋮		⋮		⋮
x_h	n_{h1}	…	n_{hj}	…	n_{hk}

A partir de la distribución de frecuencias bidimensional $(x_i, y_j; n_{ij})$, pueden obtenerse las distribuciones de frecuencias correspondientes a las variables X e Y, **distribuciones de frecuencias marginales**, $(x_i; n_{i.})$ e $(y_j; n_{.j})$, respectivamente, donde, para cada i,

$$n_{i.} = \sum_{j=1}^{k} n_{ij}$$

y, para cada j,

$$n_{.j} = \sum_{i=1}^{h} n_{ij}.$$

La **distribución de X condicionada por el valor y_j de la variable Y**, se expresa como $(x_i/Y = y_j; n_{i/j})$ y sus valores y frecuencias aparecen en la tabla siguiente:

$x_i/Y = y_j$	$n_{i/j}$
x_1	n_{1j}
⋮	⋮
x_i	n_{ij}
⋮	⋮
x_h	n_{hj}

La **frecuencia absoluta genérica**, $n_{i/j}$, es el número de unidades de la población que tienen el valor x_i de la variable X *dentro* de las que tienen el valor y_j de la variable Y.

Del mismo modo se define la **distribución de Y condicionada por el valor x_i de X**, $(y_j/X = x_i; n_{j/i})$, donde:

[1] En el interior de la tabla pueden disponerse frecuencias absolutas o relativas.

$y_j/X = x_i$	$n_{j/i}$
y_1	n_{i1}
⋮	⋮
y_j	n_{ij}
⋮	⋮
y_k	n_{ik}

La **frecuencia absoluta genérica**, $n_{j/i}$, es el número de unidades de la población que tienen el valor y_j de la variable *Y dentro* de las que tienen el valor x_i de la variable *X*.

Cuando las distribuciones de frecuencias están agrupadas en clases, las marcas de clase desempeñan el papel de representantes del intervalo.

Las **frecuencias relativas condicionadas** genéricas correspondientes a las distribuciones anteriores se obtienen a partir de las frecuencias absolutas condicionadas:

$$f_{i/j} = \frac{n_{i/j}}{n_{\cdot j}}$$

y

$$f_{j/i} = \frac{n_{j/i}}{n_{i\cdot}}.$$

Dada una distribución de frecuencias $(x_i, y_j; f_{ij})$, las variables *X* e *Y* son **estadísticamente independientes** o simplemente **independientes**, si, para cualesquiera *i* y *j*, se cumple:

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

esto es, cada frecuencia relativa conjunta es igual al producto de las correspondientes frecuencias relativas marginales.

Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, las variables *X* e *Y* son independientes, si y solamente si,

$$f_{i/j} = f_{i\cdot}$$

y

$$f_{j/i} = f_{\cdot j},$$

para cualesquiera *i* y *j*, es decir, cuando las frecuencias relativas condicionadas sean idénticas a sus respectivas frecuencias relativas marginales.

Al igual que en el caso de una distribución de frecuencias unidimensional, los **momentos bidimensionales** son medidas de resumen de la información proporcionada por los datos.

El **momento respecto al origen** o **momento no central de orden** (r, s) de la distribución bidimensional $(x_i, y_j; n_{ij})$ se define como

$$a_{r,s} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^s \cdot n_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^s \cdot f_{ij}.$$

El **momento respecto a las medias** o **momento central de orden** (r, s) de la distribución bidimensional $(x_i, y_j; n_{ij})$ es

$$m_{r,s} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^s \, n_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^s \, f_{ij}.$$

El momento $m_{1,1}$ se llama **covarianza** entre las variables X e Y y se denota también por $S_{X,Y}$, o simplemente, por S:

$$S = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_j - \overline{y}) \, f_{ij}.$$

La **media de la distribución condicionada** $(x_i/Y = y_j; f_{i/j})$ se define como

$$\overline{x}/(Y = y_j) = \sum_{i=1}^{h} x_i \cdot f_{i/j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{h} x_i \cdot n_{ij},$$

siendo la **varianza** de esta distribución:

$$S^2_{X/Y = y_j} = \sum_{i=1}^{h} (x_i - \overline{x}/(Y = y_j))^2 \, f_{i/j} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{h} (x_i - \overline{x}/(Y = y_j))^2 n_{ij}.$$

De igual forma, se definen la media y la varianza de la distribución condicionada $(y_j/X = x_i; f_{j/i})$.

Obsérvese que, puesto que las distribuciones condicionadas son distribuciones de frecuencias unidimensionales, es posible calcular todas las características que corresponden a este tipo de distribuciones y que ya se vieron en el capítulo anterior.

Uno de los aspectos fundamentales en el estudio conjunto de dos variables es el análisis de la posible relación existente entre ellas. La estadística permite, mediante procedimientos matemáticos, determinar si las variables tienen o no relación, así como medir el grado de la misma.

La relación entre variables contempla dos vertientes: su *forma* y su *grado*. La forma de la relación tiene que ver con el *aspecto* de la representación gráfica de los valores de la variable (X, Y), denominada **nube de puntos** o **diagrama de dispersión**. Esta forma se concreta en la **ecuación de regresión**, expresión matemática de la relación *ideal* entre las variables.

En cuanto al grado de relación, éste depende de la semejanza entre la nube de puntos y la ecuación de regresión. En este sentido, existen dos situaciones extremas: la primera se da cuando la nube de puntos se acopla perfectamente a la línea ideal, es decir, cuando cada valor de una de las variables queda perfectamente determinado con el conocimiento de la otra variable, estamos ante lo que se denomina **dependencia funcional**[2]; la segunda situación se produce cuando la nube de puntos es *amorfa*, reflejo de la existencia de independencia, definida anteriormente. Ahora bien, entre estas dos situaciones extremas hay diferentes grados de **dependencia estadística**.

Aunque la forma de la relación puede ser muy diversa, consideraremos únicamente la existencia de relación *lineal*[3], esto es, supondremos que la relación ideal entre las variables X e Y viene dada por la expresión de una recta que se denomina **recta de regresión.**

Utilizando el criterio de los **mínimos-cuadrados**, esto es, haciendo mínimas las distancias al cuadrado entre los valores de la nube de puntos —valores observados—, y_j, y los valores correspondientes a la ecuación de regresión —valores teóricos—, $\tilde{y}_i = a + b \cdot x_i$, esto es, haciendo mínima la expresión

$$\sum_{i=1}^{h} \sum_{j=1}^{k} \left(y_j - \tilde{y}_i\right)^2 f_{ij}^{\,4},$$

donde $y_j - \tilde{y}_i = e_{ij}$, son los *residuos* de la regresión, obtendremos, la **recta de regresión de *Y* sobre *X***, esto es, la mejor explicación lineal de los valores de la variable Y a partir de los valores de la variable X:

$$y - \overline{y} = \frac{S}{S_X^2}(x - \overline{x}).$$

De igual modo, la **recta de regresión de *X* sobre *Y***, es decir, la mejor explicación lineal de la variable X a partir de la variable Y responde a la ecuación:

$$x - \overline{x} = \frac{S}{S_Y^2}(y - \overline{y}).$$

Los coeficientes $b_{Y/X} = \dfrac{S}{S_X^2}$ y $b_{X/Y} = \dfrac{S}{S_Y^2}$ de las rectas de regresión reciben el nombre de **coeficientes de regresión**. Estos coeficientes tienen el mismo signo, que coincide con el de la covarianza entre X e Y, S.

---

[2] Esta situación no es habitual en ciencias sociales, pero es frecuente, en cambio, en el campo de las ciencias exactas.

[3] En los problemas que desarrollamos en este capítulo consideraremos también relaciones no lineales que pueden transformarse en lineales mediante sencillas operaciones matemáticas.

[4] Para minimizar tal expresión se deriva con respecto de a y de b, igualando los resultados a 0 y aplicando la condición de mínimo.

Para medir el grado de relación lineal, o grado de **correlación** entre las variables X e Y o, lo que es lo mismo, la bondad de la regresión lineal llevada a cabo, se utiliza el **coeficiente de determinación lineal**:

$$r^2 = \frac{S^2_{\tilde{Y}}}{S^2_Y},$$

donde $S^2_{\tilde{Y}}$ es la varianza de $\tilde{Y}$, variable cuyos valores son

$$\tilde{y}_i = \bar{y} + \frac{S}{S^2_X}(x_i - \bar{x}),$$

esto es, los valores teóricos proporcionados por la regresión lineal de Y sobre X[5].

El coeficiente de determinación lineal se interpreta, por tanto, como la proporción de la *variabilidad* de Y, o varianza de Y, explicada por la regresión, o varianza de $\tilde{Y}$[6].

El coeficiente de determinación lineal está acotado entre 0 y 1, siendo la relación entre X e Y de dependencia funcional —ajuste lineal perfecto—, si $r^2 = 1$ y de **incorrelación** o ausencia de relación lineal —ajuste lineal pésimo—, si $r^2 = 0$.

Para estudiar el grado de relación lineal también se utiliza el **coeficiente de correlación lineal**:

$$r = \frac{S}{S_X \cdot S_Y},$$

cuyo cuadrado es el coeficiente de determinación lineal.

Este coeficiente, que tiene el signo de la covarianza, S, puesto que el denominador es siempre positivo, está acotado entre -1 y 1, lo cual facilita su interpretación. Así,

• Si $r = 0$, esto es, si $r^2 = 0$, las variables X e Y están incorrelacionadas, es decir, no existe relación lineal entre ellas. En tal caso, como *necesariamente* la covarianza es cero, los coeficientes de regresión, $b_{Y/X}$ y $b_{X/Y}$, también son nulos, y las rectas de regresión son paralelas a los ejes de coordenadas:

$$y = \bar{y}$$
$$x = \bar{x}.$$

• Si $r = 1$, o bien $r = -1$, es decir, si $r^2 = 1$, las dos rectas de regresión coinciden, existiendo entonces dependencia funcional, creciente o decreciente, según el caso, entre las dos variables.

• Situaciones intermedias son indicativas de distintos grados de dependencia lineal entre las variables: creciente, si el coeficiente de correlación es positivo —covarianza positiva— o decreciente, si es negativo —covarianza negativa—.

---

[5] La interpretación del grado de correlación entre las variables X e Y se puede realizar igualmente a partir de la regresión lineal de X sobre Y.

[6] Análoga interpretación puede hacerse en el caso de la regresión lineal de X sobre Y.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

2.1 Recientemente, el departamento de Investigación y Desarrollo de los laboratorios farmacéuticos Balleras ha realizado un estudio sobre la influencia de la edad en el consumo de medicamentos. Para ello, eligió una muestra de 100 individuos, cuyas edades, junto con las cantidades, en euros, que gastaron en medicinas durante un año, aparecen recogidas en la siguiente tabla:

Edad / Gasto	0-15	15-30	30-60	60-100
0-30	5	7	5	3
30-90	12	2	15	21
90-180	3	1	10	16

a) Obténgase la distribución de frecuencias de la variable gasto en medicinas.

b) Hállese la distribución de frecuencias de la variable edad.

c) ¿Cuál es la distribución de frecuencias de la edad condicionada a un nivel de gasto comprendido entre 30 y 90 euros?

d) Calcúlese la distribución de frecuencias del gasto para una edad comprendida entre 60 y 100 años.

SOLUCIÓN

Esta tabla de doble entrada es la tabla de correlación correspondiente a la distribución de frecuencias bidimensional de las variables X, gasto, e Y, edad.

En el interior de la tabla aparecen las frecuencias absolutas conjuntas de las variables cuyas observaciones están agrupadas en clases. Así, por ejemplo, $n_{23} = 15$ significa que hay 15 individuos de la población con edades comprendidas entre 30 y 60 años y con un gasto en medicamentos entre 30 y 90 euros.

Puede comprobar el lector que la suma de las frecuencias absolutas conjuntas es igual a 100, número de unidades de la población.

a) De la distribución de frecuencias bidimensional pueden obtenerse las distribuciones de frecuencias marginales de las variables X e Y.

Los valores de la variable gasto, X, están agrupadas en los intervalos, 0-30, 30-90 y 90-180. Para obtener las frecuencias de cada intervalo hemos de fijarnos en que, por ejemplo, la última

fila de la tabla indica que hay 3 individuos con un gasto entre 90 y 180 euros y una edad entre 0 y 15 años; 1 individuo con un gasto entre 90 y 180 euros y una edad entre 15 y 30 años; 10 individuos con un gasto entre 90 y 180 euros y una edad entre 30 y 60 años, y, por último, 16 individuos con un gasto entre 90 y 180 euros y una edad entre 60 y 100 años. Puesto que los 100 individuos de la población están clasificados en grupos de edad y los intervalos considerados cubren toda la población, tendremos que

$$3 + 1 + 10 + 16 = 30$$

es el número de individuos de la población que tienen un gasto en medicinas entre 90 y 180 euros, esto es,

$$n_{3\cdot} = n_{31} + n_{32} + n_{33} + n_{34},$$

frecuencia absoluta del intervalo 90-180 de la variable gasto.

En general, la suma de los elementos de cada fila es igual a la frecuencia absoluta *marginal* correspondiente a cada intervalo de la variable X:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij\cdot}$$

Aplicando la expresión anterior para $i = 1,2$, se completa la distribución de frecuencias de la variable gasto según se recoge en la tabla siguiente.

Gastos	$n_{i\cdot}$
0-30	20
30-90	50
90-180	30

b) De igual forma, obtenemos la distribución de la variable edad, Y.

En efecto, ahora, la suma de los elementos de cada una de las columnas de la tabla de correlación proporciona la frecuencia marginal del correspondiente intervalo. Por ejemplo, si sumamos las cantidades de la segunda columna,

$$7 + 2 + 1 = 10,$$

llegamos a que hay 10 individuos de la población cuyas edades están comprendidas entre 15 y 30 años.

Este modo de proceder se resume mediante la expresión genérica:

$$n_{\cdot j} = \sum_{i=1}^{h} n_{ij},$$

que permite calcular las frecuencias marginales de la variable edad para los cuatro intervalos en los que están agrupadas las observaciones según esta variable.

Resulta, de tal modo, la distribución de frecuencias agrupada en intervalos de la variable edad:

Edad	$n_{.j}$
0-15	20
15-30	10
30-60	30
60-100	40

c) Cada fila de la tabla de correlación corresponde a un nivel de gasto y, por tanto, contiene las frecuencias absolutas de cada uno de los intervalos de edad *dentro* de dicho nivel de gasto.

Por consiguiente, la distribución de la variable edad condicionada por un gasto entre 30 y 90 euros es

Edad	$n_{j/2}$
0-15	12
15-30	2
30-60	15
60-100	21

Como puede observarse, la segunda columna de esta tabla coincide con la penúltima fila de la tabla de correlación. Por ejemplo, 15 son los individuos que tienen entre 30 y 60 años *y* unos gastos en medicinas entre 30 y 90 euros —interpretación de esta frecuencia como frecuencia absoluta conjunta—, pero también 15 es el número de individuos que tienen una edad entre 30 y 60 años *dentro* de los que tienen unos gastos en medicamentos entre 30 y 90 euros —concepción de esta frecuencia como frecuencia condicionada—.

Otro comentario de interés es que habríamos resuelto este apartado de igual manera, si se hubiera cuestionado sobre la distribución de la edad condicionada por un nivel de gasto igual a 60 euros, marca de clase del intervalo 30-90.

d) Del mismo modo que en el apartado anterior, las cifras de las columnas de la tabla de correlación se interpretan como las frecuencias de cada intervalo de la variable gasto dentro de un intervalo de edad fijo.

La distribución condicionada del gasto para una edad comprendida entre 60 y 100 años, aparece en la siguiente tabla:

Gastos	$n_{i/4}$
0-30	3
30-90	21
90-180	16

Obsérvese que, por ejemplo, la primera casilla de la tabla expresa que son 3 los individuos que, teniendo una edad comprendida entre 60 y 100 años, han gastado en medicinas entre 0 y 30 euros.

Al igual que se comentó en el apartado anterior, la distribución obtenida es también la distribución del gasto condicionada por una edad de 80 años, marca de clase del intervalo 60-100.

2.2 La siguiente tabla recoge los ingresos y los gastos en alimentación semanales, en euros, de 12 familias:

Gastos / Ingresos	30-60	60-90
120-300	4	2
300-480	1	5

Determínese el gasto medio por familia en alimentación de las familias con ingresos comprendidos entre 300 y 480 euros semanales.

SOLUCIÓN

Si denotamos por X e Y las variables ingresos y gastos en alimentación semanales, en euros, la distribución del gasto semanal en alimentación condicionada por un valor del ingreso igual a 390 euros —marca del clase del intervalo 300-480—, esto es, la distribución condicionada $(y_j/X = x_2; n_{j/2})$ es:

$y_j/X = x_2$	$n_{j/2}$
45	1
75	5

donde los valores y_j son las marcas de clase de los intervalos de la variable gasto en alimentación. Para obtener esta distribución de frecuencias *unidimensional* hemos considerado las frecuencias de los valores de la variable *Y dentro* de las familias cuyos ingresos están com-

prendidos entre 300 y 480 euros, es decir, hemos tomado la última fila de frecuencias de la tabla de correlación correspondiente a la distribución conjunta de las variables X e Y.

Así, el gasto medio pedido, es decir, la media de distribución anterior, $\bar{y}/(X = x_2)$, se calcula igual que hacíamos en el capítulo 1 para cualquier distribución de frecuencias unidimensional:

$$\bar{y}/(X = x_2) = \frac{1}{n_{2\cdot}} \sum_{j=1}^{k} y_j \cdot n_{j/2} = \frac{45 \cdot 1 + 75 \cdot 5}{6} = 70 \text{ euros.}$$

Nótese que, en este caso, el número de observaciones, 6, es $n_{21} + n_{22} = n_{2\cdot}$, frecuencia absoluta marginal del valor x_2.

2.3 Se realiza un estudio sobre la condición de los trabajadores de un sector del pequeño comercio, para lo cual se considera un grupo de 100 establecimientos. Sea X la variable que designa el número de trabajadores por establecimiento e Y la variable número de ellos que pertenecen a la familia propietaria del mismo. La siguiente tabla recoge la distribución conjunta de estas variables.

X \ Y	0	1	2
1	10	10	0
2	5	30	15
3	10	10	10

a) Hállese el número medio de trabajadores que pertenecen a la familia propietaria, dentro de los establecimientos que tienen 2 trabajadores.

b) Obténgase la mediana de la distribución calculada en el apartado anterior.

c) ¿Cuál es el número más frecuente de trabajadores que pertenecen a la familia propietaria dentro de los establecimientos que tienen 2 trabajadores?

SOLUCIÓN

a) Se trata de calcular la media de la distribución condicionada $(y_j/X = x_2;\ n_{j/2})$, con $x_2 = 2$, cuyos valores y frecuencias aparecen en la siguiente tabla:

$y_j/X = x_2$	$n_{j/2}$
0	5
1	30
2	15

Así, la media pedida es

$$\overline{y}/(X = x_2) = \frac{1}{n_{2\cdot}} \sum_{j=1}^{k} y_j \cdot n_{j/2} = \frac{0 \cdot 5 + 1 \cdot 30 + 2 \cdot 15}{50} = 1{,}2 \text{ trabajadores.}$$

b) Completamos la tabla anterior con una columna correspondiente a las frecuencias absolutas acumuladas de esta distribución condicionada.

$y_j/X = x_2$	$n_{j/2}$	$N_{j/2}$
0	5	5
1	30	35
2	15	50

La mediana de dicha distribución es el mínimo valor de la variable con frecuencia absoluta acumulada estrictamente mayor que $n_{2\cdot}/2 = 50/2 = 25$; en este caso, $y_2 = 1$ cuya frecuencia es 35. Por tanto,

$$Me_{Y/X = x_2} = 1 \text{ trabajador.}$$

c) Hemos de calcular la moda de la distribución condicionada, esto es, el valor de la variable Y con mayor frecuencia condicionada, que, para esta distribución, resulta ser $n_{2/2} = 30$, correspondiente al valor $y_2 = 1$. Por consiguiente,

$$Mo_{Y/X = x_2} = 1 \text{ trabajador.}$$

2.4 La siguiente tabla recoge la clasificación de 50 trabajadores de una empresa según el nivel de salario anual, en miles de euros, y el número de días de baja por enfermedad en un determinado año:

Salario / Días de baja	15-25	25-35	35-55
0-10	7	23	5
10-40	10	0	0
40-90	3	2	0

a) ¿Cuál es el número de días de baja esperados para un trabajador cuyo salario anual es de 20 mil euros?

b) Obténgase el número de días de baja más frecuente de los trabajadores con salarios anuales comprendidos entre 15 y 25 mil euros.

c) Hállese la mediana de la distribución calculada en el apartado ***a)***.

d) ¿Cuál de los promedios obtenidos es más representativo?

SOLUCIÓN

a) El número de días de baja esperados para un trabajador cuyo salario es de 20 mil euros puede interpretarse como el número medio de días de baja de los trabajadores con salarios comprendidos entre 15 y 25 mil euros, es decir, la media de la distribución de X condicionada por $Y = y_1$, donde $y_1 = 20$ es la marca de clase del intervalo 15-25:

Días de baja	$n_{i/1}$
0-10	7
10-40	10
40-90	3

Así, a partir de las marcas de clase, se obtiene:

$$\overline{x}/(Y = y_1) = \frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} x_i \cdot n_{i/1} = \frac{5 \cdot 7 + 25 \cdot 10 + 65 \cdot 3}{20} = 24 \text{ días.}$$

b) El intervalo modal, o intervalo de mayor densidad de frecuencia, de la distribución condicionada hallada en el apartado anterior es, como puede comprobar el lector, 0-10. Puesto que los intervalos tienen distinta amplitud, a la hora de calcular la moda de la distribución utilizaremos la expresión:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot c_i,$$

con lo cual,

$$Mo_{X/Y = y_1} = 0 + \frac{\frac{10}{30}}{0 + \frac{10}{30}} \cdot 10 = 10 \text{ días,}$$

que, por ser el intervalo modal el primer intervalo de la distribución, coincide con el extremo superior del mismo.

c) El intervalo mediano de la distribución condicionada es el intervalo 10-40, pues es el primer intervalo con frecuencia absoluta acumulada, $N_{2/1} = 17$, estrictamente mayor que $n_{\cdot 1}/2 = 20/2 = 10$.

Para calcular la mediana aplicamos la expresión habitual:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i$$

que, en el caso que nos ocupa, se convierte en

$$Me_{X/Y=y_1} = L_{i-1} + \frac{\frac{n_{\cdot 1}}{2} - N_{i-1/1}}{n_{i/1}} \cdot c_i.$$

Sustituyendo los datos del problema,

$$Me_{X/Y=y_1} = 10 + \frac{10-7}{10} \cdot 30 = 19 \text{ días.}$$

d) Para comparar la representatividad de los promedios de esta distribución condicionada utilizaremos los índices de dispersión de la media, de la moda y de la mediana, estudiados en el capítulo 1.

Así, por lo que se refiere a la media de la distribución condicionada, hallaremos:

$$I_{\bar{x}/(Y=y_1)} = \frac{\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} \left| x_i - \bar{x}/(Y=y_1) \right| \cdot n_{i/1}}{\bar{x}/(Y=y_1)};$$

en cuanto a la moda, obtendremos:

$$I_{Mo_{X/Y=y_1}} = \frac{\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} \left| x_i - Mo_{X/Y=y_1} \right| \cdot n_{i/1}}{Mo_{X/Y=y_1}};$$

y, por último, en relación a la mediana, calcularemos:

$$I_{Me_{X/Y=y_1}} = \frac{\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} \left| x_i - Me_{X/Y=y_1} \right| \cdot n_{i/1}}{Me_{X/Y=y_1}}.$$

Con los datos de la distribución condicionada, resultan los siguientes valores de las desviaciones absolutas medias:

$$\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} |x_i - \overline{x}/(Y = y_1)| \cdot n_{i/1} = \frac{1}{20} (|5 - 24| \cdot 7 + |25 - 24| \cdot 10 + |65 - 24| \cdot 3) = 13{,}3,$$

$$\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} |x_i - Mo_{X/Y = y_1}| \cdot n_{i/1} = \frac{1}{20} (|5 - 10| \cdot 7 + |25 - 10| \cdot 10 + |65 - 10| \cdot 3) = 17{,}5$$

y

$$\frac{1}{n_{\cdot 1}} \sum_{i=1}^{h} |x_i - Me_{X/Y = y_1}| \cdot n_{i/1} = \frac{1}{20} (|5 - 19| \cdot 7 + |25 - 19| \cdot 10 + |65 - 19| \cdot 3) = 14{,}8,$$

por lo que los respectivos índices de dispersión son:

$$I_{\overline{x}/(Y = y_1)} = \frac{13{,}3}{24} = 0{,}554,$$

$$I_{Mo_{X/Y = y_1}} = \frac{17{,}5}{10} = 1{,}75$$

e

$$I_{Me_{X/Y = y_1}} = \frac{14{,}8}{19} = 0{,}779.$$

Se puede afirmar, por consiguiente, que la media es el promedio más representativo de la distribución considerada, puesto que su índice de dispersión es el más pequeño.

2.5 En una empresa de limpieza, que cuenta con 100 trabajadores, se ha realizado un estudio sobre la relación entre el salario y el absentismo laboral, obteniéndose, entre otros, los resultados que aparecen en las siguientes tablas de distribuciones condicionadas:

$y_j/X = x_1$	$n_{j/1}$
y_1	5
y_2	20

$y_j/X = x_2$	$n_{j/2}$
y_1	15
y_2	10

$y_j/X = x_3$	$n_{j/3}$
y_1	50
y_2	0

La variable *Y* representa el número mensual de días de ausencia al trabajo y está distribuida en los intervalos 0-4 y 4-10; la variable *X* representa el salario mensual, en miles de euros, y está distribuida en los intervalos 0,6-1,2; 1,2-1,8 y 1,8-2,6.

a) Hállese la distribución bidimensional correspondiente.

b) Calcúlese el número medio mensual de días de absentismo por trabajador de los trabajadores con salarios comprendidos entre 1 200 y 1 800 euros.

c) Obténgase la varianza de la distribución del salario mensual de los trabajadores que se han ausentado del trabajo entre 4 y 10 días.

SOLUCIÓN

a) La siguiente tabla de correlación corresponde a la distribución de frecuencias bidimensional de las variables X e Y:

Absentismo / Salario	0-4	4-10
0,6-1,2	5	20
1,2-1,8	15	10
1,8-2,6	50	0

Como puede observarse, la primera columna y la primera fila de la tabla son, respectivamente, los intervalos en los que están agrupados los valores de las variables X, salario, e Y, absentismo.

Para construir el resto de la tabla, hemos tenido en cuenta que las frecuencias de la primera distribución condicionada que proporciona el enunciado $(y_j/X = x_1; n_{j/1})$ corresponden al número de observaciones iguales a y_j dentro de las que tienen un valor de la variable X igual a x_1, o, equivalentemente, según se comentó en **2.1**, a las frecuencias n_{11} y n_{12}; en general, $n_{j/1} = n_{1j}$, por lo que, variando el subíndice j, se obtienen las frecuencias de la primera fila de la tabla.

De igual forma, se obtienen la segunda y la tercera fila de la tabla anterior, a partir de las otras distribuciones condicionadas del enunciado.

b) El número medio de días de absentismo de los trabajadores con salarios entre 1 200 y 1 800 euros es la media aritmética de la distribución condicionada:

$y_j/X = x_2$	$n_{j/2}$
2	15
7	10

donde los valores de la variable Y son las marcas de clase de los intervalos 0-4 y 4-10.

Aplicando la expresión de la media aritmética a esta distribución de frecuencias unidimensional, se tiene:

$$\overline{y}/(X = x_2) = \frac{1}{n_{2\cdot}} \sum_{j=1}^{k} y_j \cdot n_{j/2} = \frac{2 \cdot 15 + 7 \cdot 10}{25} = 4 \text{ días.}$$

c) La varianza pedida es la correspondiente a la distribución condicionada de la variable X por el valor de Y igual a 7, marca de clase del intervalo 4-10:

$$S^2_{X/Y = y_2} = \frac{1}{n_{\cdot 2}} \sum_{i=1}^{h} x_i^2 \cdot n_{i/2} - (\overline{x}/(Y = y_2))^2,$$

donde el segundo sumando de la expresión anterior, media de la distribución condicionada, se obtiene como

$$\overline{x}/(Y = y_2) = \frac{1}{n_{\cdot 2}} \sum_{i=1}^{h} x_i \cdot n_{i/2}.$$

Sustituyendo, por los valores de la distribución condicionada, resulta:

$$S^2_{X/Y = y_2} = \frac{0{,}9^2 \cdot 20 + 1{,}5^2 \cdot 10 + 2{,}2^2 \cdot 0}{30} - \left(\frac{0{,}9 \cdot 20 + 1{,}5 \cdot 10 + 2{,}2 \cdot 0}{30}\right)^2 = 0{,}08.$$

2.6 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, demuéstrese que, para cualesquiera i y j,

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}}$$

y

$$f_{j/i} = \frac{f_{ij}}{f_{i\cdot}}.$$

SOLUCIÓN

La demostración es inmediata, teniendo en cuenta la definición de frecuencia relativa condicionada,

$$f_{i/j} = \frac{n_{i/j}}{n_{\cdot j}} = \frac{n_{ij}}{n_{\cdot j}},$$

y dividiendo numerador y denominador por N,

$$f_{i/j} = \frac{n_{ij}/N}{n_{\cdot j}/N} = \frac{f_{ij}}{f_{\cdot j}}.$$

Invitamos al lector a que resuelva la segunda parte del problema, aplicando análogo procedimiento.

2.7 La siguiente tabla recoge la distribución de frecuencias bidimensional de las variables X, ingresos, en millones de euros, en concepto de impuestos sobre vehículos, e Y, gastos en inversión de viales, en millones de euros, de un grupo de ayuntamientos.

Y / X	6,5-13,5	13,5-14,5	14,5-15,5
5-55	0,08	0,02	0,06
55-65	0,20	0,13	0,04
65-75	0,12	0,13	0,22

a) ¿Cuál es el ingreso medio por ayuntamiento en concepto de impuestos sobre vehículos en los ayuntamientos cuyos gastos en viales están comprendidos entre 13,5 y 14,5 millones de euros?

b) Obténgase el gasto medio por ayuntamiento en inversión de viales entre los ayuntamientos con ingresos por impuestos sobre vehículos entre 65 y 75 millones de euros

c) ¿Cuál de las dos medias es más representativa?

SOLUCIÓN

a) El ingreso medio pedido es la media de la distribución condicionada $(x_i/Y = y_2; f_{i/2})$, donde y_2 es la marca de clase del intervalo 13,5-14,5.

Como es sabido, los valores de esta distribución de frecuencias unidimensional son los valores de la variable X, siendo la frecuencia relativa del valor genérico x_i, marca de clase del intervalo genérico, igual a

$$f_{i/2} = \frac{f_{i2}}{f_{\cdot 2}},$$

según se demostró en **2.6**.

Se obtiene, así, la distribución de frecuencias condicionada que figura en la siguiente tabla:

$x_i/Y = y_2$	$f_{i/2}$
30	0,02/0,28=0,072
60	0,13/0,28=0,464
70	0,13/0,28=0,464

En definitiva, el ingreso medio de los ayuntamientos con una inversión en viales comprendida entre 13,5 y 14,5, media de la distribución condicionada, es

$$\overline{x}/(Y = y_2) = \sum_{i=1}^{h} x_i \cdot f_{i/2} = 62{,}48 \text{ millones de euros.}$$

b) El gasto medio por ayuntamiento en inversión de viales entre los ayuntamientos con ingresos por impuestos sobre vehículos entre 65 y 75 millones de euros es la media de la distribución condicionada ($y_j/X = x_3$; $f_{j/3}$), con $x_3 = 70$, marca de clase del intervalo 65-75.

Los valores de esta distribución son 10, 14 y 15, marcas de clase de los intervalos en los que están agrupados los datos de la variable *Y*, respondiendo las frecuencias relativas a la expresión genérica:

$$f_{j/3} = \frac{f_{3j}}{f_{3\cdot}},$$

como se demostró en **2.6**.
El resultado de aplicar la relación anterior a cada uno de los valores de la variable se recoge en la segunda columna de la tabla de la distribución condicionada.

$y_j/X = x_3$	$f_{j/3}$
10	0,12/0,47 = 0,255
14	0,13/0,47 = 0,277
15	0,22/0,47 = 0,468

Por consiguiente, la media de la distribución condicionada es

$$\overline{y}/(X = x_3) = \sum_{j=1}^{k} y_j \cdot f_{j/3} = 10 \cdot 0{,}255 + 14 \cdot 0{,}277 + 15 \cdot 0{,}468 = 13{,}448 \text{ millones de euros.}$$

c) Para estudiar la representatividad de las dos medias obtenidas en los apartados anteriores, hallaremos el coeficiente de variación de cada una de sus correspondientes distribuciones *unidimensionales* pues, según vimos en el capítulo anterior, ésta es una medida relativa de dispersión que permite realizar comparaciones.

El cálculo del coeficiente de variación de la primera distribución condicionada,

$$V_{X/Y=y_2} = \frac{S_{X/Y=y_2}}{\overline{x}/(Y=y_2)},$$

requiere la obtención de la desviación típica de dicha distribución, $S_{X/Y=y_2}$, a partir de la varianza de la misma. Así,

$$S^2_{X/Y=y_2} = \sum_{i=1}^{h} x_i^2 \cdot f_{i/2} - \left(\overline{x}/(Y=y_2)\right)^2 = 30^2 \cdot 0{,}072 + 60^2 \cdot 0{,}464 + 70^2 \cdot 0{,}464 - 62{,}48^2 = 105{,}05,$$

con lo cual, la desviación típica es

$$S_{X/Y=y_2} = \sqrt{S^2_{X/Y=y_2}} = \sqrt{105{,}05} = 10{,}25 \text{ millones de euros.}$$

En conclusión, el coeficiente de variación de la distribución de los ingresos en concepto de impuestos sobre vehículos en los ayuntamientos cuyos gastos en viales están comprendidos entre 13,5 y 14,5 millones de euros es

$$V_{X/Y=y_2} = \frac{10{,}25}{62{,}48} = 0{,}16.$$

Un proceso análogo para la distribución condicionada $(y_j/X = x_3; f_{j/3})$ permite calcular, también, su coeficiente de variación,

$$V_{Y/X=x_3} = \frac{S_{Y/X=x_3}}{\overline{y}/(X=x_3)},$$

sin más que tener en cuenta que

$$S^2_{Y/X=x_3} = \sum_{j=1}^{k} y_j^2 \cdot f_{j/3} - \left(\overline{y}/(X=x_3)\right)^2 = 10^2 \cdot 0{,}255 + 14^2 \cdot 0{,}277 + 15^2 \cdot 0{,}468 - 13{,}448^2 = 4{,}24,$$

con lo cual,

$$S_{Y/X=x_3} = \sqrt{S^2_{Y/X=x_3}} = \sqrt{4{,}24} = 2{,}06 \text{ millones de euros}$$

y, por tanto,

$$V_{Y/X=x_3} = \frac{2{,}06}{13{,}448} = 0{,}15.$$

Aunque este coeficiente de variación es más pequeño que el de la distribución anterior, por lo que la correspondiente media es más representativa, en realidad, al ser la diferencia tan escasa, podemos afirmar que ambas medias tienen prácticamente la misma representatividad en sus respectivas distribuciones.

2.8 De un estudio realizado para la revista Inversión sobre una muestra de suscriptores de dicha publicación, se ha observado que el 30 por ciento de ellos tiene una renta anual entre 36 y 60 mil euros y el 55 por ciento invierte anualmente en bolsa entre 900 y 3 000 euros.

Además, el 9 por ciento invierte de 600 a 900 euros en bolsa y tiene una renta entre 12 y 24 mil euros; el 20 por ciento invierte de 900 a 3 000 euros y percibe una renta entre 12 y 24 mil euros, por último, el 6 por ciento invierte de 600 a 900 euros en bolsa y tiene una renta entre 36 y 60 mil euros.

a) Calcúlese la cantidad media por individuo invertida en bolsa de los individuos con una renta comprendida entre 24 y 36 mil euros.

b) Obténgase la varianza de la distribución de la inversión condicionada a un valor de renta igual a 48 mil euros.

SOLUCIÓN

La información concerniente a las dos variables se recoge en la siguiente tabla de doble entrada donde los porcentajes se han sustituido por proporciones, esto es, por frecuencias relativas. Como puede observarse, los datos 30 por ciento y 55 por ciento, 0,3 y 0,55, son frecuencias marginales, mientras que el resto de los porcentajes corresponden a frecuencias conjuntas.

Inversión / Renta	600-900	900-3 000	
12-24	0,09	0,20	
24-36			
36-60	0,06		0,30
		0,55	

El hecho de que, por un lado, tanto las frecuencias relativas conjuntas como las frecuencias relativas marginales de cada variable sumen la unidad y de que, por otro lado, las frecuencias relativas marginales se obtengan a partir de las frecuencias relativas conjuntas, permite completar esta tabla, es decir, obtener la distribución conjunta de las variables renta, X, en miles de euros, e inversión, Y, en euros, con las cantidades que aparecen en negrita.

X \ Y	600-900	900-3 000	$f_{i\cdot}$
12-24	0,09	0,20	0,29
24-36	0,30	0,11	0,41
36-60	0,06	0,24	0,30
$f_{\cdot j}$	0,45	0,55	1

La última columna y la última fila son, respectivamente, las frecuencias relativas marginales de las variables X e Y.

a) La cantidad media por individuo invertida en bolsa de los individuos con una renta comprendida entre 24 y 36 mil euros es la media de la distribución condicionada de la variable Y por el valor $x_2 = 30$, marca de clase del intervalo 24-36; esta distribución se recoge en la siguiente tabla:

$y_j/X = x_2$	$f_{j/2}$
750	0,30/0,41=0,73
1 950	0,11/0,41=0,27

Los elementos de la segunda columna de la tabla anterior, frecuencias relativas de la distribución condicionada, se han calculado teniendo en cuenta la expresión general:

$$f_{j/2} = \frac{f_{2j}}{f_{2\cdot}}.$$

Obtenida la distribución condicionada, la media de dicha distribución es

$$\overline{y}/(X = x_2) = \sum_{j=1}^{k} y_j \cdot f_{j/2},$$

que, para los datos del problema, toma el valor:

$$\overline{y}/(X = x_2) = 750 \cdot 0{,}73 + 1\,950 \cdot 0{,}27 = 1\,074 \text{ euros}.$$

b) La distribución de la inversión condicionada a un valor de renta igual a 48 mil euros es la que figura en la siguiente tabla:

$y_j/X = x_3$	$f_{j/3}$
600-900	0,06/0,30=0,20
900-3 000	0,24/0,30=0,80

Como puede comprobar el lector, los elementos de la segunda columna de esta tabla se obtienen según la expresión:

$$f_{j/3} = \frac{f_{3j}}{f_{3\cdot}}.$$

La varianza de la distribución condicionada es

$$S^2_{Y/X\,=\,x_3} = \sum_{j\,=\,1}^{k} y_j^2 \cdot f_{j/3} - \left(\overline{y}/(X = x_3)\right)^2,$$

donde $\overline{y}/(X = x_3)$ es la media de la distribución condicionada.

En definitiva, sustituyendo los valores calculados, se tiene que

$$S^2_{Y/X\,=\,x_3} = 750^2 \cdot 0{,}20 + 1\,950^2 \cdot 0{,}80 - (750 \cdot 0{,}2 + 1\,950 \cdot 0{,}80)^2 = 230\,400.$$

2.9 Una promotora considera que las familias adquieren viviendas de mayor tamaño según sus ingresos. Para confirmar este hecho se han considerado los datos correspondientes a su última promoción de 210 viviendas, analizándose el nivel de ingresos anuales de las familias que han adquirido una vivienda de esta promoción, *X*, en miles de euros, así como el tamaño de la vivienda comprada, *Y,* en metros cuadrados.

Tamaño / Ingresos	40-100	100-200
12-24	90	10
24-30	15	20
30-40	5	70

¿Confirma esta información la hipótesis de la promotora?

SOLUCIÓN

La promotora sospecha que hay dependencia entre las dos variables. Para comprobar si está en lo cierto, hay que tener en cuenta que dos variables son independientes si, para cualesquiera *i* y *j*, se cumple que

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$$

o, equivalentemente,

$$\frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N} \cdot \frac{n_{\cdot j}}{N}.$$

Simplificando, la igualdad anterior se convierte en

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

condición que denominaremos *condición de independencia* y que, cuando las variables son independientes, se cumple para todos los pares (i, j).

En este caso, si tomamos, por ejemplo, $i = 1$ y $j = 1$ y calculamos

$$n_{1\cdot} = n_{11} + n_{12} = 90 + 10 = 100$$

y

$$n_{\cdot 1} = n_{11} + n_{21} + n_{31} = 90 + 15 + 5 = 110,$$

resulta que, por un lado,

$$n_{11} = 90$$

y, por otro,

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{100 \cdot 110}{210} = 52{,}38,$$

con lo que podemos afirmar que las variables no son independientes, confirmándose la hipótesis de la promotora.

2.10 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, pruébese que la condición necesaria y suficiente para que las variables X e Y sean independientes es que, para cualesquiera i y j:

$$f_{i/j} = f_{i\cdot}$$

y

$$f_{j/i} = f_{\cdot j}.$$

SOLUCIÓN

Si X e Y son independientes se cumple, por definición, que, para cualesquiera i y j,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

por lo cual, teniendo en cuenta lo visto en **2.6**, resulta, de modo inmediato, que

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{f_{\cdot j}} = f_{i\cdot},$$

según queríamos demostrar.

Recíprocamente, si, para cualesquiera i y j,

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}}$$

se tiene, sin más que despejar, que

$$f_{ij} = f_{i/j} \cdot f_{\cdot j}.$$

Ahora bien, si

$$f_{i/j} = f_{i\cdot},$$

entonces, sustituyendo en la expresión de la frecuencia relativa conjunta, se concluye que, para cualesquiera i y j,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

es decir, las variables X e Y son independientes.

Proponemos al lector la demostración de la doble implicación tomando como referencia la otra distribución condicionada.

2.11 La siguiente tabla refleja el salario mensual, X, en miles de euros, y el gasto médico al mes en odontólogos, Y, en euros, de un grupo de 200 familias.

Gasto / Salario	0-50	50-100	100-200
1-2	15	24	21
2-4	35	56	49

¿Son las variables X e Y independientes?

SOLUCIÓN

En **2.10** se demostró que las variables X e Y son independientes si, y solamente si, para cada par (i, j), se cumple que

$$f_{i/j} = f_{i\cdot},$$

condición que, en este caso, se traduce en las siguientes igualdades:

$$f_{1/1} = f_{1/2} = f_{1/3} = f_{1\cdot}$$

y

$$f_{2/1} = f_{2/2} = f_{2/3} = f_{2\cdot},$$

igualdades que, sustituyendo por frecuencias absolutas, son equivalentes a

$$\frac{n_{11}}{n_{\cdot 1}} = \frac{n_{12}}{n_{\cdot 2}} = \frac{n_{13}}{n_{\cdot 3}} = \frac{n_{1\cdot}}{N}$$

y

$$\frac{n_{21}}{n_{\cdot 1}} = \frac{n_{22}}{n_{\cdot 2}} = \frac{n_{23}}{n_{\cdot 3}} = \frac{n_{2\cdot}}{N}.$$

Estas expresiones indican que, si las variables son independientes, la proporción de unidades cuyo valor de la variable X está en un determinado intervalo se mantiene constante dentro de cada uno de los intervalos en los que están agrupados los valores de la variable Y; además, esta proporción coincide con la proporción de unidades de la población que tienen valores de X en dicho intervalo. Y, si esto es así, el hecho de que las variables sean independientes implica que las distribuciones de X condicionadas por los distintos valores —marcas de clase de Y— son todas *iguales* pues tienen los mismos valores —los de la variable X— y las mismas frecuencias *relativas*.

La siguiente tabla contiene las frecuencias marginales de las dos variables:

Y / X	0-50	50-100	100-200	$n_{i.}$
1-2	15	24	21	60
2-4	35	56	49	140
$n_{.j}$	50	80	70	200

A partir de estos datos se comprueba que

$$\frac{15}{50} = \frac{24}{80} = \frac{21}{70} = \frac{60}{200}$$

y

$$\frac{35}{50} = \frac{56}{80} = \frac{49}{70} = \frac{140}{200},$$

por lo que las variables X e Y son independientes.

De forma alternativa, se puede comprobar que estas variables son independientes utilizando la segunda condición equivalente a la condición de independencia de **2.10**, esto es, X e Y son independientes si, y solamente si, para cualesquiera i y j, se cumple que

$$f_{j/i} = f_{.j}.$$

En definitiva, es condición necesaria y suficiente para que dos variables sean independientes que las filas y las columnas de la tabla de correlación de su distribución conjunta sean *proporcionales*.

2.12 Sobre una población de N familias se ha realizado un estudio sobre la relación entre el número mensual de llamadas telefónicas nacionales (urbanas e interurbanas), X, y las internacionales, Y, y se han obtenido, entre otros resultados, las dos distribuciones de Y condicionadas por valores de X, tal y como se refleja en las siguientes tablas del mes de diciembre del pasado año:

$y_j/X = x_1$	$n_{j/1}$
y_1	12
y_2	24
y_3	36

$y_j/X = x_2$	$n_{j/2}$
y_1	10
y_2	a
y_3	b

a) Suponiendo que X está distribuida en los intervalos 0-60 y 60-240, y la variable Y en 0-20, 20-40 y 40-60, calcúlese el número medio por familia de llamadas internacionales de las familias que han realizado 30 llamadas nacionales.

b) Si las variables X e Y son independientes, ¿cuánto valen a y b?

SOLUCIÓN

a) Un número igual a 30 llamadas internacionales se corresponde con la marca de clase del primer intervalo de la variable X, por tanto, hay que calcular la media de la distribución de Y condicionada por $X = x_1$, es decir, $\overline{y}/(X = x_1)$.

Colocando las marcas de clase de los intervalos 0-20, 20-40 y 40-60 en la primera distribución condicionada que proporciona el enunciado, tendremos:

$y_j/X = x_1$	$n_{j/1}$
10	12
30	24
50	36

En definitiva,

$$\overline{y}/(X = x_1) = \frac{1}{n_{1\cdot}} \sum_{j=1}^{k} y_j \cdot n_{j/1} = \frac{1}{72} (10 \cdot 12 + 30 \cdot 24 + 50 \cdot 36) = 36{,}67,$$

número medio por familia de llamadas internacionales de las familias que han realizado 30 llamadas nacionales.

b) Según vimos en **2.11**, si X e Y son independientes, las dos distribuciones condicionadas $(y_j/X = x_1; n_{j/1})$ e $(y_j/X = x_2; n_{j/2})$, distribuciones unidimensionales, habrán de ser iguales y, por tanto, tendrá que cumplirse, para $j = 1, 2, 3$, que

$$f_{j/1} = f_{j/2}$$

o, equivalentemente, para $j = 1, 2, 3$, que

$$\frac{n_{j/1}}{n_{1\cdot}} = \frac{n_{j/2}}{n_{2\cdot}}.$$

Si aplicamos esta condición para $j = 2$ y $j = 3$, tendremos que, para que X e Y sean independientes, debería verificarse, por un lado,

$$\frac{n_{1/1}}{n_{1\cdot}} = \frac{n_{1/2}}{n_{2\cdot}}$$

y, por otro,

$$\frac{n_{2/1}}{n_{1.}} = \frac{n_{2/2}}{n_{2.}}.$$

Sustituyendo las frecuencias condicionadas por los datos de las distribuciones y teniendo en cuenta que $n_{1.}$ y $n_{2.}$ se hallan sumando las frecuencias absolutas condicionadas de las dos distribuciones condicionadas que proporciona el enunciado, esto es, $n_{1.} = 12 + 24 + 36 = 72$ y $n_{2.} = 10 + a + b$ se tiene que

$$\frac{24}{72} = \frac{a}{10 + a + b}$$

y

$$\frac{36}{72} = \frac{b}{10 + a + b}.$$

Resulta, de este modo, un sistema de dos ecuaciones con dos incógnitas:

$$\begin{aligned} 2 \cdot a - b &= 10 \\ a - b &= 10, \end{aligned}$$

cuya resolución conduce a los valores:

$$a = 20$$

y

$$b = 30.$$

Otra forma de resolver este problema consiste en obtener, en primera instancia, la distribución bidimensional correspondiente a las distribuciones condicionadas del enunciado. Así, colocando en la primera fila de la tabla de correlación las frecuencias de la primera distribución condicionada ($y_j/X = x_1$; $n_{j/1}$), y, en la segunda fila, las de la segunda distribución condicionada ($y_j/X = x_2$; $n_{j/2}$), resulta:

Y / X	y_1	y_2	y_3	$n_{i.}$
x_1	12	24	36	72
x_2	10	a	b	$10 + a + b$
$n_{.j}$	22	$24 + a$	$36 + b$	N

Si las variables X e Y son independientes, ha de cumplirse, para cualesquiera i y j, la condición de independencia:

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}.$$

Aplicando esta condición a $i = 1$ y $j = 1$, por ejemplo, y despejando N se tiene que

$$N = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n_{11}} = \frac{72 \cdot 22}{12} = 132.$$

Si aplicamos de nuevo la condición de independencia, ahora para $i = 1$ y $j = 2$,

$$n_{12} = \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N},$$

obtendremos:

$$24 = \frac{72\ (24 + a)}{132},$$

de donde resulta un valor de a igual a 20.

Para hallar b, basta tener en cuenta que

$$N = 132 = n_{1\cdot} + n_{2\cdot} = 72 + 10 + a + b + = 72 + 10 + 20 + b,$$

de lo cual se deduce que b es igual a 30, como ya sabíamos.

Observe el lector que, para que las variables sean independientes, las distribuciones condicionadas, según comentamos al principio de este apartado, han de ser iguales y, por tanto, han de ser proporcionales las columnas —distribuciones de la variable X condicionadas por los distintos valores de la variable Y— y las filas —distribuciones de Y condicionadas a los distintos valores de X— de la tabla de correlación.

En este sentido, si nos fijamos en las columnas de la tabla, vemos que los primeros elementos de cada columna se obtienen unos a partir de otros: la segunda casilla es dos veces la primera y la tercera, tres veces la primera. Por tanto, para que se mantenga esa proporcionalidad, necesaria y suficiente para que las variables sean independientes, tiene que ser a igual a dos veces 10 y b igual a tres veces 10, como hemos probado por otros caminos.

2.13 Obténganse las expresiones de los momentos $a_{r,0}$ y $a_{0,s}$. En particular, calcúlense los momentos, $a_{0,0}$, $a_{1,0}$, $a_{0,1}$, $a_{2,0}$ y $a_{0,2}$.

SOLUCIÓN

El momento bidimensional respecto al origen de orden (r, s) de la distribución bidimensional $(x_i, y_j; f_{ij})$ se define como

$$a_{r,s} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^s \cdot f_{ij}.$$

Si sustituimos por el valor $s = 0$, tendremos:

$$a_{r,0} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^0 \cdot f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot f_{ij}.$$

Puesto que x_i^r no depende de j y, además, $\sum_{j=1}^{k} f_{ij} = f_{i\cdot}$, se tiene que

$$a_{r,0} = \sum_{i=1}^{h} x_i^r \sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h} x_i^r \cdot f_{i\cdot} = a_r\,(x),$$

momento unidimensional respecto al origen de orden r de la distribución marginal $(x_i; f_{i\cdot})$.

En particular,

$$a_{1,0} = a_1\,(x) = \bar{x}$$

y

$$a_{2,0} = a_2\,(x).$$

Análogamente, al reemplazar el valor $r = 0$ en la expresión del momento bidimensional respecto al origen de orden (r, s) de la distribución bidimensional $(x_i, y_j; f_{ij})$, resulta:

$$a_{0,s} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^0 \cdot y_j^s \cdot f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} y_j^s \cdot f_{ij} = \sum_{j=1}^{k} y_j^s \sum_{i=1}^{h} f_{ij} = \sum_{j=1}^{k} y_j^s \cdot f_{\cdot j},$$

expresión del momento unidimensional respecto al origen de orden s de la distribución marginal $(y_j; f_{\cdot j})$, esto es, $a_s\,(y)$.

En tal caso,

$$a_{0,1} = a_1(y) = \overline{y}$$

y

$$a_{0,2} = a_2(y).$$

Por último,

$$a_{0,0} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^0 \cdot y_j^0 \cdot f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = 1.$$

2.14 Obténgase las expresiones de los momentos $m_{r,0}$ y $m_{0,s}$. En particular, calcúlense los momentos, $m_{0,0}$, $m_{1,0}$, $m_{0,1}$, $m_{2,0}$, $m_{0,2}$, y $m_{1,1}$.

SOLUCIÓN

El momento bidimensional respecto a las medias de orden (r, s) de la distribución bidimensional $(x_i, y_j; f_{ij})$ es

$$m_{r,s} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^s f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r f_{ij}.$$

Al sustituir s por 0 se obtiene:

$$m_{r,0} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^0 f_{ij}.$$

Puesto que $(x_i - \overline{x})^r$ no depende de j y $\sum_{j=1}^{k} f_{ij} = f_{i\cdot}$, resulta:

$$m_{r,0} = \sum_{i=1}^{h} (x_i - \overline{x})^r \sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h} (x_i - \overline{x})^r f_{i\cdot},$$

expresión que corresponde al momento respecto a la media de orden r de la distribución marginal $(x_i; f_{i\cdot})$, es decir, $m_r(x)$.

De manera semejante, sustituyendo por $r = 0$, se obtiene el momento respecto a la media de orden s de la distribución marginal $(y_j; f_{\cdot j})$, $m_s(y)$:

$$m_{0,s} = \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})^0 \cdot (y_j - \overline{y})^s f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j - \overline{y})^s f_{ij} = \sum_{j=1}^{k}(y_j - \overline{y})^s \sum_{i=1}^{h} f_{ij} = \sum_{j=1}^{k}(y_j - \overline{y})^s f_{\cdot j},$$

con lo cual,

$$m_{1,0} = m_1(x) = 0$$

y

$$m_{0,1} = m_1(y) = 0.$$

Además, los momentos de orden 2 son

$$m_{2,0} = m_2(x) = S_X^2$$

y

$$m_{0,2} = m_2(y) = S_Y^2,$$

es decir, las varianzas de X e Y.

Por último,

$$m_{0,0} = \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})^0 \cdot (y_j - \overline{y})^0 f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij} = 1$$

y

$$m_{1,1} = \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})^1 \cdot (y_j - \overline{y})^1 f_{ij} = S,$$

covarianza entre X e Y.

2.15 Demuéstrese la siguiente relación:

$$m_{1,1} = a_{1,1} - a_{1,0} \cdot a_{0,1}.$$

SOLUCIÓN

El momento bidimensional con respecto a las medias de orden (1, 1) de la distribución bidimensional $(x_i, y_j; f_{ij})$, esto es, la covarianza entre X e Y, es, por definición,

$$m_{1,1} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_j - \overline{y}) f_{ij}.$$

Operando en la expresión anterior y dividiendo el doble sumatorio en cuatro sumatorios dobles, resulta:

$$m_{1,1} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i \cdot y_j - x_i \cdot \overline{y} - \overline{x} \cdot y_j + \overline{x} \cdot \overline{y}) f_{ij} =$$

$$= \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{ij} - \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot \overline{y} \cdot f_{ij} - \sum_{i=1}^{h} \sum_{j=1}^{k} \overline{x} \cdot y_j \cdot f_{ij} + \sum_{i=1}^{h} \sum_{j=1}^{k} \overline{x} \cdot \overline{y} \cdot f_{ij}.$$

Puesto que $\overline{x}$ y $\overline{y}$ son constantes y, además, x_i no depende de j e y_j no depende de i pueden colocarse fuera de los correspondientes sumatorios. Así,

$$m_{1,1} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{ij} - \overline{y} \sum_{i=1}^{h} x_i \sum_{j=1}^{k} f_{ij} - \overline{x} \sum_{j=1}^{k} y_j \sum_{i=1}^{h} f_{ij} + \overline{x} \cdot \overline{y} \sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij}.$$

Teniendo en cuenta, además, que

$$\sum_{j=1}^{k} f_{ij} = f_{i\cdot}, \sum_{i=1}^{h} f_{ij} = f_{\cdot j} \text{ y } \sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = 1,$$

entonces,

$$m_{1,1} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{ij} - \overline{y} \sum_{i=1}^{h} x_i \cdot f_{i\cdot} - \overline{x} \sum_{j=1}^{k} y_j \cdot f_{\cdot j} + \overline{x} \cdot \overline{y}.$$

Considerando, por último, que $\sum_{i=1}^{h} x_i \cdot f_{i\cdot} = \overline{x}$ y $\sum_{j=1}^{k} y_j \cdot f_{\cdot j} = \overline{y}$ y simplificando, resulta:

$$m_{1,1} = a_{1,1} - \overline{x} \cdot \overline{y} - \overline{x} \cdot \overline{y} + \overline{x} \cdot \overline{y} = a_{1,1} - a_{1,0} \cdot a_{0,1},$$

según queríamos demostrar.

2.16 Sea $(x_i, y_j; f_{ij})$ una distribución de frecuencias bidimensional. Demuéstrese que las variables X e Y son independientes si, y solamente si, para cualesquiera i y l, el cociente

$$\frac{f_{ij}}{f_{lj}}$$

es constante para todo j.

SOLUCIÓN

Si las variables X e Y son independientes, el cociente de frecuencias conjuntas resulta ser igual a

$$\frac{f_{ij}}{f_{lj}} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{f_{l\cdot} \cdot f_{\cdot j}} = \frac{f_{i\cdot}}{f_{l\cdot}},$$

valor constante, sea cual sea j.
Recíprocamente, si el cociente f_{ij}/f_{lj} es constante para todo j, se cumple que[1]

$$\frac{f_{i1}}{f_{l1}} = \ldots = \frac{f_{ik}}{f_{lk}} = \frac{\sum_{j=1}^{k} f_{ij}}{\sum_{j=1}^{k} f_{lj}}.$$

Ahora bien, el numerador del último miembro de la igualdad anterior es

$$\sum_{j=1}^{k} f_{ij} = f_{i\cdot}$$

y, el denominador

$$\sum_{j=1}^{k} f_{lj} = f_{l\cdot},$$

con lo cual, para todo j, se cumple que

$$\frac{f_{ij}}{f_{lj}} = \frac{f_{i\cdot}}{f_{l\cdot}}.$$

---

[1] Por las propiedades de las fracciones, se sabe que, si

$$\frac{a_1}{b_1} = \ldots = \frac{a_n}{b_n},$$

entonces, estos cocientes son iguales a

$$\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}.$$

En definitiva, reordenando los términos de la igualdad anterior, se tiene, para cualquier j y para cualesquiera i y l, que

$$\frac{f_{ij}}{f_{i\cdot}} = \frac{f_{lj}}{f_{l\cdot}},$$

o, lo que es igual,

$$\frac{f_{1j}}{f_{1\cdot}} = \ldots = \frac{f_{hj}}{f_{h\cdot}} = \frac{\sum_{i=1}^{h} f_{ij}}{\sum_{i=1}^{h} f_{i\cdot}},$$

siendo la última igualdad el resultado de aplicar la nota anterior.

Como $\sum_{i=1}^{h} f_{ij} = f_{\cdot j}$ y $\sum_{i=1}^{h} f_{i\cdot} = 1$, entonces, para cualesquiera i y j,

$$\frac{f_{ij}}{f_{i\cdot}} = \frac{f_{\cdot j}}{1} = f_{\cdot j},$$

es decir, para cualesquiera i y j,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

quedando demostrado que las variables X e Y son independientes.

2.17 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, obténgase la expresión del momento bidimensional con respecto al origen de orden (r, s) en el caso de que las variables X e Y sean independientes.

SOLUCIÓN

Si X e Y son independientes, entonces, para cualesquiera i y j,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

con lo cual, sustituyendo la igualdad anterior en la expresión del momento bidimensional con respecto al origen de orden (r, s) de la distribución bidimensional, se tiene que

$$a_{r,s} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^s \cdot f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \cdot y_j^s \cdot f_{i\cdot} \cdot f_{\cdot j} = \left(\sum_{i=1}^{h} x_i^r \cdot f_{i\cdot} \right) \cdot \left(\sum_{j=1}^{k} y_j^s \cdot f_{\cdot j} \right),$$

siendo la última igualdad resultado de agrupar términos afines.

En consecuencia, si X e Y son independientes, se cumple:

$$a_{r,s} = a_{r,0} \cdot a_{0,s}.$$

2.18 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, obténgase la expresión del momento bidimensional con respecto a las medias de orden (r, s) en el caso de que las variables X e Y sean independientes.

SOLUCIÓN

Aplicando la condición de independencia entre las variables X e Y,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

para cualesquiera i y j, a la expresión del momento bidimensional con respecto a las medias de orden (r, s) de la distribución bidimensional, obtenemos que

$$m_{r,s} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^s f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x})^r \cdot (y_j - \overline{y})^s f_{i\cdot} \cdot f_{\cdot j}.$$

Agrupando términos semejantes, resulta que, si X e Y son independientes, entonces,

$$m_{r,s} = \left[\sum_{i=1}^{h} (x_i - \overline{x})^r f_{i\cdot} \right] \cdot \left[\sum_{j=1}^{k} (y_j - \overline{y})^s f_{\cdot j} \right] = m_{r,0} \cdot m_{0,s}.$$

2.19 ¿Cuánto vale la covarianza de una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, cuando X e Y son variables independientes?

SOLUCIÓN

Hemos demostrado, por un lado, que

$$m_{1,1} = a_{1,1} - a_{1,0} \cdot a_{0,1}$$

y, por otro lado, cuando X e Y son independientes, aplicando **2.17**, se tiene:

$$a_{1,1} = a_{1,0} \cdot a_{0,1}.$$

Por lo tanto, $m_{1,1} = a_{1,0} \cdot a_{0,1} - a_{1,0} \cdot a_{0,1} = 0$, es decir, si X e Y son independientes su covarianza es cero.

Aunque más adelante estudiaremos con detalle el coeficiente de correlación lineal:

$$r = \frac{S}{S_X \cdot S_Y},$$

merece la pena que el lector caiga en la cuenta de que la independencia entre las variables X e Y implica que r es igual a cero, existiendo incorrelación entre las variables: si entre X e Y *no* existe ningún tipo de relación no puede existir relación *lineal*.

2.20 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, interprétese el signo de la covarianza, S.

SOLUCIÓN

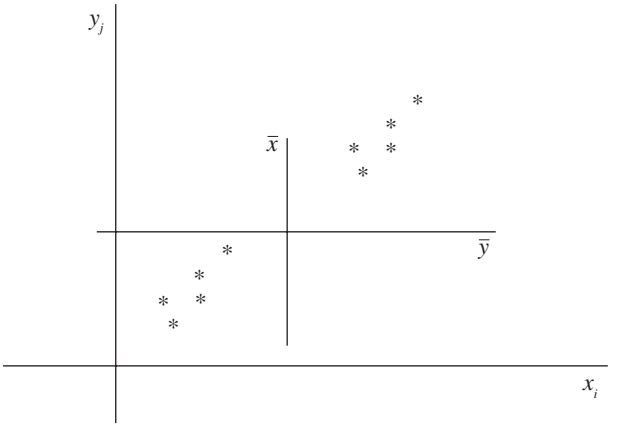
La covarianza, o *varianza conjunta* de dos variables,

$$S = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \bar{x}) \cdot (y_j - \bar{y}) f_{ij},$$

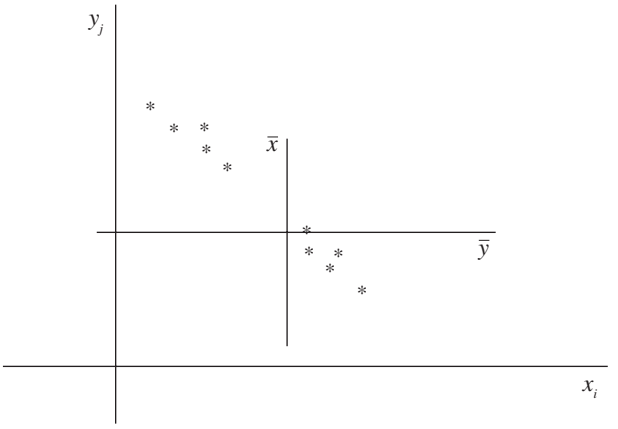
mide la variación *lineal* entre las variables, puesto que con ella se calculan diferencias *de primer orden* entre los valores de las distribuciones de cada variable y su respectiva media.

Cuando la nube de puntos adopta un aspecto como el de la figura siguiente, entonces, las desviaciones positivas de los valores de la distribución de X con respecto a $\bar{x}$ se acompañan con desviaciones también positivas de los valores de la distribución de Y con respecto a $\bar{y}$; a su vez, las desviaciones negativas de los valores de la distribución de X con respecto a su media se acompañan con desviaciones negativas de los valores de la distribución de Y con respecto a la suya, con lo cual, los factores $(x_i - \bar{x}) \cdot (y_j - \bar{y})$ serán positivos en ambos casos, siendo enton-

ces la covarianza igualmente *positiva*. Estamos, en tal caso, ante una relación *creciente* entre las variables.

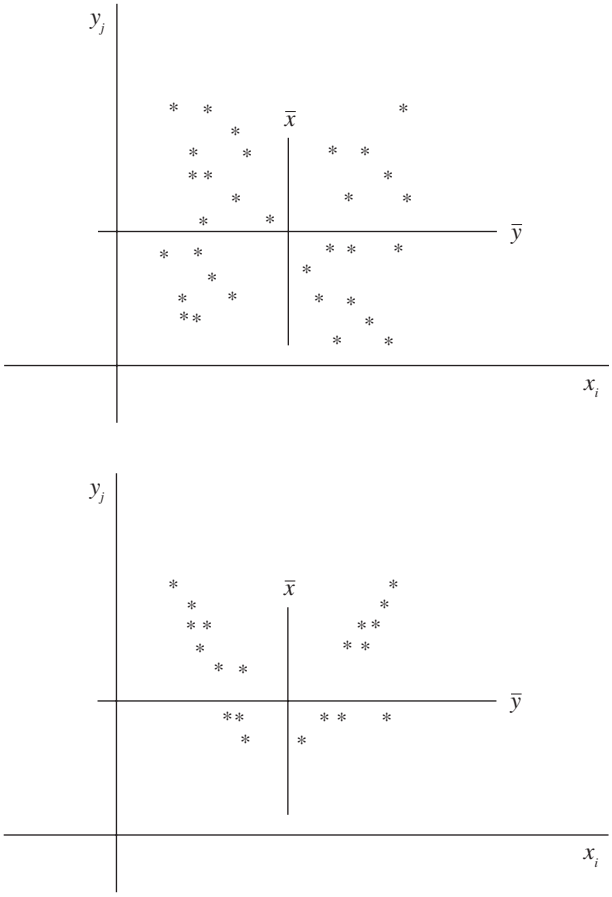


Cuando, por el contrario, el diagrama de dispersión de la distribución bidimensional se asemeja al de la siguiente figura, entonces, las diferencias $(x_i - \bar{x})$ positivas se acompañan con diferencias $(y_j - \bar{y})$ negativas y, recíprocamente, diferencias $(x_i - \bar{x})$ negativas se acompañan con diferencias $(y_j - \bar{y})$ positivas, siendo en tal situación, la covarianza *negativa* y la relación entre las variables *decreciente*.



Por último, si desviaciones positivas de los valores de la distribución de la variable X en relación con su media se acompañan con desviaciones, unas veces positivas y otras veces negativas, de los valores de la distribución de la variable Y con respecto a su media, y, viceversa, desviaciones positivas $(x_i - \bar{x})$ se acompañan con desviaciones $(y_j - \bar{y})$ unas veces positivas y otras negativas, entonces, la nube de puntos tendrá un aspecto bien *amorfo*, bien en torno a una línea no recta, según se observa en las dos figuras siguientes. En estos casos, la covarianza tomará un valor próximo a cero.

Téngase en cuenta que, aunque estos dos casos reflejan *incorrelación*, esto es, ausencia de relación *lineal*, en el primero, la nube de puntos sugiere una situación de independencia entre las variables —que, según hemos visto en el problema anterior, implica incorrelación—, y, en el segundo, en cambio, la existencia de una dependencia funcional *no lineal* entre ellas.



2.21 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, cuya covarianza es S, obténgase la covarianza de la distribución de frecuencias $(a \cdot x_i + b, c \cdot y_j + d; f_{ij})$, S', siendo a y b números reales positivos. En particular, calcúlese la covarianza de la distribución transformada por un cambio de origen y de escala en cada una de las dos variables.

SOLUCIÓN

Si $\bar{x}$ e $\bar{y}$ son, respectivamente, las medias de las distribuciones marginales $(x_i; f_{i\cdot})$ e $(y_j; f_{\cdot j})$, entonces, según vimos en el capítulo anterior, las medias de $(a \cdot x_i + b; f_{i\cdot})$ y $(c \cdot y_j + d; f_{\cdot j})$, dis-

tribuciones marginales correspondientes a la distribución bidimensional transformada son iguales a

$$a \cdot \bar{x} + b$$

y

$$c \cdot \bar{y} + d.$$

Por tanto, sustituyendo en la expresión general, la covarianza de la nueva distribución es

$$S' = \sum_{i=1}^{h} \sum_{j=1}^{k} [(a \cdot x_i + b) - (a \cdot \bar{x} + b)] \cdot [(c \cdot y_j + d) - (c \cdot \bar{y} + d)] f_{ij}.$$

Operando en la expresión anterior, se tiene que

$$S' = \sum_{i=1}^{h} \sum_{j=1}^{k} a(x_i - \bar{x})\, c\, (y_j - \bar{y})\, f_{ij} = a \cdot c \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \bar{x}) \cdot (y_j - \bar{y})\, f_{ij} = a \cdot c \cdot S.$$

En particular, si $a = 1/e_1$, $b = -o_1/e_1$, $c = 1/e_2$ y $d = -o_2/e_2$, la covarianza de la distribución obtenida tras realizar un cambio de origen y de escala es

$$S' = \frac{1}{e_1} \cdot \frac{1}{e_2} \cdot S = \frac{S}{e_1 \cdot e_2},$$

de lo cual se deduce que la covarianza solamente se ve afectada por cambios de escala.

2.22 El Departamento de Marketing de un grupo financiero ha realizado un estudio sobre la influencia de la renta en las decisiones de inversión de sus clientes. Para ello eligió una muestra de 20 clientes, cuya renta anual, junto con las cantidades invertidas en un cierto año, en miles de euros, aparecen recogidas en la siguiente tabla:

Inversión / Renta	0-4	4-8	8-12
6-14	4	2	0
14-26	2	2	3
26-34	0	1	6

a) Hallénse las medias y las varianzas de las variables consideradas.

b) ¿Cuál es la covarianza entre la inversión y la renta?

c) ¿Cuál sería el valor de la covarianza si cada cliente aumentara su inversión en mil euros? ¿Qué valor tendría la covarianza si la renta de cada cliente se incrementara en un 6 por ciento?

SOLUCIÓN

Cuando se ha de realizar el cálculo de momentos unidimensionales y bidimensionales a partir de tablas de correlación, resulta cómodo utilizar un diagrama de apoyo como el que describimos a continuación.

X \ Y	2	6	10	$n_{i\cdot}$	$x_i \cdot n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$	$\sum_{j=1}^{k} y_j \cdot n_{ij}$	$x_i \sum_{j=1}^{k} y_j \cdot n_{ij}$
10	4	2	0	6	60	600	20	200
20	2	2	3	7	140	2 800	46	920
30	0	1	6	7	210	6 300	66	1 980
$n_{\cdot j}$	6	5	9	20	410	9 700	132	3 100
$y_j \cdot n_{\cdot j}$	12	30	90	132				
$y_j^2 \cdot n_{\cdot j}$	24	180	900	1 104				
$\sum_{i=1}^{h} x_i \cdot n_{ij}$	80	90	240	410				
$y_j \sum_{i=1}^{h} x_i \cdot n_{ij}$	160	540	2 400	3 100				

Fijémonos, en primer lugar, en las columnas que aparecen a la derecha de las que inicialmente conforman la tabla de correlación. Así, los elementos de la quinta columna se corresponden, como ya sabemos, con las frecuencias marginales de la variable X, siendo $n_{i\cdot}$ la frecuencia absoluta genérica; en la sexta columna aparecen los productos de cada valor de la variable X, junto con su frecuencia, esto es, $x_i \cdot n_{i\cdot}$, con lo cual, la suma de los elementos de esa columna es

$$\sum_{i=1}^{h} x_i \cdot n_{i\cdot} = 410;$$

la sexta columna se obtiene multiplicando el cuadrado de cada valor de la variable X, x_i^2, por su frecuencia, $n_{i\cdot}$, con lo que la suma de las cantidades de esa columna es

$$\sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = 9\,700;$$

para obtener los elementos de la penúltima columna, hay que fijar cada valor de X, esto es, cada subíndice i, y hallar $\sum_{j=1}^{h} y_j \cdot n_{ij}$, con lo cual, por ejemplo, el primer elemento de dicha columna es igual a $y_1 \cdot n_{11} + y_2 \cdot n_{12} + y_3 \cdot n_{13} = 2 \cdot 4 + 6 \cdot 2 + 10 \cdot 0 = 20$, siendo la suma de los elementos de la columna:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot n_{ij} = 132;$$

finalmente, la última columna se halla de modo sencillo, multiplicando cada elemento de la columna anterior, $\sum_{j=1}^{k} y_j \cdot n_{ij}$, por el correspondiente valor de la variable X, x_i, por lo que la suma de las cantidades de esta columna es

$$\sum_{i=1}^{h} x_i \sum_{j=1}^{k} y_j \cdot n_{ij} = 3\ 100.$$

Observemos el diagrama de apoyo *hacia abajo*, fijándonos en las filas que se han añadido a partir de la tabla de contingencia inicial. Podemos ver que la quinta fila se corresponde, como es sabido, con las frecuencias marginales de la variable Y, $n_{\cdot j}$; los elementos de la sexta fila se calculan multiplicando cada valor de la variable Y, y_j, por su respectiva frecuencia, $n_{\cdot j}$, obteniéndose como suma de los elementos de esta fila:

$$\sum_{j=1}^{k} y_j \cdot n_{\cdot j} = 132;$$

la siguiente fila, la séptima, se obtiene mediante el producto de cada valor al cuadrado de la variable Y, y_j^2, por su frecuencia, $n_{\cdot j}$, con lo que la suma de esta fila es

$$\sum_{j=1}^{k} y_j^2 \cdot n_{\cdot j} = 1\ 104;$$

para calcular cada cantidad de la octava fila se fija cada valor de la variable Y, esto es, el subíndice j, y se halla $\sum_{i=1}^{h} x_i \cdot n_{ij}$, siendo, por ejemplo, el primer elemento de esta fila $x_1 \cdot n_{11} + x_2 \cdot n_{21} + x_3 \cdot n_{31} = 10 \cdot 4 + 20 \cdot 2 + 30 \cdot 0 = 80$ y la suma de sus elementos igual a

$$\sum_{j=1}^{k} \sum_{i=1}^{h} x_i \cdot n_{ij} = 410;$$

por último, la novena fila del diagrama anterior se obtiene como producto entre cada uno de los elementos de la octava fila, $\sum_{i=1}^{h} x_i \cdot n_{ij}$, por el respectivo valor de la variable Y, y_j, y la suma de sus elementos es

$$\sum_{j=1}^{k} y_j \sum_{i=1}^{h} x_i \cdot n_{ij} = 3\,100.$$

Si miramos las cantidades resultantes de sumar elementos de filas y columnas, observamos que hay una serie de coincidencias, que no son, en absoluto, fruto del azar. En efecto, $\sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot n_{ij}$, suma de los elementos de la octava columna, coincide con $\sum_{j=1}^{k} y_j \cdot n_{\cdot j}$, suma de los elementos de la sexta fila, pues

$$\sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot n_{ij} = \sum_{j=1}^{k} y_j \sum_{i=1}^{h} n_{ij} = \sum_{j=1}^{k} y_j \cdot n_{\cdot j}.$$

Siguiendo un proceso análogo, puede analizar el lector la razón de la coincidencia entre $\sum_{j=1}^{k} \sum_{i=1}^{h} x_i \cdot n_{ij}$, suma de los elementos de la penúltima fila, y $\sum_{i=1}^{h} x_i \cdot n_{i\cdot}$, resultado de sumar los elementos de la sexta columna.

Comprobamos también, de modo inmediato, que se llega a idéntico resultado sumando los elementos de la última fila y de la última columna del diagrama anterior:

$$\sum_{i=1}^{h} x_i \sum_{j=1}^{k} y_j \cdot n_{ij} = \sum_{j=1}^{k} y_j \sum_{i=1}^{h} x_i \cdot n_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij}.$$

Aunque en la introducción de este problema hemos obtenido todos los elementos del diagrama, es evidente que, dadas las coincidencias, y siempre en función de los momentos que se requieran en cada caso, bastará con calcular aquellas filas y columnas que se necesiten.

a) Para hallar la inversión media por cliente, media de la variable Y, y la renta media por cliente, media de la variable X, nos apoyamos en el diagrama, seleccionando las sumas convenientes. Así, se tiene que

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_{i\cdot} = \frac{410}{20} = 20{,}5 \text{ miles de euros},$$

para cuyo cálculo hemos utilizado las suma de los elementos de la sexta columna, habiendo podido llegar al mismo resultado —dada la coincidencia— con la suma de los elementos de la penúltima fila.

Análogamente, con la suma de los elementos de la sexta fila, o bien con los correspondientes a la penúltima columna, se obtiene la media de la variable *Y*:

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{k} y_j \cdot n_{\cdot j} = \frac{132}{20} = 6{,}6 \text{ miles de euros.}$$

En cuanto a la varianza de la variable *X*, tomando la suma de los elementos de la séptima columna, se tiene que

$$a_{2,0} = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = \frac{9\,700}{20} = 485,$$

con lo cual,

$$S_X^2 = a_{2,0} - \overline{x}^2 = 485 - 20{,}5^2 = 64{,}75.$$

De modo semejante resulta la varianza de la variable *Y*, utilizando, en este caso, la suma de los elementos de la séptima fila:

$$a_{0,2} = \frac{1}{N}\sum_{j=1}^{k} y_i^2 \cdot n_{\cdot j} = \frac{1\,104}{20} = 55{,}2,$$

siendo, por tanto, la varianza:

$$S_Y^2 = a_{0,2} - \overline{y}^2 = 55{,}2 - 6{,}6^2 = 11{,}64.$$

b) La covarianza entre renta e inversión se calcula con la suma de los elementos de la última fila o bien de la última columna. En efecto, el momento de orden (1, 1) con respecto al origen es

$$a_{1,1} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = \frac{3\,100}{20} = 155,$$

por lo que

$$S = a_{1,1} - \overline{x} \cdot \overline{y} = 155 - 20{,}5 \cdot 6{,}6 = 19{,}7.$$

c) Un aumento en la inversión de mil euros supone una transformación lineal en la variable *Y*: de la distribución bidimensional $(x_i, y_j; f_{ij})$, a la distribución $(x_i, y_j + 1; f_{ij})$.

Ahora bien, según hemos visto en **2.21**, este tipo de transformación no afecta a la covarianza, por lo cual, la covarianza de la nueva distribución sigue siendo 19,7.

La segunda transformación propuesta supone un cambio en la variable X, de modo que, de la distribución bidimensional inicial $(x_i, y_j; f_{ij})$, pasamos a la distribución $(1{,}06 \cdot x_i, y_j; f_{ij})$. En esta situación, según se comprobó en el citado problema, la nueva covarianza es $1{,}06 \cdot S = 1{,}06 \cdot 19{,}7 = 20{,}88$.

2.23 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, obténgase la expresión de la recta de regresión lineal mínimo-cuadrática de Y respecto de X.

SOLUCIÓN

Mediante el criterio de los mínimos cuadrados se obtienen los valores de los parámetros a y b de la recta $\tilde{y}_i = a + b \cdot x_i$ para que las diferencias al cuadrado entre los valores observados de la variable que queremos explicar —en este caso la variable Y—, y_j, y los valores teóricos de dicha variable dados por la recta anterior sean lo más pequeñas posible. Se trata, por tanto, de hacer mínima la función:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} (y_j - \tilde{y}_i)^2 f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} [y_j - (a + b \cdot x_i)]^2 f_{ij}.$$

En consecuencia, derivando respecto de a y de b la función anterior resulta[2]:

$$\frac{d}{da}\left[\sum_{i=1}^{h} \sum_{j=1}^{k} [y_j - (a + b \cdot x_i)]^2 \ f_{ij}\right] = -2 \sum_{i=1}^{h} \sum_{j=1}^{k} [y_j - (a + b \cdot x_i)] \ f_{ij}$$

$$\frac{d}{db}\left[\sum_{i=1}^{h} \sum_{j=1}^{k} [y_j - (a + b \cdot x_i)]^2 \ f_{ij}\right] = -2 \sum_{i=1}^{h} \sum_{j=1}^{k} [y_j - (a + b \cdot x_i)] \ x_i \cdot f_{ij}.$$

Igualando a cero y desarrollando estas expresiones se tiene:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot f_{ij} - a \sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} - b \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot f_{ij} = 0$$

$$\sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot x_i \cdot f_{ij} - a \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot f_{ij} - b \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^2 \cdot f_{ij} = 0.$$

El sistema anterior, que recibe el nombre de *sistema de ecuaciones normales*, puede expresarse en función de los momentos respecto al origen unidimensionales y bidimensionales:

$$a_{0,1} - a - b \cdot a_{1,0} = 0$$
$$a_{1,1} - a \cdot a_{1,0} - b \cdot a_{2,0} = 0.$$

[2] Como puede comprobar el lector, desarrollando los sumatorios, la derivada de un sumatorio es igual al sumatorio de las derivadas, según se deduce de modo inmediato de las reglas de derivación.

Multiplicando la primera ecuación por $-a_{1,0}$ y sumando ambas ecuaciones, se tienen los valores:

$$b = \frac{a_{1,1} - a_{1,0} \cdot a_{0,1}}{a_{2,0} - a^2_{1,0}} = \frac{m_{1,1}}{m_{2,0}}$$

y

$$a = a_{0,1} - \frac{m_{1,1}}{m_{2,0}} \cdot a_{1,0},$$

o, lo que es lo mismo,

$$b = \frac{S}{S_X^2}$$

y

$$a = \overline{y} - \frac{S}{S_X^2} \cdot \overline{x}.$$

En realidad, para que el ejercicio estuviera completo debería comprobarse la condición de mínimo, aunque por tratarse éste de un libro de estadística y no de matemáticas, obviaremos esta comprobación.

En definitiva, la recta de regresión mínimo-cuadrática de Y sobre X, resultante de sustituir los valores a y b calculados, es

$$\tilde{y}_i = \overline{y} - \frac{S}{S_X^2} \cdot \overline{x} + \frac{S}{S_X^2} \cdot x_i,$$

o, equivalentemente, la que es su expresión más habitual, prescindiendo de los subíndices,

$$y - \overline{y} = \frac{S}{S_X^2} (x - \overline{x}).$$

Invitamos al lector a que compruebe, utilizando el mismo procedimiento, que la recta de regresión mínimo-cuadrática de X sobre Y responde a la expresión:

$$x - \overline{x} = \frac{S}{S_Y^2} (y - \overline{y}).$$

2.24 A partir de la regresión lineal de Y, ahorro anual, sobre X, renta mensual de un grupo de familias (ambas variables en miles de euros), se ha estimado que el ahorro correspondiente a una renta de 3 mil euros es de 0,4 miles de euros, mientras que, si la renta es de 2,5 miles de euros, el ahorro es de 0,3 miles de euros. Con estos datos, hállese la ecuación de la recta de regresión de Y sobre X.

SOLUCIÓN

El enunciado indica que la recta de regresión de Y sobre X, $y = a + b \cdot x$, pasa por los puntos (3; 0,4) y (2,5; 0,3). Sustituyendo en la expresión de dicha recta de regresión los valores de estos pares de puntos, se tiene el sistema de ecuaciones:

$$0{,}4 = a + b \cdot 3$$
$$0{,}3 = a + b \cdot 2{,}5.$$

Restando ambas ecuaciones,

$$0{,}4 - 0{,}3 = b\,(3 - 2{,}5),$$

y despejando, se halla el valor

$$b = 0{,}2,$$

que, sustituido en cualquiera de ellas, conduce al valor

$$a = -0{,}2.$$

En definitiva, la recta de regresión de Y sobre X, esto es, la mejor explicación lineal del ahorro a partir de la renta es

$$y = -0{,}2 + 0{,}2 \cdot x.$$

El valor $b = 0{,}2$, pendiente de la recta de regresión, tiene una clara interpretación: un incremento de una unidad, es decir, de mil euros, en la renta de una familia, supondría un aumento del 20 por ciento —$0{,}2 \times 100$— en el ahorro anual, es decir, de 200 euros.

2.25 Obténgase la media y la varianza de los residuos en la regresión lineal de Y sobre X.

SOLUCIÓN

Los residuos de la regresión lineal de Y sobre X son, por definición, la diferencia entre los valores observados de la variable Y y los valores teóricos estimados mediante la recta de regresión, es decir,

$$e_{ij} = y_j - \tilde{y}_i,$$

donde

$$\tilde{y}_i = \overline{y} + \frac{S}{S_X^2}(x_i - \overline{x})$$

es el valor teórico de la variable Y correspondiente a un valor x_i de la variable X.
La media de los residuos será, entonces, sin más que sustituir,

$$\overline{e} = \sum_{i=1}^{h}\sum_{j=1}^{k} e_{ij} \cdot f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - \tilde{y}_i) f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} \left[y_j - \left(\overline{y} + \frac{S}{S_X^2}(x_i - \overline{x}) \right) \right] f_{ij}.$$

Reagrupando términos y colocando fuera de los sumatorios los valores constantes, la media de los residuos es

$$\overline{e} = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - \overline{y}_i) f_{ij} - \frac{S}{S_X^2} \sum_{i=1}^{h}\sum_{j=1}^{k} (x_i - \overline{x}) f_{ij}.$$

Ahora bien, el primer sumando de la expresión anterior es

$$\sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - \overline{y}) f_{ij} = \sum_{j=1}^{k} (y_j - \overline{y}) \sum_{i=1}^{h} f_{ij} = \sum_{j=1}^{k} (y_j - \overline{y}) f_{\cdot j} = 0,$$

pues, según vimos en el capítulo anterior, se trata de la media aritmética de las desviaciones de los valores de la distribución de la variable Y con respecto a su media.

Y, análogamente,

$$\frac{S}{S_X^2} \sum_{i=1}^{h}\sum_{j=1}^{k} (x_i - \overline{x}) f_{ij} = \frac{S}{S_X^2} \cdot 0 = 0,$$

con lo cual, la media aritmética de los residuos resulta:

$$\overline{e} = 0.$$

En consecuencia, la varianza de los residuos, o varianza residual, es

$$S_e^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} (e_{ij} - \overline{e})^2 f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} e_{ij}^2 \cdot f_{ij}.$$

Sustituyendo e_{ij} por su valor y posteriormente $\tilde{y}_i$ por el suyo, se obtiene:

$$S_e^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - \tilde{y}_i)^2 f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} \left[y_j - \left(\overline{y} + \frac{S}{S_X^2}(x_i - \overline{x}) \right) \right]^2 f_{ij} =$$

$$= \sum_{i=1}^{h}\sum_{j=} \left[(y_j - \overline{y}) - \frac{S}{S_X^2}(x_i - \overline{x}) \right]^2 f_{ij}.$$

Desarrollando el binomio y descomponiendo el doble sumatorio anterior en tres sumandos, se tiene que la varianza residual es

$$S_e^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}(y_j-\overline{y})^2 f_{ij} + \left(\frac{S}{S_X^2}\right)^2 \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i-\overline{x})^2 f_{ij} - 2\cdot\frac{S}{S_X^2}\sum_{i=1}^{h}\sum_{j=1}^{k}(y_j-\overline{y})\cdot(x_i-\overline{x})\, f_{ij}.$$

Si tenemos en cuenta que

$$\sum_{i=1}^{h}\sum_{j=1}^{k}(y_j-\overline{y})^2 f_{ij} = \sum_{j=1}^{k}(y_j-\overline{y})^2\sum_{i=1}^{h} f_{ij} = \sum_{j=1}^{k}(y_j-\overline{y})^2 f_{\cdot j} = S_Y^2,$$

que, además,

$$\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i-\overline{x})^2 f_{ij} = \sum_{i=1}^{h}(x_i-\overline{x})^2\sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h}(x_i-\overline{x})^2 f_{i\cdot} = S_X^2,$$

y que, por último,

$$\sum_{i=1}^{h}\sum_{j=1}^{k}(y_j-\overline{y})\cdot(x_i-\overline{x})\, f_{ij} = S,$$

la expresión de la varianza residual resulta ser:

$$S_e^2 = S_Y^2 + \frac{S^2}{(S_X^2)^2}\cdot S_X^2 - 2\cdot\frac{S}{S_X^2}\cdot S = S_Y^2 + \frac{S^2}{S_X^2} - 2\cdot\frac{S^2}{S_X^2},$$

esto es,

$$S_e^2 = S_Y^2 - \frac{S^2}{S_X^2}.$$

Es importante observar que, aunque los residuos y sus frecuencias dependen de i y de j, por definición, lo cual nos obliga a trabajar con sumatorios dobles, en realidad, hemos calculado una media y una varianza de una distribución de frecuencias unidimensional.

Se puede comprobar, siguiendo un desarrollo análogo al de este problema, que la varianza residual en la regresión lineal de X sobre Y es

$$S_{e'}^2 = S_X^2 - \frac{S^2}{S_Y^2},$$

donde $S_{e'}^2$ es la varianza residual de dicha regresión.

2.26 Obténgase la media y la varianza de los valores teóricos en la regresión lineal de Y sobre X.

SOLUCIÓN

Ya que entre los residuos, los valores observados y los valores teóricos de la variable Y en la regresión lineal de Y sobre X, existe la siguiente relación:

$$e_{ij} = y_j - \tilde{y}_i,$$

entonces, despejando,

$$\tilde{y}_i = y_j - e_{ij},$$

con lo cual, la media de los valores teóricos es

$$\bar{\tilde{y}} = \sum_{i=1}^{h} \sum_{j=1}^{k} (y_j - e_{ij}) f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} y_j \cdot f_{ij} - \sum_{i=1}^{h} \sum_{j=1}^{k} e_{ij} \cdot f_{ij} = \bar{y} - \bar{e},$$

simplemente con aplicar resultados ya comentados en el problema anterior.

Por último, y puesto que la media de los residuos es cero, se tiene que

$$\bar{\tilde{y}} = \bar{y} - \bar{e} = \bar{y},$$

esto es, *la media de los valores teóricos coincide con la media de los valores observados*, es decir, con la media de la variable Y.

Por tanto, a la hora de calcular la varianza de los valores teóricos tendremos que

$$S_{\tilde{Y}}^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} (\tilde{y}_i - \bar{\tilde{y}})^2 f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} (\tilde{y}_i - \bar{y})^2 f_{ij}.$$

Como los valores teóricos de la regresión de Y sobre X responden a la expresión genérica:

$$\tilde{y}_i = \bar{y} + \frac{S}{S_X^2} (x_i - \bar{x}),$$

entonces,

$$\tilde{y}_i - \bar{y} = \bar{y} + \frac{S}{S_X^2} (x_i - \bar{x}) - \bar{y} = \frac{S}{S_X^2} (x_i - \bar{x}),$$

por lo que la varianza de esta variable, tras realizar oportunas operaciones, es

$$S^2_{\tilde{Y}} = \sum_{i=1}^{h}\sum_{j=1}^{k}\left[\frac{S}{S^2_X}(x_i - \overline{x})\right]^2 f_{ij} = \left(\frac{S}{S^2_X}\right)^2 \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})^2 f_{ij} = \frac{S^2}{(S^2_X)^2}\cdot S^2_X,$$

esto es,

$$S^2_{\tilde{Y}} = \frac{S^2}{S^2_X}.$$

Podríamos haber prescindido de trabajar con frecuencias conjuntas, ya que la varianza de los valores teóricos,

$$S^2_{\tilde{Y}} = \sum_{i=1}^{h}\sum_{j=1}^{k}(\tilde{y}_i - \overline{\tilde{y}})^2 f_{ij} = \sum_{i=1}^{h}(\tilde{y}_i - \overline{\tilde{y}})^2 \sum_{j=1}^{k} f_{ij} = \sum_{i=1}^{h}(\tilde{y}_i - \overline{\tilde{y}})^2 f_{i\cdot},$$

es, en realidad, la varianza de la distribución $(\tilde{y}_i; f_{i\cdot})$, distribución de frecuencias unidimensional de la variable $\tilde{Y}$.

Otra vía alternativa para resolver este ejercicio pasa por considerar que

$$\tilde{y}_i = \overline{y} + \frac{S}{S^2_X}(x_i - \overline{x}),$$

o, lo que es lo mismo,

$$\tilde{y}_i = \frac{S}{S^2_X}\cdot x_i + \overline{y} - \frac{S}{S^2_X}\cdot \overline{x},$$

por lo cual, cuando hallamos los valores teóricos de la recta de regresión de Y sobre X, estamos obteniendo, realmente, los valores de la distribución transformada

$$\left(\frac{S}{S^2_X}\cdot x_i + \overline{y} - \frac{S}{S^2_X}\cdot \overline{x}; f_{i\cdot}\right),$$

a partir de la distribución $(x_i; f_{i\cdot})$, siendo, en este caso[3], $a = S/S^2_X$ y $b = \overline{y} - (S/S^2_X)\,\overline{x}$.

[3] No hay que confundir con los valores a y b de la recta de regresión, pues, en esta ocasión, estamos siguiendo las notación del capítulo 1, correspondiente a transformaciones lineales.

En consecuencia, aplicando los resultados conocidos del capítulo 1 sobre el cálculo de la media y de la varianza de una distribución transformada, tendremos que, por un lado, la media de los valores teóricos es

$$\bar{\tilde{y}} = \frac{S}{S_X^2} \cdot \bar{x} + \bar{y} - \frac{S}{S_X^2} \cdot \bar{x} = \bar{y}$$

y, por otro lado, la varianza es

$$S_{\tilde{Y}}^2 = \left(\frac{S}{S_X^2} \right)^2 S_X^2 = \frac{S^2}{S_X^2}.$$

Puede el lector comprobar, razonando de modo análogo, que la varianza de $\tilde{X}$, varianza de los valores teóricos en la regresión lineal de X sobre Y, es

$$S_{\tilde{X}}^2 = \frac{S^2}{S_Y^2},$$

siendo la media de $\tilde{X}$ igual a $\bar{x}$.

2.27 Demuéstrese que el coeficiente de determinación lineal responde a la expresión:

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2}.$$

SOLUCIÓN

Sustituyendo en la definición de coeficiente de determinación lineal en la regresión de Y sobre X,

$$r^2 = \frac{S_{\tilde{Y}}^2}{S_Y^2},$$

el valor $S_{\tilde{Y}}^2 = S^2/S_X^2$, obtenido en **2.26**, resulta, de modo inmediato, que

$$r^2 = \frac{S^2/S_X^2}{S_Y^2} = \frac{S^2}{S_X^2 \cdot S_Y^2},$$

expresión habitual de este coeficiente.

Si el punto de partida para esta demostración hubiera sido la regresión lineal de X sobre Y, habríamos llegado a idéntico resultado:

$$r^2 = \frac{S_{\tilde{X}}^2}{S_X^2} = \frac{S^2}{S_X^2 \cdot S_Y^2}.$$

En consecuencia el mismo (único) coeficiente de determinación lineal, r^2, sirve para interpretar la bondad del ajuste de Y sobre X y de X sobre Y. Esta conclusión es coherente con el hecho de que con este coeficiente estamos midiendo el *grado de relación lineal* entre las variables X e Y.

2.28 En la regresión lineal de Y sobre X, demuéstrese la siguiente relación denominada *descomposición de la varianza*:

$$S_Y^2 = S_{\tilde{Y}}^2 + S_e^2.$$

SOLUCIÓN

Teniendo en cuenta los resultados de los problemas anteriores, la demostración es inmediata, ya que, por un lado,

$$S_e^2 = S_Y^2 - \frac{S^2}{S_X^2}$$

y, por otro lado,

$$S_{\tilde{Y}}^2 = \frac{S^2}{S_X^2}.$$

Con lo cual, resulta evidente, sumando ambas ecuaciones miembro a miembro, que la varianza de la variable Y se descompone en la varianza de $\tilde{Y}$ y en la varianza de e, según queríamos demostrar.

Si consideramos la regresión lineal de X sobre Y, con un razonamiento análogo demostraríamos que

$$S_X^2 = S_{\tilde{X}}^2 + S_{\tilde{e}}^2,$$

donde S_X^2 es la varianza de X, $S_{\tilde{X}}^2$ es la varianza de los valores teóricos y $S_{\tilde{e}}^2$ es la varianza residual.

2.29 A partir de la relación demostrada en el problema anterior, justifíquese la expresión del coeficiente de determinación lineal de la regresión lineal de Y sobre X y coméntese, utilizando dicho coeficiente, las diferentes situaciones que pueden plantearse en el estudio de la bondad del ajuste.

SOLUCIÓN

La relación

$$S_Y^2 = S_{\tilde{Y}}^2 + S_e^2,$$

correspondiente a la regresión lineal de Y sobre X, indica que toda la *variabilidad* de Y, variable que queremos explicar, queda determinada por la varianza de los valores teóricos, es decir, por la regresión realizada, junto con la varianza de los residuos. Ello significa que, cuanto menor sea la varianza residual, mayor será la varianza de $\tilde{Y}$ en relación con la varianza de Y, o lo que es igual, mayor será la varianza de Y que habremos conseguido explicar con la regresión efectuada. Este razonamiento justifica la definición del coeficiente de determinación lineal como la proporción de varianza de Y explicada por la regresión, es decir, la proporción que la varianza de $\tilde{Y}$ representa sobre la varianza total:

$$r^2 = \frac{S_{\tilde{Y}}^2}{S_Y^2}.$$

La descomposición de la varianza de Y en suma de dos cantidades positivas explica igualmente el hecho de que el numerador de r^2 sea siempre menor que el denominador y el que, por tanto,

$$0 \leq r^2 \leq 1.$$

En cuanto a la interpretación de los diferentes valores de este coeficiente, consideremos las siguientes situaciones:

- Si $r^2 = 0$, entonces, el numerador de su expresión será igualmente nulo, es decir, $S_{\tilde{Y}}^2 = 0$. Ello quiere decir que, por la relación existen entre S_Y^2, $S_{\tilde{Y}}^2$ y S_e^2, necesariamente se cumple que

  $$S_Y^2 = S_e^2,$$

  por lo que, en este caso, resulta nula la parte de la variabilidad de Y que ha quedado explicada por la regresión: el ajuste lineal es pésimo, no existiendo relación lineal entre las variables X e Y.

  Obsérvese, además que, al ser cero la varianza de los valores teóricos, $S_{\tilde{Y}}^2$, no hay dispersión, con lo cual, la variable $\tilde{Y}$ es constante, coincidiendo con su media:

  $$\tilde{y}_i = \bar{\tilde{y}} = \bar{y},$$

  para todo $i = 1, ..., h$, siendo la recta de regresión de Y sobre X:

  $$y = \bar{y}.$$

Este razonamiento es coherente con el hecho de que, si $r^2 = 0$, y puesto que $r^2 = S^2/S_X^2 \cdot S_Y^2$, entonces, necesariamente, $S = 0$, con lo cual, la expresión de la recta de regresión de Y sobre X es $y = \overline{y}$.

Además, al ser la covarianza, S, igual a cero, la recta de regresión de X sobre Y es $x = \overline{x}$, conclusión a la que llegaríamos igualmente, partiendo de la descomposición de la varianza en la regresión de X sobre Y, $S_X^2 = S_{\tilde{X}}^2 + S_{e'}^2$, y siguiendo un razonamiento paralelo al efectuado en este punto con la regresión de Y sobre X.

- Si $r^2 = 1$, numerador y denominador del coeficiente de determinación lineal coinciden, $S_Y^2 = S_{\tilde{Y}}^2$, por lo que el ajuste lineal es perfecto, al conseguir explicar toda la varianza de Y mediante la regresión realizada.

  Nótese, además, que por la relación existente entre las tres varianzas, en este caso resulta ser nula la varianza residual, lo que supone que la variable e es constante y coincide con su valor medio:

  $$e_{ij} = \overline{e} = 0,$$

  para cualesquiera i y j.

  Pero como

  $$e_{ij} = y_j - \tilde{y}_i,$$

  se deduce, que, para cada valor x_i de la variable X, existe un valor y_j de la variable Y tal que $y_j = \tilde{y}_i$, situándose los puntos del diagrama de dispersión perfectamente alineados y existiendo, por tanto, dependencia lineal perfecta entre X e Y.

- En la medida en que r^2 se acerque a cero, peor será el ajuste, esto es, menor el grado de dependencia lineal entre las variables, y viceversa, cuanto más se aproxime a 1, mejor será la regresión y, por tanto, mayor el grado de dependencia lineal entre X e Y.

A partir de la descomposición de la varianza, el coeficiente de determinación lineal admite la expresión:

$$r^2 = \frac{S_{\tilde{Y}}^2}{S_Y^2} = \frac{S_Y^2 - S_e^2}{S_Y^2} = 1 - \frac{S_e^2}{S_Y^2},$$

en la regresión de Y sobre X, y

$$r^2 = \frac{S_{\tilde{X}}^2}{S_X^2} = \frac{S_X^2 - S_{e'}^2}{S_X^2} = 1 - \frac{S_{e'}^2}{S_X^2},$$

en la regresión de X sobre Y.

2.30 Demuéstrese que, si existe dependencia lineal perfecta entre las variables X e Y, esto es, si

$$Y = a + b \cdot X,$$

donde a y b son números reales, $b \neq 0$, entonces,

$$|S| = S_X \cdot S_Y.$$

SOLUCIÓN

Por las propiedades de la varianza de una variable vistas en el capítulo 1, si S_X^2 es la varianza de la variable X, entonces, la varianza de la variable Y es

$$S_Y^2 = b^2 \cdot S_X^2,$$

y, en consecuencia, su desviación típica es

$$S_Y = |b| \cdot S_X.$$

Para calcular la covarianza entre X e Y, S, hay que considerar que, para cada valor de la variable X, x_i, existe un valor de la variable Y, $a + b \cdot x_i$, con lo cual, puede escribirse un único sumatorio en la expresión de S; además, por las propiedades de la media aritmética, se cumple que $\overline{y} = a + b \cdot \overline{x}$. Teniendo en cuenta estos comentarios, la covarianza entre las variables X e Y es

$$S = \sum_{i=1}^{h} (x_i - \overline{x}) \cdot [(a + b \cdot x_i) - (a + b \cdot \overline{x})] f_{i\cdot} = b \sum_{i=1}^{h} (x_i - \overline{x}) f_{i\cdot} = b \cdot S_X^2.$$

Por tanto, tomando módulos en la expresión anterior, se tiene, por un lado,

$$|S| = |b| \cdot S_X^2,$$

y, por otro lado, el producto de las desviaciones típicas es

$$S_X \cdot S_Y = S_X \cdot |b| \cdot S_X = |b| \cdot S_X^2.$$

En definitiva, comparando ambas expresiones:

$$|S| = S_X \cdot S_Y,$$

según queríamos demostrar.

Se concluye, por tanto, que, si la relación entre las variables es creciente, esto es, si $b > 0$, entonces,

$$S = b \cdot S_X^2$$

es una cantidad positiva, con lo cual, $|S| = S$ y

$$S = S_X \cdot S_Y,$$

siendo, en tal caso, el coeficiente de correlación lineal,

$$r = \frac{S}{S_X \cdot S_Y},$$

igual a 1.

Por el contrario, si la relación entre X e Y es decreciente, es decir, si $b < 0$, entonces,

$$S = b \cdot S_X^2$$

es menor que cero, siendo, en ese caso, $|S| = -S$ y verificándose que

$$S = -S_X \cdot S_Y,$$

con lo cual, el coeficiente de correlación lineal, r, toma el valor -1.

2.31 En el departamento comercial de una empresa, con restaurantes de comida rápida repartidos por la geografía de una gran ciudad, se sospecha el número de personas que consumen diariamente el «menú de la casa» depende del precio de éste, puesto que al variar los precios en 10 establecimientos se han obtenido los siguientes resultados en un cierto día:

Establecimientos	1	2	3	4	5	6	7	8	9	10
Precio menú (en euros)	4,5	4,6	4,7	4,8	4,9	5,0	5,1	5,2	5,3	5,4
N.º comensales	80	79	72	65	70	64	61	50	45	43

a) Obténgase la ecuación lineal que exprese la dependencia estadística intuida.

b) Represéntese la nube de puntos de la distribución y la recta de regresión obtenida.

c) Calcúlese una medida de la bondad del ajuste.

d) Hállese una predicción del número diario de comensales, si el precio del menú fuera de 6 euros.

SOLUCIÓN

a) Denotando por X la variable precio del menú y por Y la variable número diario de comensales, la ecuación que expresa la relación lineal entre ambas variables, intuida por el departamento comercial de la empresa, es la recta de regresión:

$$y - \overline{y} = \frac{S}{S_X^2}(x - \overline{x}).$$

Al ser las frecuencias unitarias, es decir, al no repetirse los pares de observaciones de las unidades de la población analizada, que en este caso son los establecimientos, la tabla de correlación es:

X \ Y	43	45	50	61	64	65	70	72	79	80
4,5	0	0	0	0	0	0	0	0	0	1
4,6	0	0	0	0	0	0	0	0	1	0
4,7	0	0	0	0	0	0	0	1	0	0
4,8	0	0	0	0	0	1	0	0	0	0
4,9	0	0	0	0	0	0	1	0	0	0
5,0	0	0	0	0	1	0	0	0	0	0
5,1	0	0	0	1	0	0	0	0	0	0
5,2	0	0	1	0	0	0	0	0	0	0
5,3	0	1	0	0	0	0	0	0	0	0
5,4	1	0	0	0	0	0	0	0	0	0

Se trata, por tanto, de una tabla en la que cada fila y columna tiene un uno y sólo un uno y el resto son ceros; en este caso, los valores de la variable bidimensional coinciden con las observaciones y cada dato de la variable X se corresponde con un dato, y sólo uno, de la variable Y. Esta situación permite que las notaciones y los cálculos de los momentos necesarios para la obtención de las rectas de regresión y coeficientes de bondad de ajuste sean más sencillos.

Así, la media, suma de las observaciones entre el total de datos, de cada una de las variables, es

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{10}(4{,}5 + 4{,}6 + 4{,}7 + 4{,}8 + 4{,}9 + 5 + 5{,}1 + 5{,}2 + 5{,}3 + 5{,}4) = 4{,}95 \text{ euros}$$

y

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{10}(80 + 79 + 72 + 65 + 70 + 64 + 61 + 50 + 45 + 43) = 62{,}9 \text{ comensales.}$$

Para obtener el coeficiente de regresión $b_{Y/X} = S/S_X^2$, calculamos numerador y denominador, apoyándonos en los momentos no centrales:

$$a_{1,1} = \frac{1}{N}\sum_{i=1}^{N} x_i \cdot y_i$$

y

$$a_{2,0} = \frac{1}{N}\sum_{i=1}^{N} x_i^2,$$

que, con los datos del problema, son

$$a_{1,1} = \frac{4{,}5\cdot 80 + 4{,}6\cdot 79 + 4{,}7\cdot 72 + 4{,}8\cdot 65 + 4{,}9\cdot 70 + 5\cdot 64 + 5{,}1\cdot 61 + 5{,}2\cdot 50 + 5{,}3\cdot 45 + 5{,}4\cdot 43}{10},$$

es decir,

$$a_{1,1} = 307{,}86,$$

y

$$a_{2,0} = \frac{1}{10}\,(4{,}5^2 + 4{,}6^2 + 4{,}7^2 + 4{,}8^2 + 4{,}9^2 + 5^2 + 5{,}1^2 + 5{,}2^2 + 5{,}3^2 + 5{,}4^2) = 24{,}585.$$

En las expresiones genéricas de todos los momentos calculados los sumatorios toman valores desde 1 hasta N, puesto que, según se ha dicho, el número de valores de ambas variables coincide en este caso con el número de datos.

En definitiva,

$$b_{Y/X} = \frac{S}{S_X^2} = \frac{a_{1,1} - \bar{x} \cdot \bar{y}}{a_{2,0} - \bar{x}^2} = \frac{307{,}86 - 4{,}95 \cdot 62{,}9}{24{,}585 - 4{,}95^2} = \frac{-\,3{,}495}{0{,}0825} = -\,42{,}36.$$

Por tanto, la recta de regresión de Y sobre X es

$$y - 62{,}9 = -\,42{,}36\,(x - 4{,}95)$$

o, lo que es igual,

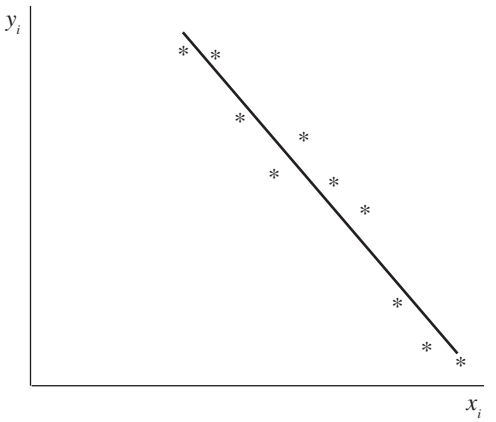
$$y = 272{,}582 - 42{,}36 \cdot x.$$

El coeficiente de regresión obtenido, pendiente de la recta de regresión, $-42{,}36$, indica que un aumento unitario de la variable X, en este caso un incremento de un euro en el precio del menú, produciría una disminución de 42,36 unidades en la variable Y, es decir, un descenso de más de 42 clientes diarios en un establecimiento.

b) En la siguiente tabla figuran los valores de la variable X, junto con los valores teóricos proporcionados por la regresión lineal, es decir, los valores de la variable $\tilde{Y}$:

x_i	$\tilde{y}_i$
4,5	81,962
4,6	77,726
4,7	73,490
4,8	69,254
4,9	65,018
5,0	60,782
5,1	56,546
5,2	52,310
5,3	48,074
5,4	43,838

Se observa que, por ejemplo, el valor $\tilde{y}_7 = 56{,}546$ se obtiene como $272{,}582 - 42{,}36 \cdot 5{,}1$, esto es, sustituyendo el valor $x_7 = 5{,}1$ en la recta de regresión calculada. Los pares de puntos $(x_i, \tilde{y}_i)$, hallados mediante este proceso y que conforman la recta de regresión mínimo cuadrática de Y sobre X, aparecen representados en la siguiente gráfica, junto con la nube de puntos de los pares (x_i, y_i).



Puede apreciarse que existe mucha similitud entre la nube de puntos y la recta de regresión, hecho que se constatará de modo objetivo en el siguiente apartado con la obtención de los coeficientes de determinación y de correlación lineal. Además, tanto la nube de puntos, como la recta de regresión, cuya pendiente tiene signo negativo, muestran la relación decreciente entre el precio del menú y el número de comensales que acuden a un establecimiento.

c) Una medida de la bondad del ajuste es el coeficiente de determinación lineal:

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2}.$$

Puesto que del apartado ***a)*** se tienen los valores de la covarianza, S, y de la varianza de X, S_X^2, únicamente resta calcular el valor de la varianza de la variable Y:

$$S_Y^2 = a_{0,2} - \overline{y}^2,$$

con

$$a_{0,2} = \frac{1}{N} \sum_{i=1}^{N} y_i^2.$$

Sustituyendo los datos del problema, resulta que

$$a_{0,2} = \frac{1}{10}\,(80^2 + 79^2 + 72^2 + 65^2 + 70^2 + 64^2 + 61^2 + 50^2 + 45^2 + 43^2) = 4\,114{,}1,$$

siendo, por tanto, la varianza de la variable Y:

$$S_Y^2 = 4\,114{,}1 - 62{,}9^2 = 157{,}69.$$

De este modo, el coeficiente de determinación lineal es

$$r^2 = \frac{(-3{,}495)^2}{0{,}0825 \cdot 157{,}69} = 0{,}9389$$

y el coeficiente de correlación lineal, raíz cuadrada del coeficiente de determinación lineal, toma el valor

$$r = \frac{S}{S_X \cdot S_Y} = -0{,}968.$$

El signo negativo del coeficiente de correlación lineal, signo de la covarianza, expresa la existencia de una relación decreciente entre las variables X e Y.

Los coeficientes calculados son indicativos de un alto grado de correlación lineal entre las variables, reflejo, igualmente, de un buen ajuste.

d) Para resolver este apartado basta con sustituir el valor de la variable X, en este caso, $x = 6$, en la recta de regresión obtenida:

$$y = 272{,}582 - 42{,}36 \cdot 6,$$

con lo cual, resulta un valor

$$y = 18{,}42,$$

esto es, un número diario de comensales igual a 19.

Dado que, según se ha comprobado en el apartado anterior, existe un alto grado de dependencia lineal entre las variables, tiene sentido la estimación planteada.

2.32 De un estudio elaborado sobre la relación entre la renta per cápita mensual, X, en euros, y el número de vehículos matriculados por cada 100 habitantes, Y, en 12 ciudades de un país, se ha obtenido la siguiente distribución de frecuencias bidimensional:

Ciudades	1	2	3	4	5	6	7	8	9	10	11	12
Renta	900	1 000	1 025	1 050	1 200	1 500	1 700	2 000	2 100	2 150	2 300	2 400
N.º vehículos	16,5	17	17,5	17,25	19	19,5	20,5	22	22,5	22,75	25	26

a) Estímese la relación de dependencia lineal.

b) Represéntese gráficamente la ecuación obtenida en el apartado anterior.

c) Calcúlese la bondad del ajuste.

SOLUCIÓN

a) Las variables del enunciado nos hacen pensar que lo más sensato a la hora de estimar la relación de dependencia lineal entre las variables es explicar Y a partir de X, esto es, obtener la recta de regresión de Y sobre X,

$$y - \overline{y} = \frac{S}{S_X^2}(x - \overline{x}).$$

Para ello, y teniendo en cuenta que, al igual que ocurría en el problema anterior, se trata de una distribución de frecuencias unitaria, construiremos, en esta ocasión, una tabla de apoyo que puede resultar de interés al lector para la resolución de este tipo de problemas.

Como puede verse, además de la primera y segunda columnas de la tabla que contienen, respectivamente, los valores de las variables X e Y, en la tercera columna se incluyen los productos entre dichos valores; en la cuarta y quinta columna están los cuadrados de los valores de cada variable. Cada casilla de la última fila de la tabla es la suma de los elementos de su columna.

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
900	16,50	14 850,0	810 000	272,2500
1 000	17,00	17 000,0	1 000 000	289,0000
1 025	17,50	17 937,5	1 050 625	306,2500
1 050	17,25	18 112,5	1 102 500	297,5625
1 200	19,00	22 800,0	1 440 000	361,0000
1 500	19,50	29 250,0	2 250 000	380,2500
1 700	20,50	34 850,0	2 890 000	420,2500
2 000	22,00	44 000,0	4 000 000	484,0000
2 100	22,50	47 250,0	4 410 000	506,2500
2 150	22,75	48 912,5	4 622 500	517,5625
2 300	25,00	57 500,0	5 290 000	625,0000
2 400	26,00	62 400,0	5 760 000	676,0000
19 325	**245,50**	**414 862,5**	**34 625 625**	**5 135,3750**

Las medias de las variables son

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{19\,325}{12} = 1\,610{,}417 \text{ euros}$$

y

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{245{,}5}{12} = 20{,}458 \text{ vehículos.}$$

Numerador y denominador del coeficiente de regresión, $b_{Y/X} = S/S_X^2$, se calculan mediante los momentos respecto al origen:

$$a_{1,1} = \frac{1}{N}\sum_{i=1}^{N} x_i \cdot y_i = \frac{414\,862{,}5}{12} = 34\,571{,}875$$

y

$$a_{2,0} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 = \frac{34\,625\,625}{12} = 2\,885\,468{,}75.$$

Por tanto,

$$b_{Y/X} = \frac{S}{S_X^2} = \frac{a_{1,1} - \overline{x}\cdot\overline{y}}{a_{2,0} - \overline{x}^2} = \frac{34\,571{,}875 - 1\,610{,}417 \cdot 20{,}458}{2\,885\,468{,}75 - 1\,610{,}417^2} = \frac{1\,625{,}964}{292\,025{,}836} = 0{,}005567,$$

y la recta de regresión de Y sobre X es

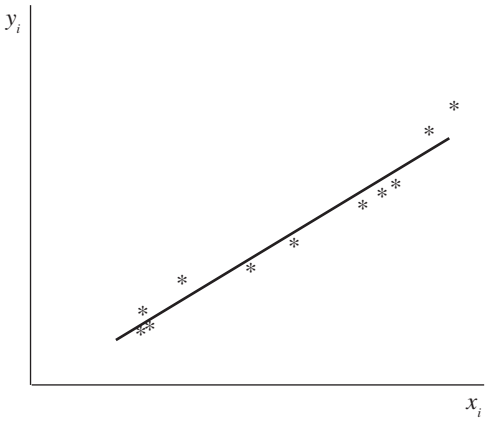
$$y - 20{,}458 = 0{,}005567\ (x - 1610{,}417),$$

es decir,

$$y = 11{,}4928 + 0{,}005567 \cdot x.$$

b) En la siguiente gráfica se recoge, tanto la nube de puntos de los pares de valores de las variables consideradas, (x_i, y_i), como la recta de regresión hallada en el apartado anterior formada por los puntos $(x_i, \tilde{y}_i)$.

Como puede verse, existe una fuerte dependencia lineal *creciente* entre las variables.



c) El coeficiente de determinación lineal,

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2},$$

toma, para los datos del problema, el valor

$$r^2 = \frac{1\,625{,}964^2}{292\,025{,}836 \cdot 9{,}42} = 0{,}96,$$

donde la varianza de Y se ha calculado como

$$S_Y^2 = a_{0,2} - \overline{y}^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \overline{y}^2 = 427{,}948 - 20{,}458^2 = 9{,}42.$$

La raíz cuadrada positiva —pues positiva es la covarianza— del coeficiente de determinación lineal es el coeficiente de correlación lineal:

$$r = \sqrt{0{,}96} = 0{,}979.$$

Aunque el estudio de la bondad de la regresión realizada ya estaría terminado, con un elevado coeficiente de determinación lineal indicativo de un buen ajuste, vamos a ver con este ejemplo algunos de los resultados teóricos estudiados. Para ello, completamos la tabla anterior con las columnas correspondientes a los valores al cuadrado de la variable $\widetilde{Y}$ y a los valores de la variable e, junto con sus cuadrados.

x_i	$\tilde{y}_i$	$\tilde{y}_i^2$	e_i	e_i^2
900	16,5031	272,3523	−0,0031	0,0000
1 000	17,0598	291,0368	−0,0598	0,0036
1 025	17,1990	295,8047	0,3010	0,0906
1 050	17,3382	300,6114	−0,0881	0,0078
1 200	18,1732	330,2652	0,8268	0,6836
1 500	19,8433	393,7566	−0,3433	0,1179
1 700	20,9567	439,1833	−0,4567	0,2086
2 000	22,6268	511,9721	−0,6268	0,3929
2 100	23,1835	537,4747	−0,6835	0,4672
2 150	23,4619	550,4584	−0,7118	0,5067
2 300	24,2969	590,3393	0,7031	0,4943
2 400	24,8536	617,7014	1,1464	1,3142
19 325	**245,5**	**5 130,96**	**0**	**4,2874**

Los valores de la variable e se han obtenido como

$$e_i = y_i - \tilde{y}_i.$$

Con todos estos datos, son varios los resultados que podemos comprobar. Así, por ejemplo, vemos que la media de los valores teóricos,

$$\bar{\tilde{y}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i = \frac{245{,}5}{12} = 20{,}458,$$

coincide, efectivamente, con la media de los valores observados o media de la variable Y, $\bar{y}$.

Comprobamos, también, que la media de los residuos en la regresión lineal es 0:

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{0}{12} = 0.$$

Además, si calculamos la varianza de los valores teóricos,

$$S^2_{\tilde{Y}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i^2 - \bar{\tilde{y}}^2 = \frac{5\ 130{,}96}{12} - 20{,}458^2 = 427{,}58 - 418{,}53 = 9{,}05,$$

y la varianza residual,

$$S_e^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{4{,}2874}{12} = 0{,}36,$$

observamos que se cumple la descomposición de la varianza,

$$S_Y^2 = S_{\hat{Y}}^2 + S_e^2.$$

Por último, el coeficiente de determinación lineal puede hallarse también como

$$r^2 = \frac{S_{\hat{Y}}^2}{S_Y^2} = \frac{9{,}05}{9{,}42} = 0{,}96,$$

coeficiente que nos indica que el 96 por ciento de la varianza total ha sido explicada por la regresión lineal efectuada.

2.32 En la regresión lineal de Y sobre X, demuéstrese la siguiente relación:

$$S_e^2 = S_Y^2\,(1 - r^2).$$

SOLUCIÓN

Hay que tener en cuenta que, según vimos en **2.25**,

$$S_e^2 = S_Y^2 - \frac{S^2}{S_X^2},$$

con lo cual, sacando factor común a S_Y^2, resulta:

$$S_e^2 = S_Y^2\left(1 - \frac{S^2}{S_X^2 \cdot S_Y^2}\right) = S_Y^2\,(1 - r^2),$$

quedando, así, demostrada la igualdad.

Puede intentar el lector plantear y resolver un problema análogo a propósito de la regresión lineal de X sobre Y.

2.34 Pruébese que

$$b_{Y/X} \cdot b_{X/Y} = r^2,$$

donde $b_{Y/X}$ y $b_{X/Y}$ son los coeficientes de regresión de las rectas de regresión.

SOLUCIÓN

Puesto que, por definición,

$$b_{Y/X} = \frac{S}{S_X^2},$$

y

$$b_{X/Y} = \frac{S}{S_Y^2}$$

entonces, de modo inmediato, se obtiene que

$$b_{Y/X} \cdot b_{X/Y} = \frac{S}{S_X^2} \cdot \frac{S}{S_Y^2} = \frac{S^2}{S_X^2 \cdot S_Y^2} = r^2.$$

2.35 En la regresión lineal de Y sobre X, demuéstrese la siguiente expresión de la recta de regresión a partir de coeficiente de correlación lineal, r.

$$y = \bar{y} + r \cdot \frac{S_Y}{S_X}(x - \bar{x}).$$

SOLUCIÓN

Multiplicando numerador y denominador del coeficiente de regresión de la recta de regresión de Y sobre X, S/S_X^2, por S_Y, en la expresión de dicha recta, se tiene:

$$y = \bar{y} + \frac{S}{S_X^2} \cdot \frac{S_Y}{S_Y}(x - \bar{x}) = \bar{y} + \frac{S}{S_X \cdot S_Y} \cdot \frac{S_Y}{S_X}(x - \bar{x}) = \bar{y} + r \cdot \frac{S_Y}{S_X}(x - \bar{x}),$$

expresión de la recta de regresión de Y sobre X en función del coeficiente de correlación, r.

Invitaremos al lector a que demuestre que la recta de regresión de X sobre Y puede escribirse como

$$y = \bar{y} + \frac{1}{r} \cdot \frac{S_Y}{S_X}(x - \bar{x}).$$

2.36 Analícese la relación que existe entre las pendientes de las rectas de regresión de Y sobre X y de X sobre Y y los correspondientes coeficientes de regresión.

SOLUCIÓN

La recta de regresión de Y sobre X es

$$y = \overline{y} + \frac{S}{S_X^2}(x - \overline{x}),$$

cuya pendiente, cantidad que multiplica a x, $p_{Y/X}$, es S/S_X^2, por lo que

$$p_{Y/X} = b_{Y/X}.$$

En cuanto a la recta de regresión de X sobre Y,

$$x = \overline{x} + \frac{S}{S_Y^2}(y - \overline{y}),$$

o, lo que es igual, despejando,

$$y = \overline{y} + \frac{S_Y^2}{S}(x - \overline{x}),$$

su pendiente, cantidad que multiplica a x, tras despejar la variable y, $p_{X/Y}$, es S_Y^2/S, con lo cual se deduce que, en este caso,

$$p_{X/Y} = \frac{1}{b_{X/Y}}.$$

Como puede observarse, los coeficientes de regresión tienen el mismo signo que las pendientes de las rectas de regresión, signo que coincide, a su vez, con el de la covarianza, S, y, consecuentemente, con el signo del coeficiente de correlación, r.

Comparemos ahora las pendientes de ambas rectas de regresión, partiendo de sus expresiones en función del coeficiente de correlación lineal, obtenidas es **2.35**:

$$y = \overline{y} + r \cdot \frac{S_Y}{S_X}(x - \overline{x}),$$

recta de regresión de Y sobre X e

$$y = \bar{y} + \frac{1}{r} \cdot \frac{S_Y}{S_X}(x - \bar{x}),$$

recta de regresión de X sobre Y.

Puesto que

$$|r| \leq 1,$$

entonces, invirtiendo ambos miembros de la desigualdad anterior y cambiando, por tanto, el sentido de la misma,

$$\left|\frac{1}{r}\right| \geq 1,$$

se concluye que

$$\left|\frac{1}{r}\right| \geq |r|.$$

Consecuentemente, multiplicando ambos miembros por la cantidad positiva S_Y/S_X,

$$\left|\frac{1}{r}\right| \cdot \frac{S_Y}{S_X} \geq |r| \cdot \frac{S_Y}{S_X},$$

resulta que entre los valores absolutos de las pendientes de las rectas de regresión se cumple la siguiente relación:

$$|p_{X/Y}| \geq |p_{Y/X}|,$$

dándose la igualdad, si, y solamente si,

$$\left|\frac{1}{r}\right| = |r|,$$

hecho que ocurre cuando $|r| = 1$, es decir, cuando existe dependencia lineal perfecta y ambas rectas de regresión coinciden.

Desde el punto de vista práctico no suele ser razonable el cálculo de las dos rectas de regresión, ya que el sentido de la causalidad generalmente es único —o bien X depende de Y, o bien

Y depende de *X*—. Sin embargo, resulta de interés teórico analizar las relaciones existentes entre ambas rectas.

2.37 Sobre una muestra de 94 empresas se realiza un estudio sobre la situación laboral de los trabajadores. Sea *X* la variable que designa el número de trabajadores por empresa e *Y* la variable número de ellos con contrato temporal. La siguiente tabla recoge la distribución conjunta de estas variables.

Y / *X*	1	2	3
1-3	25	0	0
3-5	4	25	5
5-7	0	0	35

a) Calcúlese el número medio de trabajadores por empresa.

b) Hállese la función de regresión lineal de *Y* sobre *X*.

c) Estúdiese la bondad del ajuste realizado.

SOLUCIÓN

a) En la siguiente tabla se recoge la distribución marginal de la variable *X*:

x_i	$n_{i\cdot}$
2	25
4	34
6	35
	94

donde los valores x_i son las marcas de clase de los intervalos de la variable y $n_{i\cdot}$ las correspondientes frecuencias.

Por tanto, la media de esta distribución de frecuencias es

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_{i\cdot} = \frac{2 \cdot 25 + 4 \cdot 34 + 6 \cdot 35}{94} = 4{,}21 \text{ trabajadores.}$$

b) Para hallar la recta de regresión de Y sobre X obtenemos los momentos no centrales y centrales a partir del siguiente diagrama de apoyo, en el que se recogen solamente las filas y columnas necesarias para realizar los cálculos oportunos.

X \ Y	1	2	3	$n_{i\cdot}$	$x_i \cdot n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$	$\sum_{j=1}^{k} y_j \cdot n_{ij}$	$x_i \sum_{j=1}^{k} y_j \cdot n_{ij}$
2	25	0	0	25	50	100	25	50
4	4	25	5	34	136	544	69	276
6	0	0	35	35	210	1 260	105	630
$n_{\cdot j}$	29	25	40	94	396	1 904	199	956
$y_j^2 \cdot n_{\cdot j}$	29	100	360	489				

Así, los momentos no centrales son:

$$a_{0,1} = \frac{1}{N}\sum_{j=1}^{k} y_j \cdot n_{\cdot j} = \frac{199}{94} = 2{,}12,$$

$$a_{2,0} = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = \frac{1\,904}{94} = 20{,}26$$

y

$$a_{1,1} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = \frac{956}{94} = 10{,}17.$$

A partir de estos momentos, hallamos los momentos centrales, covarianza y varianza de X:

$$m_{1,1} = S = a_{1,1} - a_{1,0} \cdot a_{0,1} = 10{,}17 - 4{,}21 \cdot 2{,}12 = 1{,}25$$

y

$$m_{2,0} = S_X^2 = a_{2,0} - a_{1,0}^2 = 20{,}26 - 4{,}21^2 = 2{,}54.$$

En definitiva, la mejor explicación lineal del número de trabajadores con contrato laboral temporal con respecto al número total de trabajadores,

$$y - \bar{y} = \frac{S}{S_X^2}(x - \bar{x}),$$

es, sustituyendo por los valores calculados:

$$y - 2{,}12 = \frac{1{,}25}{2{,}54}(x - 4{,}21),$$

esto es,

$$y = 0{,}057 + 0{,}49 \cdot x.$$

c) Para hallar el coeficiente de determinación lineal,

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2},$$

además de los momentos obtenidos en el apartado anterior, necesitamos calcular la varianza de la variable *Y*:

$$S_Y^2 = a_{0,2} - \overline{y}^2 = 5{,}2 - 2{,}12^2 = 0{,}71.$$

Por consiguiente,

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2} = \frac{1{,}25^2}{2{,}54 \cdot 0{,}71} = 0{,}866,$$

lo cual significa que el 86,6 por ciento de la variabilidad de la variable *Y* ha resultado explicada por la regresión lineal realizada.

2.38 En la siguiente tabla se recogen los datos, en miles de euros, correspondientes al pasado año, referentes a gastos de personal y al beneficio anual de 200 empresas dedicadas al sector servicios.

Beneficio / Gastos	20-60	60-70	70-140
6-10	90	1	0
10-14	4	30	1
14-18	1	0	73

a) Obténgase la recta de regresión del beneficio sobre el gasto.

b) Calcúlese una medida de bondad del ajuste.

SOLUCIÓN

a) Los momentos necesarios para hallar la recta de regresión del beneficio, Y, sobre el gasto, X,

$$y - \overline{y} = \frac{S}{S_X^2}(x - \overline{x}),$$

se obtendrán utilizando el siguiente diagrama de apoyo en el que, prescindiendo de coincidencias, aparecen solo aquellas filas y columnas que intervienen en el cálculo de dichos momentos.

X \ Y	20-60	60-70	70-140	$n_{i\cdot}$	$x_i \cdot n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$	$\sum_{j=1}^{k} y_j \cdot n_{ij}$	$x_i \sum_{j=1}^{k} y_j \cdot n_{ij}$
6-10	90	1	0	91	728	5 824	3 665	29 320
10-14	4	30	1	35	420	5 040	2 230	26 760
14-18	1	0	73	74	1 184	18 944	8 800	140 800
$n_{\cdot j}$	95	31	74	200	2 332	29 808	14 695	196 880
$y_j^2 \cdot n_{\cdot j}$	152 000	130 975	1 065 600	1 348 575				

Los momentos no centrales,

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{h} x_i \cdot n_{i\cdot} = \frac{2\,332}{200} = 11{,}66 \text{ miles de euros},$$

$$\overline{y} = \frac{1}{N}\sum_{j=1}^{k} y_j \cdot n_{\cdot j} = \frac{14\,695}{200} = 73{,}475 \text{ miles de euros},$$

$$a_{2,0} = \frac{1}{N}\sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = \frac{29\,808}{200} = 149{,}04$$

y

$$a_{1,1} = \frac{1}{N}\sum_{i=1}^{h}\sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = \frac{196\,880}{200} = 984{,}4,$$

permiten hallar la varianza de X y la covarianza entre las variables:

$$S_X^2 = a_{2,0} - \overline{x}^2 = 149{,}04 - 11{,}66^2 = 13{,}08$$

y

$$S = a_{1,1} - \overline{x} \cdot \overline{y} = 984{,}4 - 11{,}66 \cdot 73{,}475 = 127{,}68.$$

Sustituyendo en la expresión de la recta de regresión, se obtiene que la mejor explicación lineal de Y sobre X es

$$y - 73{,}475 = \frac{127{,}68}{13{,}08}(x - 11{,}66),$$

esto es,

$$y = -40{,}32 + 9{,}76 \cdot x.$$

b) Con la varianza de la variable Y,

$$S_Y^2 = \frac{1}{N}\sum_{j=1}^{k} y_j^2 \cdot n_{\cdot j} - \bar{y}^2 = \frac{1\ 348\ 575}{200} - 73{,}475^2 = 1\ 344{,}3,$$

junto con la varianza de X y la covarianza entre X e Y, ya calculadas, se obtiene el coeficiente de determinación lineal,

$$r^2 = \frac{S^2}{S_X^2 \cdot S_Y^2} = \frac{127{,}68^2}{13{,}08 \cdot 1\ 344{,}3} = 0{,}927,$$

valor que muestra un buen grado de relación lineal entre las variables, indicando, así, que el ajuste realizado es correcto. Además, el signo positivo de la covarianza expresa que la relación lineal entre las variables es creciente.

2.39 Se considera la distribución de frecuencias:

x_i	– 1	0	1
$f_{i\cdot}$	1/3	1/3	1/3

Demuéstrese que las variables X e $Y = X^2$ están incorrelacionadas pero son dependientes.

SOLUCIÓN

Antes de resolver este ejercicio, proponemos al lector que dé respuesta a la siguiente pregunta: ¿en qué problema anterior se ha comentado ya la idea fundamental de que la incorrelación no es condición suficiente para la independencia entre variables?

La distribución conjunta de estas variables aparece en la siguiente tabla:

X \ Y	0	1
– 1	0	1/3
0	1/3	0
1	0	1/3

Las frecuencias conjuntas iguales a 1/3 corresponden a aquellos pares de valores cuya segunda componente, valor de Y, es igual al cuadrado de la primera componente, valor de X.

Puesto que $\overline{x} = 0$ y $\overline{y} = 2/3$, como puede comprobar el lector, se tiene que la covarianza de esta distribución es

$$S = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_j - \overline{y}) f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{ij} = -1 \cdot 1 \cdot \frac{1}{3} + 0 \cdot 0 \cdot \frac{1}{3} + 1 \cdot 1 \cdot \frac{1}{3} = 0,$$

con lo cual, las variables X e Y están incorrelacionadas.

Por otro lado, es obvio que las variables X e Y son dependientes, puesto que $Y = X^2$, con lo cual, existe dependencia funcional *perfecta* entre ellas. En cualquier caso, puede comprobarse del modo habitual, que estas variables no son independientes, ya que, por ejemplo, $f_{11} = 0$ no coincide con

$$f_{1\cdot} \cdot f_{\cdot 1} = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}.$$

Si nuestro interés fuera obtener la recta de regresión de Y sobre X, llegaríamos a que ésta es $y = \overline{y} = 2/3$, con pésimo ajuste, pues r^2 es igual a cero: ¡estamos empeñándonos en explicar *linealmente* la variable Y en función de la X cuando la forma de la relación es *parabólica*!

2.40 Las siguientes distribuciones representan, en cientos de euros, la renta, Y, y el gasto en ocio, X, de 100 individuos, en un cierto año.

Renta	80-100	100-150	150-300
N.º individuos	20	50	30

Gasto en ocio	1-2	2-4	4-8
N.º individuos	15	75	10

La mejor explicación lineal del gasto en función de la renta viene dada por la ecuación $x = 3{,}075$.

a) ¿Es esta información suficiente para afirmar que estas variables son independientes?

b) Calcúlese la recta de regresión de la renta en función del gasto. Desde el punto de vista económico, ¿está justificado el sentido de esta relación?

SOLUCIÓN

a) Al ser la recta de regresión de X sobre Y igual a una constante, la covarianza entre las variables es cero, es decir, las variables están incorrelacionadas, lo cual, como sabemos, implica que no existe relación *lineal* entre ellas, no que sean independientes. Puesto que la incorrelación es condición *necesaria* pero *no suficiente* para la independencia, podemos decir que esta información no basta para afirmar que estas variables sean independientes.

En cualquier caso, el resto del enunciado tampoco proporciona información para poder analizar la posible independencia entre las variables, ya que dicho análisis requiere la comparación de los productos de las frecuencias marginales, que aparecen en las tablas anteriores, con las correspondientes frecuencias de la distribución conjunta de las que no se dispone.

b) Como la covarianza entre las variables es igual a cero, la recta de regresión de la renta, Y, en función del gasto, X, será $y = \overline{y}$, paralela al eje de ordenadas.

Utilizando las marcas de clase de la distribución agrupada en intervalos correspondiente a la variable Y, puede comprobar el lector que se obtiene un valor medio, $\overline{y} = 148$, con lo cual la recta de regresión de Y sobre X es

$$y = 148.$$

En realidad, no parece tener demasiado sentido, desde el punto de vista económico, que la renta de los individuos dependa de lo que éstos gasten en ocio.

2.41 El gerente del servicio de transportes urbanos de la comarca de Villamayor cree que el número de viajeros depende del precio del billete. En el municipio hay 15 líneas, que recorren las diferentes zonas de la comarca, variando en cada una de las líneas el precio del billete en función del tipo de recorrido y de la distancia máxima de la misma.

En un análisis de la relación existente entre el precio del billete, en céntimos de euro, X, y el número de viajeros, en cientos, Y, que utilizan diariamente estos servicios, se obtiene que el coeficiente de correlación lineal es igual a -1, hecho que resulta suficiente para que el gerente afirme que el aumento de precio de billete es el único motivo del descenso en el número de viajeros. ¿Qué opinión estadística merece esta conclusión?

SOLUCIÓN

Cualquier análisis de regresión tiene que tener un fundamento teórico que profundice en la naturaleza de las relaciones entre las variables y que apoye el ejercicio estadístico llevado a cabo. Esto es necesario pues, en caso contrario, podríamos encontrarnos con ajustes óptimos que son, en realidad, resultado de la *casualidad* y no de una verdadera relación *causa-efecto* entre las variables estudiadas.

Conviene también alertar al lector sobre el hecho de que un elevado coeficiente de correlación lineal entre dos variables puede ser la consecuencia de la influencia implícita de una tercera variable. Así, en este ejemplo, podría suceder que hubiera otra variable que influyera en el aumento del precio del billete que sería la que, realmente, produciría el descenso en el número de viajeros.

2.41 Demuéstrese que las rectas de regresión de X sobre Y y de Y sobre X se cortan en el punto $(\overline{x}, \overline{y})$.

SOLUCIÓN

Si sustituimos el valor de y de la recta de regresión de Y sobre X,

$$y = \overline{y} + \frac{S}{S_X^2} \cdot x - \frac{S}{S_X^2} \cdot \overline{x},$$

en la recta de regresión de X sobre Y,

$$x - \overline{x} = \frac{S}{S_Y^2} (y - \overline{y}),$$

tendremos:

$$x - \overline{x} = \frac{S}{S_Y^2} \left(\overline{y} + \frac{S}{S_X^2} \cdot x - \frac{S}{S_X^2} \cdot \overline{x} - \overline{y} \right).$$

Simplificando y operando, resulta que

$$x - \overline{x} = x \cdot \frac{S^2}{S_X^2 \cdot S_Y^2} - \overline{x} \cdot \frac{S^2}{S_X^2 \cdot S_Y^2}.$$

Por último, agrupando términos semejantes,

$$x\left(1-\frac{S^2}{S_X^2\cdot S_Y^2}\right)=\overline{x}\left(1-\frac{S^2}{S_X^2\cdot S_Y^2}\right),$$

es decir,

$$x\,(1-r^2)=\overline{x}\,(1-r^2).$$

Si suponemos que

$$1-r^2=0,$$

esto es, si $r^2=1$, entonces, el ajuste el perfecto y, como sabemos, las dos rectas de regresión coinciden.

Si, por el contrario,

$$1-r^2\neq 0,$$

entonces, dividiendo por esa cantidad los dos miembros de la igualdad se obtiene que x es $\overline{x}$ y, sustituyendo, por ejemplo, en la recta de regresión de Y sobre X resulta un valor de y igual a $\overline{y}$; en definitiva, las rectas de regresión tienen como punto de corte $(\overline{x}, \overline{y})$, punto denominado *centro de gravedad*.

2.43 Al finalizar la campaña publicitaria de Navidad, y de cara a preparar la del próximo año, la empresa de productos cosméticos Santa Lorena analiza los siguientes datos correspondientes a 40 fragancias. La variable X indica el número de anuncios emitidos de dichas fragancias durante la Navidad, y la variable Y refleja el número de unidades, en miles, vendidas en esta época navideña. Sean las rectas de regresión:

$$y=4{,}3+5{,}1\cdot x$$
$$y=-6{,}5+5{,}7\cdot x.$$

Calcúlese:

a) La media de unidades vendidas por fragancia en ese periodo.

b) El número medio de anuncios emitidos por fragancia.

c) La proporción de la variabilidad de la variable Y que viene explicada por la correspondiente regresión.

SOLUCIÓN

a) La media de unidades vendidas por fragancia en ese periodo y el número medio de anuncios emitidos se obtienen resolviendo el anterior sistema de ecuaciones, ya que los valores de este modo calculados, punto de corte de las dos rectas de regresión, son las medias de ambas distribuciones.

Así, igualando las ecuaciones,

$$4{,}3 + 5{,}1 \cdot x = -6{,}5 + 5{,}7 \cdot x,$$

y, agrupando términos semejantes,

$$4{,}3 + 6{,}5 = 5{,}7 \cdot x - 5{,}1 \cdot x,$$

resulta el valor

$$x = \frac{10{,}8}{0{,}6} = 18,$$

es decir, el número medio de anuncios emitidos por fragancia, $\overline{x}$, es igual a 18.

b) Para calcular la media de unidades vendidas por fragancia, sustituimos el valor hallado en el apartado anterior, por ejemplo, en la primera ecuación:

$$y = 4{,}3 + 5{,}1 \cdot 18 = 96{,}1,$$

con lo cual, el número medio de unidades vendidas por fragancia, $\overline{y}$, es de 96 100.

c) La proporción de la variabilidad de la variable Y explicada por la correspondiente regresión es, por definición, el coeficiente de determinación.

Puesto que como dato del problema tenemos las dos rectas de regresión, el procedimiento de cálculo de este coeficiente consiste en el empleo de los coeficientes de regresión:

$$r^2 = b_{Y/X} \cdot b_{X/Y}.$$

Ahora bien, como sabemos por **2.36**, la recta de regresión de Y sobre X es la que tiene menor pendiente en valor absoluto, por lo cual, dicha recta de regresión es

$$y = 4{,}3 + 5{,}1 \cdot x,$$

y su coeficiente de regresión, que coincide con la pendiente, es

$$b_{Y/X} = 5{,}1.$$

Sin embargo, en la recta de regresión de X sobre Y,

$$y = -6{,}5 + 5{,}7 \cdot x,$$

el coeficiente de regresión es el inverso de la pendiente, por lo que

$$b_{X/Y} = \frac{1}{5{,}7}.$$

En definitiva, el coeficiente de determinación será

$$r^2 = 5{,}1 \cdot \frac{1}{5{,}7} = 0{,}89,$$

quedando, por tanto, un 89 por ciento de la variable Y explicada por la regresión lineal de Y sobre X.

2.41 Sean X e Y las variables que designan, en euros, la renta y el consumo en alimentación mensual, respectivamente, de un grupo de familias. Las rectas de regresión mínimo-cuadráticas correspondientes a estas variables son:

$$y = \frac{x}{10}$$

$$y = \frac{x}{9} - 20.$$

Hállese:

a) La renta media y el consumo medio en alimentación mensual por familia.

b) Los coeficientes de regresión de cada una de las rectas.

c) El coeficiente de correlación lineal.

SOLUCIÓN

a) La renta media y el consumo medio en alimentación mensual por familia, es decir, $\bar{x}$ e $\bar{y}$, se obtienen resolviendo el sistema de ecuaciones correspondiente a las dos rectas de regresión. Así, igualando ambas ecuaciones, se tiene que

$$\frac{x}{10} = \frac{x}{9} - 20,$$

expresión que, tras sencillas operaciones, conduce al valor

$$x = 1\ 800.$$

Sustituyendo esta cantidad, por ejemplo, en la primera ecuación, resulta

$$y = \frac{1\ 800}{10} = 180.$$

En definitiva, la renta media, $\overline{x}$, y el consumo medio en alimentación mensual por familia, $\overline{y}$, son, respectivamente, 1 800 y 180 euros.

b) Para identificar las rectas anteriores, hemos de considerar, según vimos en **2.36**, que la recta de regresión de Y sobre X es la de menor pendiente en valor absoluto. Como la recta

$$y = \frac{x}{10}$$

tiene una pendiente igual a 1/10 y a la recta

$$y = \frac{x}{9} - 20$$

le corresponde una pendiente de 1/9, se concluye que

$$y = \frac{x}{10}$$

es la recta de regresión de Y sobre X, siendo $b_{Y/X} = 1/10$ el correspondiente coeficiente de regresión.

Además,

$$y = \frac{x}{9} - 20$$

es la recta de regresión de X sobre Y y su coeficiente de regresión es $b_{X/Y} = 9$, valor inverso a la pendiente de dicha recta.

c) El coeficiente de correlación lineal, raíz cuadrada del coeficiente de determinación lineal, con signo igual al de las pendientes de las rectas —en este caso positivo—, se calcula a partir de los coeficientes de regresión:

$$r = \sqrt{b_{Y/X} \cdot b_{X/Y}} = \sqrt{\frac{1}{10} \cdot 9} = 0{,}95.$$

2.45 Demuéstrese que, si, para cualquier j, $\overline{x}/(Y = y_j) = \overline{x}$, entonces, las variables X e Y están incorrelacionadas.

SOLUCIÓN

Puesto que, dada la distribución condicionada $(x_i/Y = y_j; f_{i/j})$, se verifica que

$$f_{i/j} = \frac{f_{ij}}{f_{\cdot j}},$$

despejando la frecuencia relativa conjunta, se tiene:

$$f_{ij} = f_{i/j} \cdot f_{\cdot j}.$$

Sustituyendo f_{ij} de la igualdad anterior en la definición de covarianza,

$$S = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{ij} - \overline{x} \cdot \overline{y} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot f_{i/j} \cdot f_{\cdot j} - \overline{x} \cdot \overline{y},$$

y reagrupando términos semejantes dentro de los sumatorios, resulta:

$$S = \sum_{j=1}^{k} y_j \cdot f_{\cdot j} \sum_{i=1}^{h} x_i \cdot f_{i/j} - \overline{x} \cdot \overline{y} = \sum_{j=1}^{k} y_j \cdot f_{\cdot j} \cdot (\overline{x}/(Y = y_j)) - \overline{x} \cdot \overline{y},$$

pues, por definición,

$$\overline{x}/(Y = y_j) = \sum_{j=1}^{h} x_i \cdot f_{i/j}.$$

Reemplazando $\overline{x}/(Y = y_j)$ por $\overline{x}$ en la expresión anterior, se obtiene que

$$S = \sum_{j=1}^{k} y_j \cdot f_{\cdot j} \cdot \overline{x} - \overline{x} \cdot \overline{y} = \overline{x} \sum_{j=1}^{k} y_j \cdot f_{\cdot j} - \overline{x} \cdot \overline{y} = \overline{x} \cdot \overline{y} - \overline{x} \cdot \overline{y} = 0,$$

estando incorrelacionadas las variables X e Y.

2.46 Dada la distribución de frecuencias bidimensional:

X \ Y	5	7
2	1	0
3	0	1

Dígase, sin hacer operaciones, cuál es el valor del coeficiente de determinación lineal entre X e Y.

SOLUCIÓN

La tabla anterior muestra que el diagrama de dispersión de la distribución de frecuencias bidimensional está formado por dos únicos puntos: (2;5) y (3;7). Como se sabe, por dos puntos pasa una sola recta, por lo que existe un ajuste lineal *perfecto* entre X e Y y, en definitiva, el coeficiente de determinación lineal entre estas dos variables es igual a 1.

En todo caso, este ejercicio está planteado solamente desde un punto de vista didáctico con objeto de que el lector fije los conceptos estudiados, pues el coeficiente de determinación lineal es tanto más fiable cuanto mayor sea el número de observaciones[4].

2.47 Dada una distribución de frecuencias bidimensional $(x_i, y_j; f_{ij})$, cuyo coeficiente de correlación lineal es r, obténgase el coeficiente de correlación lineal de la distribución de frecuencias $(a \cdot x_i + b, c \cdot y_j + d; f_{ij})$, siendo a y b números reales positivos. En particular, calcúlese el coeficiente de correlación lineal de la distribución transformada por un cambio de origen y de escala en cada una de las variables.

SOLUCIÓN

Denotando por S_X y S_Y a las desviaciones típicas de las distribuciones marginales $(x_i, f_{i\cdot})$ e $(y_j; f_{\cdot j})$, resulta, según se demostró en el capítulo 1, que las desviaciones típicas de las distribuciones transformadas, $(a \cdot x_i + b; f_{i\cdot})$ y $(c \cdot y_j + d; f_{\cdot j})$, son, respectivamente,

$$|a| \cdot S_X$$

y

$$|c| \cdot S_Y \cdot$$

Por otro lado, en el problema **2.21** se probó que la covarianza de la variable transformada es

$$S' = a \cdot c \cdot S,$$

donde S es la covarianza de la distribución bidimensional $(x_i, y_j; f_{ij})$.

---

[4] Este hecho tiene que ver con el concepto de *grados de libertad* que el lector puede consultar en cualquier libro de inferencia estadística.

En consecuencia, el coeficiente de correlación de la nueva distribución es, sin mas que sustituir,

$$r' = \frac{a \cdot c \cdot S}{|a| \cdot |c| \cdot S_X \cdot S_Y} = \frac{a \cdot c}{|a| \cdot |c|} \cdot r,$$

con lo cual, si a y c tienen el mismo signo,

$$r' = r$$

y, si tienen distinto signo,

$$r' = -r.$$

En particular, si $a = 1/e_1$, $b = -o_1/e_1$, $c = 1/e_2$ y $d = -o_2/e_2$, es decir, si realizamos un cambio de origen y de escala sobre las variables, teniendo en cuenta que $e_1, e_2 > 0$, el coeficiente de correlación de la distribución transformada coincide con r. En consecuencia, el coeficiente de correlación lineal no se ve afectado por cambios ni de origen ni de escala en las variables.

2.48 La empresa Eduarsa, dedicada a la plantación de kiwis, posee 20 fincas distribuidas por el territorio nacional. El rendimiento de la finca, Y, en toneladas, así como la superficie de la misma, X, en hectáreas, se refleja en la siguiente tabla:

X \ Y	10	11	12
1	4	0	0
2	0	5	1
3	0	0	10

a) Sabiendo que el rendimiento de una finca depende de la superficie de ésta, obténgase una medida del grado de relación lineal existente entre las variables.

b) Si el kilo de kiwis se vende a mayoristas a 1,5 euros, hállese el grado de dependencia lineal de los ingresos y la superficie.

SOLUCIÓN

a) Para obtener el coeficiente de correlación lineal, medida del grado de relación lineal entre las variables,

$$r = \frac{S}{S_X \cdot S_Y},$$

hallaremos los momentos no centrales y centrales, apoyándonos en el siguiente diagrama.

X \ Y	10	11	12	$n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$
1	4	0	0	4	4
2	0	5	1	6	24
3	0	0	10	10	90
$n_{\cdot j}$	4	5	11	20	118
$y_j \cdot n_{\cdot j}$	40	55	132	227	
$y_j^2 \cdot n_{\cdot j}$	400	605	1 584	2 589	
$\sum_{i=1}^{h} x_i \cdot n_{ij}$	4	10	32	46	
$y_j \sum_{i=1}^{h} x_i \cdot n_{ij}$	40	110	384	534	

Así, la superficie media por finca es

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_{i\cdot} = \frac{46}{20} = 2{,}3 \text{ hectáreas}$$

y el rendimiento medio

$$\bar{y} = \frac{1}{N} \sum_{j=1}^{k} y_j \cdot n_{\cdot j} = \frac{227}{20} = 11{,}35 \text{ toneladas.}$$

Además,

$$a_{2,0} = \frac{1}{N} \sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = \frac{118}{20} = 5{,}9,$$

con lo cual, la varianza de la variable X es

$$S_X^2 = a_{2,0} - \bar{x}^2 = 5{,}9 - 2{,}3^2 = 0{,}61.$$

De modo análogo,

$$a_{0,2} = \frac{1}{N} \sum_{j=1}^{k} y_j^2 \cdot n_{\cdot j} = \frac{2\,589}{20} = 129{,}45$$

y, en consecuencia, la varianza de Y es

$$S_Y^2 = a_{0,2} - \bar{y}^2 = 129{,}45 - 11{,}35^2 = 0{,}6275.$$

Por último, el momento de orden (1, 1) respecto al origen,

$$a_{1,1} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = \frac{534}{20} = 26{,}7,$$

permite calcular la covarianza de las variables:

$$S = a_{1,1} - \bar{x} \cdot \bar{y} = 26{,}7 - 2{,}3 \cdot 11{,}35 = 0{,}595.$$

En definitiva, el coeficiente de correlación lineal es

$$r = \frac{S}{S_X \cdot S_Y} = \frac{0{,}595}{0{,}78 \cdot 0{,}79} = 0{,}965.$$

La interpretación de este coeficiente es clara: existe un alto grado de dependencia lineal entre la superficie de la finca y el rendimiento de la misma, siendo, además, su signo positivo reflejo de una relación creciente entre ambas variables.

Puede comprobar el lector que, si suponemos que es la variable Y, rendimiento, la que depende de X, superficie, la recta de regresión de Y sobre X es

$$y = 9{,}1075 + 0{,}975 \cdot x,$$

recta cuyo coeficiente de regresión expresa que un incremento de una unidad en la variable X, esto es, de una hectárea, supondría un incremento de 0,975 unidades en la variable Y, es decir, de 975 kilos. Además, dado que $r^2 = 0{,}9312$, el 93,12 por ciento de la varianza de Y está explicada por la regresión lineal.

b) De la distribución inicial $(x_i, y_j; n_{ij})$, hemos pasado a una distribución transformada, $(x_i, 1\,500 \cdot y_j; n_{ij})$, distribución bidimensional de las variables, superficie de una finca, en hectáreas, X, e ingresos, en miles de euros, $1{,}5 \cdot Y$.

La transformación realizada en la variable Y es un cambio de escala con $e_2 = 1/1\,500$, por lo cual, según vimos en el problema anterior, el coeficiente de correlación entre la superficie de la finca y los ingresos coincide con el coeficiente de correlación entre la superficie de la finca y el rendimiento, esto es, 0,965.

2.49 Demuéstrese que el coeficiente de correlación lineal de la distribución $(x_i, y_j; f_{ij})$ es igual a la covarianza de las variables tipificadas.

SOLUCIÓN

Sean

$$U = \frac{X - \overline{x}}{S_X}$$

y

$$V = \frac{Y - \overline{y}}{S_Y}$$

las variables tipificadas de X e Y, respectivamente.

La covarianza de U y V es, por definición,

$$S_{U,V} = \sum_{i=1}^{h} \sum_{j=1}^{k} (u_i - \overline{u}) \cdot (v_j - \overline{v}) f_{ij}.$$

Ahora bien, según se demostró en el capítulo anterior, $\overline{u} = \overline{v} = 0$, con lo cual, sustituyendo,

$$S_{U,V} = \sum_{i=1}^{h} \sum_{j=1}^{k} u_i \cdot v_j \cdot f_{ij} = \sum_{i=1}^{h} \sum_{j=1}^{k} \left(\frac{x_i - \overline{x}}{S_X} \right) \cdot \left(\frac{y_j - \overline{y}}{S_Y} \right) f_{ij},$$

siendo el último miembro de la igualdad resultado de sustituir los valores de las variables U y V en función de los valores de las variables X e Y.

Operando se obtiene:

$$S_{U,V} = \frac{1}{S_X \cdot S_Y} \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{x}) \cdot (y_j - \overline{y}) f_{ij} = \frac{S}{S_X \cdot S_Y} = r_{X,Y},$$

como pretendíamos probar.

2.50 Obténgase la mejor explicación de la variable Y en función de la variable X según el modelo *potencial*:

$$y = a \cdot x^b,$$

aplicando el criterio de los mínimos cuadrados.

SOLUCIÓN

Para obtener los valores de a y b por aplicación del criterio de los mínimos cuadrados, se puede trabajar de igual modo que en la regresión lineal, esto es, haciendo mínima la suma de los

cuadrados de las distancias entre los valores observados y los valores teóricos del modelo o, lo que es igual, minimizando los residuos al cuadrado:

$$\sum_{i=1}^{h}\sum_{j=1}^{k} e_{ij}^2 \cdot f_{ij} = \sum_{i=1}^{h}\sum_{j=1}^{k} (y_j - a \cdot x^b)^2 f_{ij}.$$

Existe, sin embargo, otra forma de trabajar que resulta más cómoda: se trata de *linealizar* la función del modelo de regresión planteado. Pasamos así de un modelo de regresión *potencial* a un modelo de regresión *lineal*, para el cual los parámetros están calculados.

En efecto, si

$$y = a \cdot x^b,$$

entonces, tomando logaritmos [5], se tiene la relación equivalente

$$\ln y = \ln a + b \cdot \ln x$$

que, oportunos cambios de variable

$$\ln Y = V,$$
$$\ln X = U$$

y

$$\ln a = c$$

con $a > 0$, permiten escribir como

$$v = c + b \cdot u.$$

Las estimaciones de los parámetros c y b en la regresión lineal de V sobre U son, como es sabido,

$$b = \frac{S_{U,V}}{S_U^2}$$

y

$$c = \overline{v} - \frac{S_{U,V}}{S_U^2} \cdot \overline{u},$$

[5] La base de los logaritmos puede ser cualquiera.

con lo cual,

$$a = \exp(c) = \exp\left(\overline{v} - \frac{S_{U,V}}{S_U^2} \cdot \overline{u}\right).$$

2.51 Obténgase la mejor explicación de la variable Y en función de la variable X según el modelo *exponencial*:

$$y = a \cdot b^x.$$

SOLUCIÓN

Repitiendo el procedimiento llevado a cabo en el ejercicio anterior, transformamos linealmente el modelo, tomando logaritmos:

$$\ln y = \ln a + x \cdot \ln b.$$

Realizando el cambio de variable

$$\ln Y = V$$

y denotando

$$c = \ln a$$

y

$$d = \ln b,$$

con $a, b > 0$, resulta el modelo lineal:

$$v = c + d \cdot x.$$

Por aplicación del criterio de los mínimos cuadrados se obtienen, en este caso, los valores

$$d = \frac{S_{X,V}}{S_X^2}$$

y

$$c = \overline{v} - \frac{S_{X,V}}{S_X^2} \cdot \overline{x}.$$

Y, en definitiva,

$$b = \exp(d) = \exp\left(\frac{S_{X,V}}{S_X^2}\right)$$

y

$$a = \exp(c) = \exp\left(\overline{v} - \frac{S_{X,V}}{S_X^2} \cdot \overline{x}\right).$$

2.52 La empresa Telepastel, dedicada a la venta de dulces a domicilio, tiene centros de venta en 10 ciudades españolas. Durante el pasado año se repartieron folletos de propaganda por los buzones, siendo X el número de folletos repartidos, en miles, e Y, los ingresos por ventas, en miles de euros, en cada una de las ciudades.

N.º folletos	Ingresos
1	6
1,5	8
2	12
2,5	17
3	25
4	45
4,5	70
5	96
6	190
6,2	250

a) Represéntese la nube de puntos de la distribución de frecuencias bidimensional.

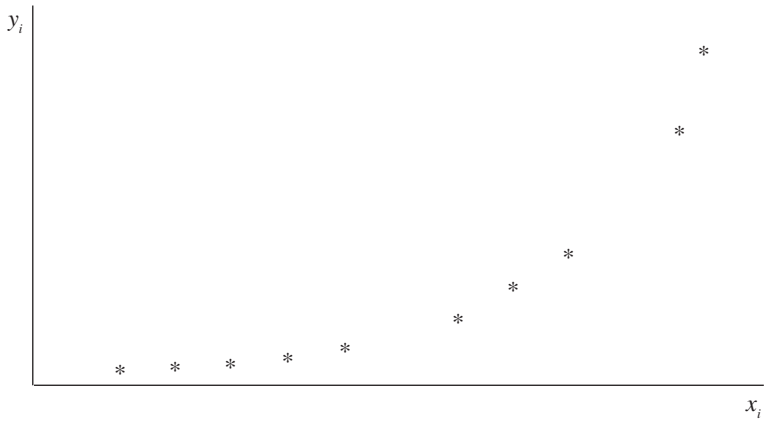
b) Obténgase, a la vista de la gráfica anterior, la ecuación de regresión que mejor refleje la dependencia de los ingresos del número de folletos de propaganda emitidos.

c) Analícese la bondad del ajuste realizado.

SOLUCIÓN

a) La representación de los pares de puntos de la distribución de frecuencias unitaria que proporciona el enunciado sugiere que éstos se alinean en torno a una curva exponencial, intu-

yéndose, por tanto, la posible existencia de una dependencia exponencial de la variable *Y*, ingresos, con respecto a *X*, número de folletos de propaganda.



b) A la vista de la gráfica anterior, lo más acertado es aplicar el criterio de los mínimos cuadrados para obtener los valores a y b que proporcionen la mejor explicación de Y sobre X, según el modelo:

$$y = a \cdot b^x.$$

Siguiendo los pasos de **2.51**, linealizamos el modelo exponencial, tomando logaritmos:

$$\ln y = \ln a + x \cdot \ln b,$$

con lo cual, realizando el cambio de variable

$$V = \ln Y$$

y denotando

$$c = \ln a$$

y

$$d = \ln b,$$

resulta el modelo:

$$v = c + d \cdot x,$$

que permite la estimación de c y d en un modelo lineal.

En la siguiente tabla se recogen los cálculos que servirán de apoyo en la obtención de los distintos momentos necesarios para hallar la ecuación de regresión exponencial. Los datos de cada casilla de la última fila son la suma de los elementos de la correspondiente columna.

x_i	y_i	$v_i = \ln y_i$	$x_i \cdot v_i$	x_i^2	v_i^2
1	6	1,79	1,790	1	3,20
1,5	8	2,08	3,120	2,25	4,33
2	12	2,48	4,960	4	6,15
2,5	17	2,83	7,075	6,25	8,01
3	25	3,22	9,660	9	10,37
4	45	3,81	15,240	16	14,52
4,5	70	4,25	19,125	20,25	18,06
5	96	4,56	22,800	25	20,79
6	190	5,25	31,500	36	27,56
6,2	250	5,52	34,224	38,44	30,47
35,7	**719**	**35,79**	**149,494**	**158,19**	**143,46**

Los parámetros de la recta de regresión de V sobre X son

$$d = \frac{S_{X,V}}{S_X^2}$$

y

$$c = \overline{v} - \frac{S_{X,V}}{S_X^2} \cdot \overline{x},$$

con lo cual, hemos de calcular las medias de las variables,

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{35,7}{10} = 3,57 \text{ miles de folletos}$$

y

$$\overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i = \frac{35,79}{10} = 3,58,$$

y la varianza de X y la covarianza entre X y V:

$$S_X^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2 = \frac{158,19}{10} - 3,57^2 = 3,074$$

y

$$S_{X,V} = \frac{1}{N}\sum_{i=1}^{N} x_i \cdot v_i - \overline{x} \cdot \overline{v} = \frac{149{,}494}{10} - 3{,}57 \cdot 3{,}58 = 2{,}1688.$$

En definitiva,

$$d = \frac{S_{X,V}}{S_X^2} = \frac{2{,}1688}{3{,}074} = 0{,}7055$$

y

$$c = \overline{v} - \frac{S_{X,V}}{S_X^2} \cdot \overline{x} = 3{,}58 - 0{,}7055 \cdot 3{,}57 = 1{,}061,$$

y, al aplicar los resultados de **2.51**, se tiene que

$$b = \exp(d) = 2{,}025$$

y

$$a = \exp(c) = 2{,}889.$$

En consecuencia, la ecuación mínimo-cuadrática que expresa la dependencia exponencial existente entre el ingreso y el número de folletos emitidos es

$$y = 2{,}889 \cdot 2{,}025^{x}.$$

c) El estudio de la bondad del ajuste de la variable Y sobre X en modelos no lineales mediante el coeficiente de determinación:

$$R^2 = \frac{S_{\hat{Y}}^2}{S_Y^2},$$

denotado con una letra mayúscula con el fin de diferenciarlo del coeficiente de correlación lineal, puede ser erróneo, porque, al no estar garantizada la descomposición de la varianza de Y, podría ocurrir que el coeficiente tomara valores no comprendidos entre 0 y 1. Este hecho hace conveniente el empleo del cociente:

$$\frac{S_e^2}{S_Y^2},$$

como medida de bondad de ajuste en modelos no lineales, pues parece coherente con el criterio de los mínimos cuadrados la comparación de la varianza residual con la varianza de la varia-

ble que queremos explicar por el procedimiento de regresión. El empleo de este coeficiente permitirá la discriminación entre modelos, considerándose más adecuado aquel cuya varianza residual sea menor en relación con la varianza de la variable Y.

Conviene también mencionar que no debe analizarse la bondad del ajuste del modelo no lineal a partir del modelo linealizado, pues, en general, $1 - r^2$, esto es, la proporción que la varianza de los residuos en la regresión lineal del modelo transformado representa sobre la varianza de Y, no coincide con S_e^2 / S_Y^2, proporción que la varianza residual del modelo no lineal representa sobre la varianza de la variable Y.

En la siguiente tabla figuran, además de los valores de las variables Y e $\tilde{Y}$, los valores de la variable e, así como sus cuadrados, lo cual facilitará el cálculo de las correspondientes varianzas.

x_i	y_i	$\tilde{y}_i$	e_i	e_i^2
1	6	5,85	0,15	0,0225
1,5	8	8,33	–0,33	0,1089
2	12	11,85	0,15	0,0225
2,5	17	16,86	0,14	0,0196
3	25	23,99	1,01	1,0201
4	45	48,58	–3,58	12,8164
4,5	70	69,13	0,87	0,7569
5	96	98,37	–2,37	5,6169
6	190	199,20	–9,20	84,64
6,2	250	229,39	20,61	424,7721
35,7	**719**	**711,55**	**7,45**	**529,7959**

Los valores teóricos, $\tilde{y}_i$, se han obtenido, aplicando la ecuación de ajuste anterior. Por ejemplo, $\tilde{y}_6 = 48{,}58$ se ha calculado mediante el valor $x_6 = 4$ como $2{,}889 \cdot 2{,}025^4$; en cuanto a los valores de la variable e se han hallado como $e_i = y_i - \tilde{y}_i$.

La varianza de Y es

$$S_Y^2 = a_{0,2} - \overline{y}^2,$$

donde

$$\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{719}{10} = 71{,}9 \text{ miles de euros}$$

y

$$a_{0,2} = \frac{1}{N} \sum_{i=1}^{N} y_i^2 = \frac{115\,899}{10} = 11\,589{,}9,$$

con lo cual,

$$S_Y^2 = 11\ 589{,}9 - 71{,}9^2 = 6\ 420{,}29.$$

En cuanto a la varianza de e, varianza residual,

$$S_e^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 - \bar{e}^2,$$

se tiene, dado que el modelo no es lineal, que, por un lado, la media de los residuos no es nula,

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{7{,}45}{10} = 0{,}745,$$

y, por otro lado,

$$\frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{529{,}7959}{10} = 52{,}97959,$$

por lo que la varianza de los residuos es

$$S_e^2 = 52{,}97959 - 0{,}745^2 = 52{,}42.$$

Finalmente,

$$\frac{S_e^2}{S_Y^2} = \frac{52{,}42}{6\ 420{,}29} = 0{,}0082,$$

valor próximo a 0 que apoya la hipótesis de relación de las variables conforme a un modelo exponencial. En concreto la varianza de los residuos representa únicamente un 0,82 por ciento de la varianza de la variable Y.

Según comentamos al principio de este apartado, en este caso, se comprueba cómo la varianza de Y,

$$S_Y^2 = 6\ 420{,}29,$$

no coincide con la suma de varianzas de los residuos y de los valores teóricos:

$$S_e^2 + S_{\hat{Y}}^2 = 52{,}42 + 5\ 958{,}952 = 6\ 011{,}372.$$

2.53 En una residencia hospitalaria se desea estudiar la posible relación entre la edad y el gasto en medicamentos. Para ello se ha elegido una muestra de 10 individuos, cuyas

edades, X, y gastos mensuales en medicamentos, Y, en euros, figuran en la siguiente tabla.

Edad	Gasto
30	27
40	60
50	120
60	200
70	350
75	500
80	510
85	610
90	740
95	900

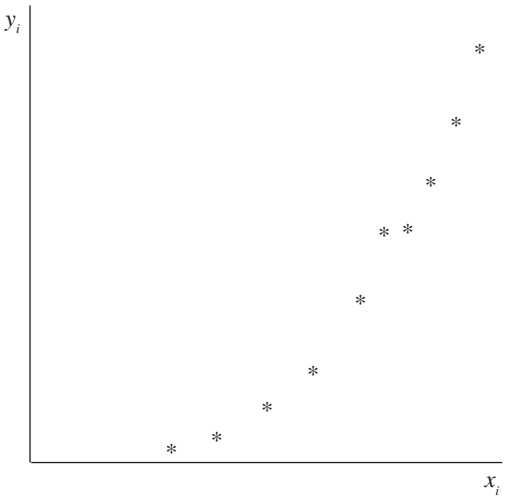
a) Represéntese el diagrama de dispersión de esta distribución de frecuencias.

b) Obténgase, a la vista de la gráfica anterior, la ecuación de regresión que mejor refleje la dependencia estadística de los gastos en medicamentos de la edad de los individuos.

c) Analícese la bondad del ajuste realizado.

SOLUCIÓN

a) Mediante la representación de los pares de puntos (x_i, y_i), se obtiene el siguiente diagrama de dispersión:



b) En el cumplimiento del primer objetivo en la resolución de un problema de regresión como es la determinación de la *forma* de la dependencia existente entre las variables, parece adecuado considerar, a la vista de la representación gráfica, una ecuación de ajuste potencial:

$$y = a \cdot x^b.$$

Según se vio en **2.50**, el procedimiento más sencillo para hallar los parámetros a y b del modelo anterior consiste en aplicar el criterio de los mínimos cuadrados al modelo linealizado obtenido a partir del modelo potencial sin más que tomar logaritmos,

$$\ln y = \ln a + b \cdot \ln x,$$

ya que, haciendo los cambios de variable

$$V = \ln X$$

y

$$U = \ln Y,$$

y denotando

$$c = \ln a,$$

resulta el modelo lineal

$$v = c + b \cdot u,$$

cuyos parámetros c y b se calculan mediante expresiones conocidas:

$$b = \frac{S_{U,V}}{S_U^2}$$

y

$$c = \overline{v} - \frac{S_{U,V}}{S_U^2} \cdot \overline{u}.$$

La siguiente tabla servirá para la obtención de los momentos no centrales y centrales. En las casillas de la última fila —marcadas en negrita— aparecen las sumas de los elementos de cada una de las columnas.

x_i	y_i	$v_i = \ln y_i$	$u_i = \ln x_i$	$u_i \cdot v_i$	u_i^2	v_i^2
30	27	3,2958	3,4012	11,2097	11,5682	10,8623
40	60	4,0943	3,6889	15,1035	13,6080	16,7633
50	120	4,7875	3,9120	18,7287	15,3037	22,9202
60	200	5,2983	4,0943	21,6928	16,7633	28,0720
70	350	5,8579	4,2485	24,8873	18,0498	34,3150
75	500	6,2146	4,3175	26,8315	18,6408	38,6213
80	510	6,2344	4,3820	27,3191	19,2019	38,8677
85	610	6,4135	4,4427	28,4933	19,7376	41,1330
90	740	6,6067	4,4998	29,7288	20,2482	43,6485
95	900	6,8024	4,5539	30,9774	20,7380	46,2726
675	**4 017**	**55,6054**	**41,5408**	**234,9721**	**173,8595**	**321,4759**

Tenemos, así, que los valores medios son

$$\overline{u} = \frac{1}{N}\sum_{i=1}^{N} u_i = \frac{41{,}5408}{10} = 4{,}15408$$

y

$$\overline{v} = \frac{1}{N}\sum_{i=1}^{N} v_i = \frac{55{,}6054}{10} = 5{,}56054.$$

Además, la varianza de U y la covarianza entre U y V son, respectivamente,

$$S_U^2 = \frac{1}{N}\sum_{i=1}^{N} u_i^2 - \overline{u}^2 = \frac{173{,}8595}{10} - 4{,}15408^2 = 0{,}1296$$

y

$$S_{U,V} = \frac{1}{N}\sum_{i=1}^{N} u_i \cdot v_i - \overline{u} \cdot \overline{v} = \frac{234{,}9721}{10} - 4{,}15408 \cdot 5{,}56054 = 0{,}3983.$$

En definitiva,

$$b = \frac{S_{U,V}}{S_U^2} = \frac{0{,}3983}{0{,}1296} = 3{,}07$$

y

$$c = \overline{v} - \frac{S_{U,V}}{S_U^2} \cdot \overline{u} = 5{,}56054 - \frac{0{,}3989}{0{,}1296} \cdot 4{,}15408 = -\,7{,}23,$$

con lo cual, despejando, se tiene que

$$a = \exp(c) = 0{,}0007.$$

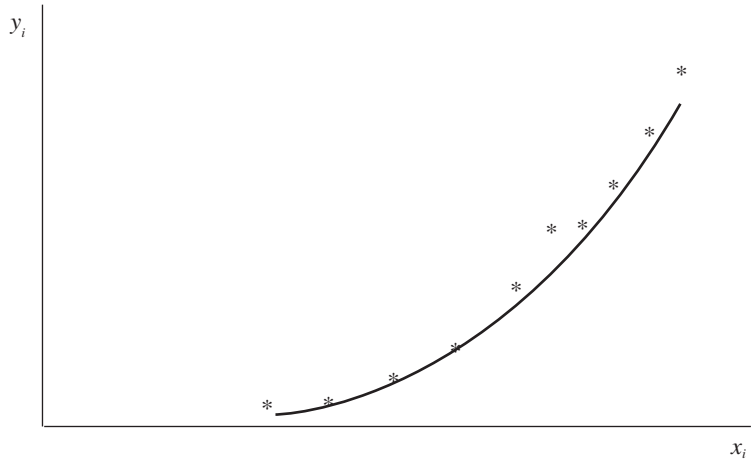
Por consiguiente, el modelo es

$$y = 0{,}0007 \cdot x^{3{,}07}.$$

c) En la tercera columna de la tabla figuran los valores de la variable $\tilde{Y}$ obtenidos por la regresión efectuada. De este modo, por ejemplo, $\tilde{y}_3 = 115{,}06$ se ha hallado a partir de $x_3 = 50$ como $0{,}0007 \cdot 50^{3{,}07}$.

x_i	y_i	$\tilde{y}_i$
30	27	23,98
40	60	58,00
50	120	115,06
60	200	201,38
70	350	323,26
75	500	399,52
80	510	487,06
85	610	586,70
90	740	699,23
95	900	825,49
675	**4 017**	**3 719,68**

Las representaciones de los pares de puntos, $(x_i, \tilde{y}_i)$, y de la nube de puntos, (x_i, y_i), aparecen en la gráfica siguiente:



Para medir la bondad del ajuste realizado, utilizaremos, como en el problema anterior, el cociente

$$\frac{S_e^2}{S_Y^2},$$

proporción que la varianza residual representa sobre la varianza de la variable Y.

Completamos la tabla con las columnas que permitirán hallar las varianzas de Y y de e; en concreto, la penúltima columna corresponde a los valores de los residuos:

$$e_i = y_i - \tilde{y}_i.$$

x_i	y_i	$\tilde{y}_i$	y_i^2	e_i	e_i^2
30	27	23,98	729	3,02	9,1204
40	60	58,00	3600	2,00	4,0000
50	120	115,06	14 400	4,94	24,4036
60	200	201,38	40 000	–1,38	1,9044
70	350	323,26	122 500	26,74	715,0276
75	500	399,52	250 000	100,48	10 096,2304
80	510	487,06	260 100	22,94	526,2436
85	610	586,70	372 100	23,30	542,8900
90	740	699,23	547 600	40,77	1 662,1929
95	900	825,49	810 000	74,51	5 551,7401
675	**4 017**	**3 719,68**	**2 421 029**	**297,32**	**19 133,7530**

A partir de estos datos se obtiene que

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{4\,017}{10} = 401{,}7 \text{ euros}$$

y

$$a_{0,2} = \frac{1}{N}\sum_{i=1}^{N} y_i^2 = \frac{2\,421\,029}{10} = 242\,102{,}9,$$

con lo cual,

$$S_Y^2 = a_{0,2} - \overline{y}^2 = 242\,102{,}9 - 401{,}7^2 = 80\,740{,}01.$$

De igual modo, para el cálculo de la varianza residual,

$$S_e^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 - \bar{e}^2,$$

se tiene, por un lado,

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{297{,}32}{10} = 29{,}732,$$

y, por otro lado,

$$\frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{19\ 133{,}753}{10} = 1\ 913{,}38,$$

siendo, en consecuencia, la varianza de e igual a

$$S_e^2 = 1\ 913{,}38 - 29{,}732^2 = 1\ 029{,}39.$$

En definitiva,

$$\frac{S_e^2}{S_Y^2} = \frac{1\ 029{,}39}{80\ 740{,}01} = 0{,}0127,$$

representando la varianza residual el 1,27 por ciento de la variabilidad de Y.

Capítulo 3

Análisis de atributos

Principales conceptos y resultados

Si la característica objeto de estudio en las unidades de una población es cualitativa, es decir, no numérica, se denomina **atributo**. Las observaciones distintas de un atributo, A, son sus **modalidades**, que se denotan por A_1, ..., A_h.

Se llama **frecuencia absoluta** de una modalidad de un atributo al número de observaciones iguales a dicha modalidad, denotándose por n_i la frecuencia absoluta genérica de la modalidad A_i. Si N es el número de observaciones, se cumple que

$$\sum_{i=1}^{h} n_i = N.$$

La **frecuencia relativa** de una modalidad de un atributo es la proporción de observaciones iguales a dicha modalidad, siendo f_i la frecuencia relativa genérica. Puesto que

$$f_i = \frac{n_i}{N},$$

se cumple, entonces, que

$$\sum_{i=1}^{h} f_i = 1.$$

Se denomina **distribución de frecuencias** del atributo A al conjunto de modalidades con sus correspondientes frecuencias, absolutas o relativas, y se denota por $(A_i; n_i)$ o bien $(A_i; f_i)$.

Al igual que ocurre con las distribuciones de frecuencias de variables, si todas las frecuencias absolutas son iguales a la unidad, la distribución de frecuencias de un atributo es una distribución de frecuencias **unitaria**.

Las representaciones gráficas más habituales de la distribución de frecuencias de un atributo son el **diagrama de barras** y el **diagrama de sectores**. Para representar un diagrama de barras se marcan segmentos sobre el eje de abscisas correspondientes a cada modalidad del atributo, elevando sobre ellos barras cuyas longitudes son iguales a las frecuencias absolutas o relativas. El diagrama de sectores es un círculo, dividido en sectores, siendo sus áreas proporcionales a las frecuencias absolutas o relativas.

Dado el carácter no cuantitativo de los atributos, no es posible obtener medidas numéricas que resuman la información proporcionada por los datos. Escasas son las excepciones, como es el caso de la **moda**, modalidad con mayor frecuencia. Cuando las modalidades de un atributo admitan una ordenación por el grado de intensidad de la característica, es posible calcular también la **mediana** de la distribución, modalidad que tiene el mismo número de observaciones «mayores» y «menores» que ella.

La observación conjunta de dos atributos, A y B, en las unidades de una población lleva a la obtención de pares de datos cuyas componentes son cualitativas, siendo (A_i, B_j) la modalidad genérica de (A, B).

La **frecuencia absoluta** de (A_i, B_j), o **frecuencia absoluta conjunta**, es el número de veces que aparecen simultáneamente A_i y B_j en las unidades de la población y se denota por n_{ij}. Si $A_1, ..., A_h$ son las modalidades del atributo A y $B_1, ..., B_k$ las modalidades del atributo B, entonces,

$$\sum_{i=1}^{h} \sum_{j=1}^{k} n_{ij} = N.$$

La **frecuencia relativa** de (A_i, B_j), o **frecuencia relativa conjunta**, f_{ij}, es la proporción de observaciones iguales a dicho par.

La **distribución de frecuencias bidimensional** correspondiente a (A, B) es el conjunto de pares de modalidades, junto con sus frecuencias. Utilizaremos indistintamente la notación $(A_i, B_j; n_{ij})$, o $(A_i, B_j; f_{ij})$ con frecuencias absolutas o relativas.

La disposición más frecuente de una distribución bidimensional de atributos es una tabla de doble entrada denominada **tabla de contingencia**, que, al igual que las tablas de correlación en el caso de variables, contiene en su interior las frecuencias conjuntas.

A \ B	B_1	...	B_j	...	B_k
A_1	n_{11}	...	n_{1j}	...	n_{1k}
⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	...	n_{ij}	...	n_{ik}
⋮	⋮	⋮	⋮	⋮	⋮
A_h	n_{h1}	...	n_{hj}	...	n_{hk}

Partiendo de la distribución de frecuencias bidimensional pueden obtenerse las **distribuciones marginales** de los atributos A y B, $(A_i; n_{i\cdot})$ y $(B_j; n_{\cdot j})$, respectivamente, donde

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$$

y

$$n_{\cdot j} = \sum_{i=1}^{h} n_{ij}$$

son las frecuencias marginales genéricas.

También es posible calcular distribuciones condicionadas partiendo de la distribución bidimensional. Así, la **distribución del atributo *A* condicionada por la modalidad B_j del atributo *B*** es $(A_i/B_j; n_{i/j})$, conforme se recoge en la siguiente tabla:

A_i/B_j	$n_{i/j}$
A_1	n_{1j}
⋮	⋮
A_i	n_{ij}
⋮	⋮
A_h	n_{hj}

De igual modo, la **distribución del atributo *B* condicionada por la modalidad A_i del atributo *A*** se denota por $(B_j/A_i; n_{j/i})$, donde

B_j/A_i	$n_{j/i}$
B_1	n_{i1}
⋮	⋮
B_j	n_{ij}
⋮	⋮
B_k	n_{ik}

A partir de las **frecuencias absolutas condicionadas** se calculan las **frecuencias relativas condicionadas**, según las relaciones genéricas:

$$f_{i/j} = \frac{n_{i/j}}{n_{\cdot j}}$$

y

$$f_{j/i} = \frac{n_{j/i}}{n_{i\cdot}}.$$

Dada una distribución de frecuencias bidimensional $(A_i, B_j; f_{ij})$, los atributos A y B son independientes, si, para cualesquiera i y j,

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j},$$

o, lo que es igual,

$$n_{ij} = \frac{n_i \cdot n_{\cdot j}}{N},$$

para cualesquiera i y j, condición que denominaremos **condición de independencia** entre las modalidades A_i y B_j.

Cuando la tabla de contingencia es de dimensión 2×2, es suficiente comprobar la condición de independencia con una pareja de modalidades.

La condición anterior es equivalente a

$$f_{i/j} = f_{i\cdot}$$

y

$$f_{j/i} = f_{\cdot j},$$

para cualesquiera i y j, es decir, que la condición necesaria y suficiente para que dos atributos sean independientes es que las frecuencias relativas condicionadas sean idénticas a sus respectivas frecuencias relativas marginales.

Cuando los atributos no son independientes, se habla de **tipo de asociación** entre sus modalidades. Así, si

$$n_{ij} > \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

esto es, si la frecuencia absoluta conjunta entre las modalidades A_i y B_j es mayor que la que existiría en el caso de que los atributos A y B fuesen independientes, se dice que entre las modalidades A_i y B_j existe **asociación positiva**.

Recíprocamente, si

$$n_{ij} < \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

entonces, entre las modalidades A_i y B_j hay **asociación negativa**.

En las tablas de contingencia de dimensión 2×2, cuando los atributos no son independientes, el estudio del tipo de asociación de un par de modalidades de la tabla determina el tipo de asociación del resto de los pares. Si, por ejemplo,

$$n_{12} > \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N},$$

es decir, si entre A_1 y B_2 la asociación es positiva, también será positiva la asociación entre A_2 y B_1, dándose, en cambio, asociación negativa entre A_1 y B_1 y entre A_2 y B_2.

Este hecho justifica la utilización del **coeficiente de asociación**[1]

$$\mathrm{H} = n_{11} - \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

cuya interpretación es la siguiente:

- Si H es cero, se cumple la condición de independencia para las modalidades A_1 y B_1 y, al ser una tabla de dimensión 2×2, ello implica que los atributos son independientes.
- Si $\mathrm{H} > 0$, entonces los atributos no son independientes, habiendo asociación positiva entre las modalidades A_1 y B_1 y entre A_2 y B_2, y asociación negativa entre A_1 y B_2 y entre A_2 y B_1.
- Si $\mathrm{H} < 0$, los atributos no son independientes y entre A_1 y B_1 y entre A_2 y B_2 hay asociación negativa, y asociación positiva entre A_1 y B_2 y entre A_2 y B_1.

El estudio de la independencia entre atributos cuando las tablas de contingencia son de dimensión $h \times k$ requiere la comprobación de la condición de independencia con todos los pares de modalidades.

Cuando los atributos no son independientes, es posible dar una medida de su *grado de asociación*, utilizando el coeficiente χ^2 **de Pearson**,

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{ij} - \dfrac{n_{i\cdot} \cdot n_{\cdot j}}{N}\right)^2}{\dfrac{n_{i\cdot} \cdot n_{\cdot j}}{N}},$$

que, según se observa, compara cada frecuencia absoluta conjunta, n_{ij}, con las frecuencias absolutas *teóricas* que corresponderían en el caso de que existiera independencia, $n_{i\cdot} \cdot n_{\cdot j} / N$. Así, si este coeficiente es cero, los atributos son independientes, siendo mayor el grado de asociación cuanto mayor sea su valor[2].

Cuando los atributos que se analizan admiten una ordenación de sus modalidades como consecuencia de la mayor o menor intensidad con la que se presenta la característica[3], es posible estudiar el grado de asociación que existe entre ellos, mediante medidas más precisas.

Para ello, definimos dos variables X e Y, con valores iguales a los respectivos rangos o números de orden de las modalidades de los atributos A y B. De este modo, de la distribución de frecuencias $(A_i, B_j; n_{ij})$, resulta la distribución de frecuencias $(x_i, y_j; n_{ij})$, a partir de la cual es posible calcular el coeficiente de correlación lineal, cuyo valor será indicativo del grado de asociación entre las intensidades de los atributos: un valor positivo y próximo a 1 del coeficiente

---

[1] Este coeficiente puede definirse a partir de una pareja cualquiera de modalidades, siendo su interpretación análoga a la realizada en el texto.

[2] De este coeficiente derivan otros, entre los que destacamos el coeficiente de contingencia de Pearson, $C = \sqrt{\chi^2 / N + \chi^2}$, coeficiente acotado entre 0 y 1.

[3] El coeficiente del que hablamos a continuación se utiliza para analizar el grado de asociación entre dos características cuyos estados admiten una ordenación por rangos; pudiendo ser dichas características numéricas, es decir, variables.

denotará una gran asociación creciente (positiva) entre las intensidades y, recíprocamente, un valor negativo y cercano a -1 será indicativo de una asociación decreciente (negativa) y elevada entre las intensidades.

Una situación particular del análisis descrito surge cuando se dispone de N unidades clasificadas según el rango o posición que tienen en relación a dos atributos[4], A y B, y les asociamos dos variables X e Y, cuyos valores son los rangos de A y B, respectivamente. Tendremos, en este caso, pares de observaciones (x_1, y_1), ..., (x_i, y_i), ..., (x_N, y_N), donde (x_i, y_i) son los rangos que tiene la unidad i-ésima, con respecto a los atributos A y B.

Sobre las variables X e Y se realiza un análisis de correlación, a partir del coeficiente de correlación lineal visto en el capítulo 2, que servirá para estudiar la *concordancia* o *discordancia* entre las ordenaciones de las unidades de la población según los rangos de los dos atributos. Una correlación positiva y alta entre ambas variables es indicativa de una fuerte concordancia entre las ordenaciones según los rangos de los dos atributos; y recíprocamente, una elevada correlación negativa nos hace pensar en una fuerte discordancia entre las ordenaciones. El coeficiente de correlación lineal en esta situación se denomina **coeficiente de rangos de Spearman** y adopta la expresión

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N^3 - N},$$

donde $d_i = x_i - y_i$ es la diferencia genérica entre los rangos de ambos atributos.

Otro procedimiento para el análisis de la concordancia o discordancia entre los rangos de dos atributos consiste en ordenar las unidades según el orden natural de los rangos de uno de los dos atributos, por ejemplo del primero, obteniéndose pares de observaciones $(1, y_1)$, ..., (i, y_i), ..., (N, y_N). Cuanto más próxima esté la ordenación y_1, ..., y_i, ..., y_N al orden natural 1, ..., i, ..., N que tienen los rangos del primer atributo, mayor la concordancia entre ambas ordenaciones y, viceversa, cuanto más próxima esté dicha ordenación a la ordenación inversa del orden natural, N, ..., i, ..., 1, mayor la discordancia. Este análisis se sintetiza mediante el **coeficiente τ de Kendall**:

$$\tau = \frac{2 \sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j)}{N(N-1)},$$

donde $\delta(y_i, y_j)$ tiene el valor 1 si entre y_i e y_j se sigue el orden natural, es decir, si $y_i < y_j$, y vale -1, en caso contrario.

Este coeficiente está acotado entre -1 y 1, tomando el valor 1 cuando la *concordancia es perfecta* y el valor -1 cuando existe una *perfecta discordancia*.

---

[4] Los coeficientes seguidamente descritos analizan el grado de concordancia entre dos ordenaciones, pudiendo provenir éstas del estudio de características numéricas, esto es, de variables, sobre las unidades de la población.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

3.1 Una encuesta realizada por la revista *El Mensual* sobre el *ranking* que ocupa un grupo de suplementos de fin de semana arroja los siguientes resultados: *El Mensual* alcanza 6 millones de lectores, siendo 4,5, 1,8, 1,5 y 1,2 millones los lectores de los dominicales *Magazine*, *La Semana*, *Comunidad y Tierra*, respectivamente.

a) Preséntese en una tabla la distribución de frecuencias del enunciado.

b) ¿Qué porcentaje de lectores lee *La Semana?* ¿Y el dominical *Tierra?*

c) Represéntese gráficamente, mediante diagrama de barras y diagrama de sectores, la distribución de frecuencias.

SOLUCIÓN

a) En la primera columna de la tabla colocamos las modalidades de la característica *suplemento que se lee el fin de semana*, *A*, esto es, las revistas del grupo considerado, dejando la segunda columna para las frecuencias absolutas de cada una de dichas modalidades, es decir, el número de lectores que lee cada una de las revistas. En la tabla siguiente se presenta, por tanto, la distribución de frecuencias $(A_i; n_i)$, donde A_i es la modalidad genérica y n_i su correspondiente frecuencia absoluta.

Suplemento	N.º lectores
El Mensual	6,0
Magazine	4,5
La Semana	1,8
Comunidad	1,5
Tierra	1,2

De este modo, de la consulta de la tabla se desprende que, por ejemplo, $n_4 = 1{,}5$ indica que la revista *Comunidad,* A_4, tiene 1,5 millones de lectores, mientras que $n_1 = 6$ muestra que son 6 los millones de personas que leen *El Mensual,* A_1.

La suma de los elementos de la segunda columna de la tabla, suma de las frecuencias absolutas de las modalidades del atributo, es $N = 15$, por tanto, 15 millones es el número de unidades de la población.

b) A partir de los datos de la tabla anterior, aplicando la expresión genérica:

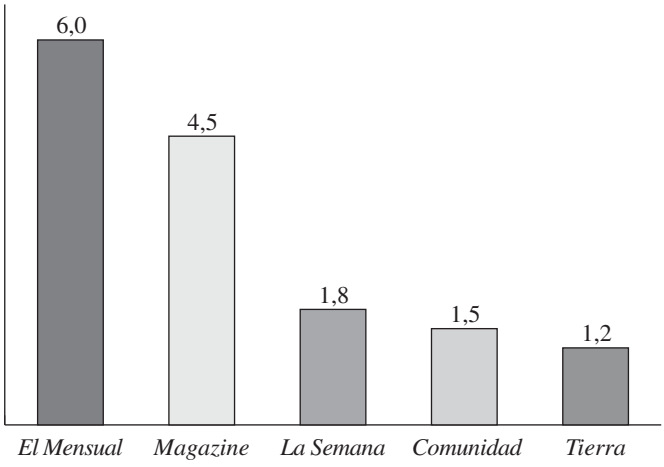
$$f_i = \frac{n_i}{N},$$

se obtiene una nueva columna de frecuencias relativas de cada una de las modalidades:

Suplemento	N.º lectores	Proporción lectores
El Mensual	6,0	0,40
Magazine	4,5	0,30
La Semana	1,8	0,12
Comunidad	1,5	0,10
Tierra	1,2	0,08

Puesto que $f_3 = 0{,}12$ es la frecuencia relativa de la modalidad A_3, *La Semana,* el 12 por ciento de los lectores leen este dominical. De igual modo se concluye que el 8 por ciento de los lectores se decantan por *Tierra,* puesto que la frecuencia relativa de esta modalidad es $f_5 = 0{,}08$.

c) Para representar con un diagrama de barras la distribución de frecuencias, colocamos en el eje de abscisas cinco segmentos de igual longitud, uno para cada una de las modalidades, en este caso revistas, del atributo considerado, elevando sobre cada segmento una barra cuya longitud es igual a la frecuencia absoluta de cada modalidad:



La siguiente representación gráfica es un diagrama de sectores, es decir, un círculo dividido en sectores, siendo el área de cada uno de ellos proporcional a la frecuencia de la respectiva modalidad. Para calcular el área de los sectores hay que considerar que el hecho de que dichas áreas

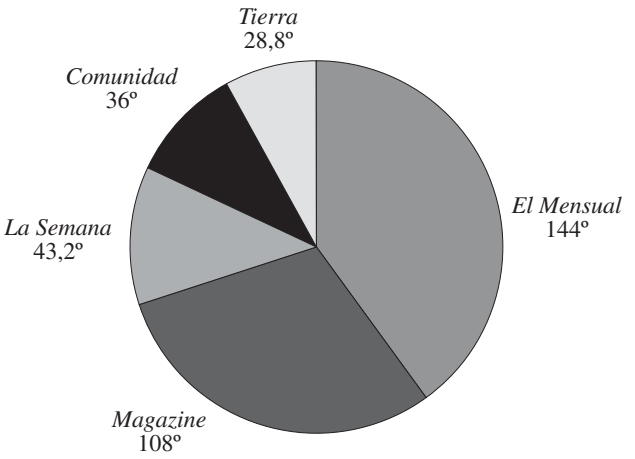
sean proporcionales a las frecuencias es equivalente a que sus ángulos lo sean. Ahora bien, si α_i es el ángulo del sector de la modalidad genérica A_i con frecuencia absoluta n_i, entonces, puesto que el ángulo de todo el círculo es igual a 360 grados, necesariamente ha de cumplirse la relación de proporcionalidad:

$$\frac{360}{N} = \frac{\alpha_i}{n_i},$$

de lo cual,

$$\alpha_i = 360 \cdot \frac{n_i}{N} = 360 \cdot f_i.$$

Así, por ejemplo, el ángulo α_2 de la modalidad, A_2, *El Magazine*, es igual a $360 \cdot 0{,}3 = 108$ grados, lo cual equivale a que el 30 por ciento del área del círculo corresponde a esta modalidad.



3.2 Una empresa dedicada a la construcción de muebles de diseño cuenta con 200 trabajadores de los cuales 100 pertenecen a la sección de carpintería, 20 a la de transporte, 50 trabajadores son de la sección de administración y el resto es personal de dirección.

a) ¿Cuál es la población objeto de estudio? ¿De cuántas unidades consta? ¿Qué tipo de característica se analiza en ella?

b) ¿Cuál es la distribución de frecuencias de la característica analizada? ¿Qué peso tiene cada sección en el conjunto de la empresa?

c) ¿En qué sección hay mayor número de trabajadores?

d) Represéntese gráficamente la distribución de frecuencias obtenida en el apartado ***b)***.

SOLUCIÓN

a) La población que se estudia está formada por los 200 trabajadores de la empresa de construcción de muebles, sobre la que se analiza la característica *sección a la que pertenece cada trabajador*. Esta característica no es numérica, pues sus distintos estados no son cuantificables; se trata, por tanto, de un atributo cuyas modalidades son: *sección de carpintería*, *sección de transporte* y *sección de administración* y *dirección*.

b) La distribución de frecuencias del atributo está formada por sus modalidades —primera columna de la tabla siguiente—, junto con sus correspondientes frecuencias; en este caso, el enunciado proporciona las frecuencias absolutas de cada modalidad, según se recoge en la segunda columna de la tabla:

Secciones	N.º trabajadores	% trabajadores
Carpintería	100	50
Transporte	20	10
Administración	50	25
Dirección	30	15

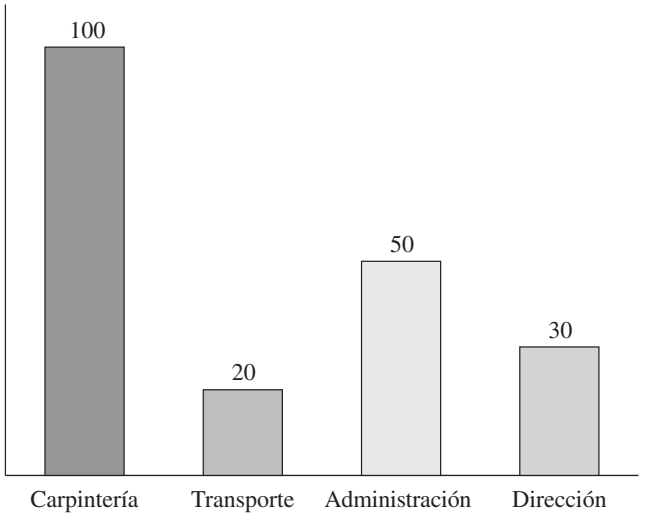
Así, puesto que 50 trabajadores de la empresa pertenecen a la sección de administración, entonces, $n_3 = 50$.

En la tercera columna aparece el peso de cada sector en el conjunto de la empresa, es decir, la frecuencia relativa de cada modalidad del atributo, obtenida como cociente entre la frecuencia absoluta y el número total de datos. Por ejemplo, $f_4 = 30/200 = 0{,}15$, con lo cual, el 15 por ciento del total de la empresa es personal de dirección.

c) La sección de carpintería, con 100 empleados, es la sección con mayor número de trabajadores; por tanto, esta modalidad es la moda de la distribución.

d) Representaremos la distribución de trabajadores por secciones, esto es, la distribución de frecuencias del atributo considerado, utilizando un diagrama de barras y un diagrama de sectores.

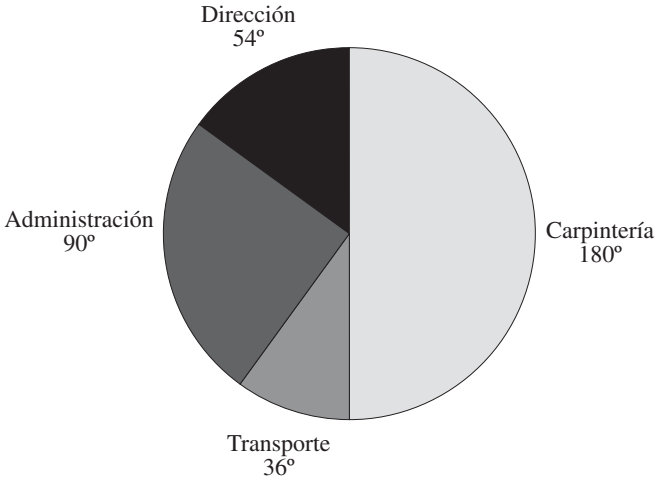
Para obtener el diagrama de barras, colocamos en el eje de abscisas segmentos de igual medida para cada modalidad del atributo; en esta ocasión, para cada sección de la empresa. A continuación, sobre cada uno de los segmentos elevamos una barra cuya altura sea igual a la correspondiente frecuencia, absoluta o relativa. Si consideramos frecuencias absolutas, se obtiene la siguiente representación gráfica.



En cuanto al diagrama de sectores, hemos de calcular el área de cada sector, teniendo en cuenta que, como sabemos, esto es equivalente a obtener la medida de su ángulo, según la expresión genérica demostrada en el problema anterior:

$$\alpha_i = 360 \cdot f_i.$$

Si se aplica esta igualdad a cada una de las modalidades, resulta la siguiente representación gráfica:



3.3 A partir de un estudio realizado sobre el plan de formación bianual 2003-2004 para funcionarios, se ha conocido que la asistencia a cursos formativos, según los distintos grupos (categorías), fue la que se presenta a continuación:

Grupo	Asistentes 2003	Asistentes 2004
Grupo A	38 704	39 704
Grupo B	51 782	53 782
Grupo C	57 007	57 507
Grupo D	87 053	90 053
Grupo E	13 990	13 790
TOTAL	248 536	254 836

a) ¿Cuál es la población que se analiza? ¿De cuántas unidades consta? ¿Cuál es la característica estudiada? Obténgase la correspondiente distribución de frecuencias.

b) Represéntese gráficamente la información de la tabla, separadamente, año por año, y de modo conjunto.

SOLUCIÓN

a) En este problema se plantean dos poblaciones de funcionarios que se presentan a cursos formativos: una de 248 536 unidades para el año 2003 y otra de 254 836 unidades para el año 2004. Sobre cada población se analiza la misma característica cualitativa o atributo, *categoría o grupo al que pertenece el funcionario*, cuyas modalidades son: *grupo A*, *grupo B*, *grupo C*, *grupo D* y *grupo E*.

Tendremos, por tanto, dos distribuciones de frecuencias:

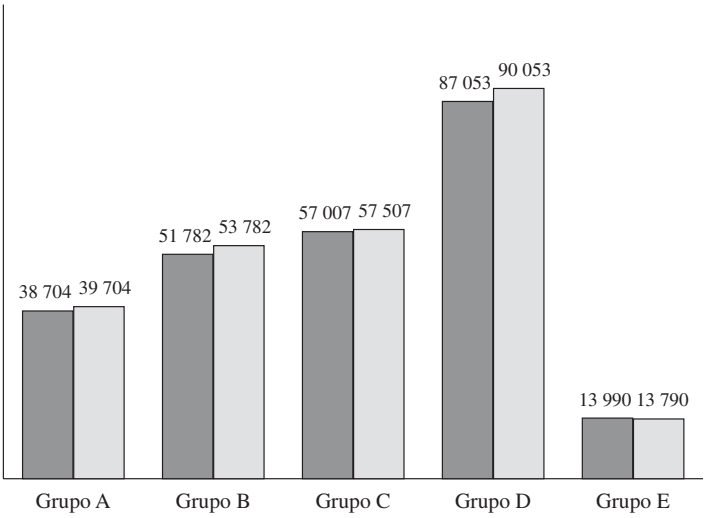
Categoría	Asistentes
Grupo A	38 704
Grupo B	51 782
Grupo C	57 007
Grupo D	87 053
Grupo E	13 900

para el año 2003, y, para el año 2004:

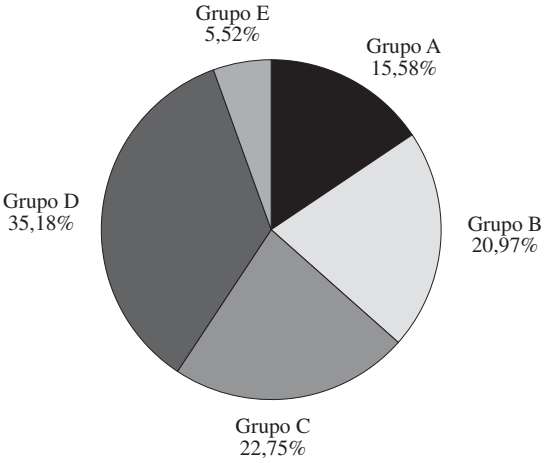
Categoría	Asistentes
Grupo A	39 704
Grupo B	53 782
Grupo C	57 507
Grupo D	90 053
Grupo E	13 790

Donde, por ejemplo, $n_3 = 57\,007$, en la primera tabla, indica que hubo 57 007 funcionarios del grupo C que asistieron a cursos de formación en 2003, y $n_1 = 39\,704$, en la segunda tabla, expresa que en 2004 fueron 39 704 los funcionarios del grupo A que se presentaron a cursos formativos.

b) En la siguiente gráfica aparecen dos diagramas de barras que se corresponden con cada una de las distribuciones. Para cada categoría profesional o modalidad del atributo se ha marcado un segmento doble sobre el cual se ha elevado una doble barra con alturas iguales a las frecuencias absolutas de cada distribución para dicha modalidad. Este análisis *conjunto* permite la comparación gráfica de las situaciones de ambos años.



Y, por último, con un diagrama sectorial representamos el número total de asistentes en ese período:



Así, por ejemplo, el 6 por ciento correspondiente al grupo E se ha obtenido como

$$\frac{13\,990 + 13\,790}{248\,536 + 254\,836} \cdot 100.$$

3.4 Una empresa con centros comerciales distribuidos por todo el territorio nacional pretende abrir nuevos mercados en el sur de Francia. Con objeto de seleccionar personal, realiza encuestas a los futuros trabajadores, presentándose a las pruebas 1 000 personas, cuyas titulaciones se reflejan en la siguiente tabla:

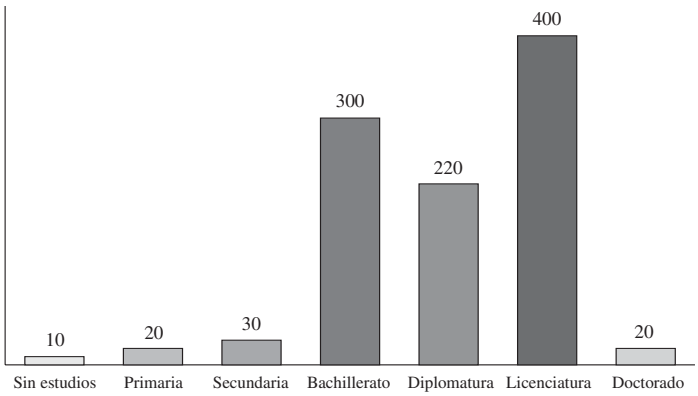
Nivel de titulación	N.º aspirantes
Sin estudios	10
Primaria	20
Secundaria	30
Bachillerato	300
Diplomatura	220
Licenciatura	400
Doctorado	20

a) Represéntese gráficamente la distribución de frecuencias.

b) Calcúlese la mediana de la distribución.

SOLUCIÓN

a) Las modalidades de la característica *nivel de titulación* admiten la ordenación que aparece en la primera columna de la tabla correspondiente a la distribución de frecuencias. Por esta razón, a la hora de representar gráficamente dicha distribución con un diagrama de barras, colocaremos las modalidades también de modo ordenado en los segmentos marcados sobre el eje de abscisas.



b) Obtendremos, en primer lugar, las frecuencias acumuladas que aparecen en la tercera columna de la tabla siguiente. Obsérvese que, aunque nuestra característica es un atributo, el hecho de que admita una ordenación de sus modalidades permite el cálculo de este tipo de frecuencias, mediante las expresiones:

$$N_1 = n_1 \text{ y } N_i = n_1 + ... + n_i, \text{ para } i = 2, ..., h.$$

Nivel de titulación	n_i	N_i
Sin estudios	10	10
Primaria	20	30
Secundaria	30	60
Bachillerato	300	360
Diplomatura	220	580
Licenciatura	400	980
Doctorado	20	1 000

Por ejemplo, $N_4 = 360$ significa que hay 360 candidatos que tienen a lo sumo estudios de Bachillerato.

Proponemos al lector la obtención de las frecuencias relativas acumuladas a partir de la definición que de esta clase de frecuencias se dio en el capítulo 1.

A continuación, siguiendo con el cálculo de la mediana, hallamos, al igual que en el análisis de variables correspondiente al capítulo 1, la cantidad $N/2$, que en este caso es igual a 500. Puesto que no existe una modalidad cuya frecuencia absoluta acumulada, N_i, sea igual a 500[1], la mediana es aquella modalidad tal que su frecuencia absoluta acumulada es estrictamente mayor que 500, es decir, el nivel de titulación de *diplomado* con $N_5 = 580$.

3.5 Finalizada la campaña de Navidad, la asociación de productores de cava de la pequeña región de Arautiol pretende hacer un estudio sobre los hábitos de consumo de esta bebida en la comarca donde sus empresas distribuyen el producto. En el estudio tienen en cuenta dos zonas totalmente diferenciadas (zona norte y zona sur), por considerar que la orografía del terreno hace que las costumbres de ambas sean distintas.

La siguiente tabla refleja el total de litros de cava (brut, seco y semiseco) vendidos en la última temporada navideña, diferenciando las ventas en las zonas norte y sur.

[1] Si hubiera existido una modalidad, A_i, cuya frecuencia absoluta acumulada, N_i, fuera igual a 500, habría dos medianas: las modalidades A_i y A_{i+1}.

Zona / Cava	Norte	Sur
Brut	1 000	2 000
Seco	600	500
Semiseco	400	1 500

a) Obténganse las distribuciones de frecuencias marginales.

b) Hállense las distribuciones condicionadas.

SOLUCIÓN

a) La tabla de contingencia del enunciado corresponde a la distribución bidimensional (A_i, B_j; n_{ij}) de los atributos, *A*, *tipo de cava*, cuyas modalidades son *brut*, *seco* y *semiseco* y *B*, *zona de venta*, con modalidades *norte* y *sur*; en las casillas de esta tabla se encuentran las frecuencias conjuntas, n_{ij}. Así, por ejemplo, el número de litros vendidos de cava semiseco en la zona sur es $n_{32} = 1\,500$.

A partir de dicha tabla, se calculan las distribuciones de frecuencias marginales de cada atributo (A_i; $n_{i\cdot}$) y (B_j; $n_{\cdot j}$), teniendo en cuenta las relaciones entre frecuencias marginales y conjuntas, para todo *i*, y, para cualquier *j*,

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij},$$

y

$$n_{\cdot j} = \sum_{i=1}^{h} n_{ij}.$$

De este modo, sumando cada elemento de la primera columna con su correspondiente elemento de la segunda, se hallan las frecuencias del atributo *A*, *tipo de cava*, según se recoge en la última columna de la siguiente tabla. Se observa, por ejemplo, que el número de litros de cava del tipo seco vendidos en la última temporada, es decir, $n_{2\cdot} = 1\,100$, se obtiene como suma de los litros vendidos de este tipo en la zonas norte y sur, esto es, $n_{21} + n_{22} = 600 + 500$.

Zona / Cava	Norte	Sur	$n_{i\cdot}$
Brut	1 000	2 000	3 000
Seco	600	500	1 100
Semiseco	400	1 500	1 900
$n_{\cdot j}$	2 000	4 000	6 000

Análogamente, las frecuencias de la distribución marginal (B_j; $n_{\cdot j}$) se calculan sumando, casilla a casilla, las filas de la tabla de contingencia, obteniéndose los datos de la última fila de la tabla anterior. De este modo, por ejemplo, el número de litros de cava vendidos en la zona sur, $n_{\cdot 2} = 4\,000$, es igual al total de litros que en esta zona se han vendido de cada uno de los tipos de cava, $n_{12} + n_{22} + n_{32} = 2\,000 + 500 + 1\,500$.

En definitiva, la distribución de frecuencias marginal del primer atributo, es decir, la distribución de ventas según el tipo de cava, es la siguiente:

Cava	$n_{i\cdot}$	$f_{i\cdot}$
Brut	3 000	0,500
Seco	1 100	0,183
Semiseco	1 900	0,317

Como puede observarse, en la última columna de la tabla figuran las frecuencias relativas de cada modalidad del atributo obtenidas según la relación genérica:

$$f_{i\cdot} = \frac{n_{i\cdot}}{N}.$$

De igual manera la distribución de frecuencias marginal del segundo atributo, esto es, la distribución de ventas de cava según la zona es la que se recoge a continuación:

Zona	$n_{\cdot j}$	$f_{\cdot j}$
Norte	2 000	0,33
Sur	4 000	0,67

Nótese que la tercera columna de la tabla anterior de frecuencias relativas de las modalidades, se ha calculado según la expresión genérica:

$$f_{\cdot j} = \frac{n_{\cdot j}}{N}.$$

b) A partir de la distribución bidimensional (A_i, B_j; n_{ij}), se hallan las distribuciones del atributo A condicionadas por cada modalidad B_j del atributo B. Así, la distribución de ventas por tipo de cava *dentro* de la zona norte, esto es, (A_i / B_1; $n_{i/1}$), tiene como modalidades, *brut*, *seco* y *semiseco*, siendo sus frecuencias las de la primera columna de la tabla de contingencia:

$$n_{i/1} = n_{i1},$$

para todo i.

En la siguiente tabla se recoge esta distribución de frecuencias unidimensional:

Cava	$n_{i/1}$	$f_{i/1}$
Brut	1 000	0,50
Seco	600	0,30
Semiseco	400	0,20

En la tercera columna de la tabla se incorporan, además, las frecuencias relativas de la distribución unidimensional calculadas según la expresión genérica:

$$f_{i/1} = \frac{n_{i/1}}{n_{\cdot 1}} = \frac{n_{i1}}{n_{\cdot 1}}.$$

Como puede observar el lector, el número de litros de cava de tipo brut vendidos dentro de la zona norte, $n_{1/1} = 1\,000$ coincide con el número de litros de cava que en esta temporada se han vendidos en la zona norte y de tipo brut, esto es, n_{11}; la proporción de litros de cava de este tipo es $f_{1/1} = 1\,000/2\,000 = 0{,}5$.

De igual modo se obtendría la distribución condicionada $(A_i \,/\, B_2;\, n_{i/2})$, es decir, la distribución de ventas por tipo de cava *dentro* de la zona sur:

Cava	$n_{i/2}$	$f_{i/2}$
Brut	2 000	0,500
Seco	500	0,125
Semiseco	1 500	0,375

Para hallar las distribuciones condicionadas del atributo B por cada modalidad A_i del atributo A, es decir $(B_j \,/\, A_i;\, n_{j/i})$, actuamos de idéntica forma. En consecuencia, la distribución de ventas por zonas de cava de tipo brut, $(B_j \,/\, A_1;\, n_{j/1})$ es

Zona	$n_{j/1}$	$f_{j/1}$
Norte	1 000	0,33
Sur	2 000	0,67

donde los elementos de la segunda columna, coinciden con la primera fila de la tabla de contingencia:

$$n_{j/1} = n_{1j},$$

para todo j.

Así, por ejemplo, del total de litros de tipo brut, los que se han vendido en la zona sur, es decir, $n_{2/1}$, coincide con los litros que se han vendido de tipo brut y en la zona sur, esto es, $n_{12} = 2\,000$, siendo la correspondiente frecuencia relativa $f_{2/1} = 2\,000/3\,000 = 0{,}67$.

Dejamos al lector la comprobación de que la distribución de ventas por zonas de cava de tipo seco (B_j / A_2; $n_{j/2}$) y la distribución de ventas por zonas de cava de tipo semiseco (B_j / A_3; $n_{j/3}$) son, respectivamente, las recogidas en las siguientes tablas.

Zona	$n_{j/2}$	$f_{j/2}$
Norte	600	0,55
Sur	500	0,45

Zona	$n_{j/3}$	$f_{j/3}$
Norte	400	0,21
Sur	1 500	0,79

3.6 La asociación de comerciantes de una ciudad realiza una consulta a todos los trabajadores de este sector para conocer sus preferencias respecto al horario de trabajo. En concreto se les plantea si desean realizar o no jornada continua, frente a la opción de jornada partida. El resultado de la encuesta indica que el 70 por ciento de las trabajadoras desea jornada continua, siendo este porcentaje de un 35 por ciento entre los varones.

a) ¿A qué distribuciones de frecuencias corresponden los porcentajes del enunciado?

b) Suponiendo que un 60 por ciento de las personas que trabajan en este sector son mujeres, obténgase el porcentaje de los que trabajan en este sector que prefieren jornada continua.

SOLUCIÓN

a) Sobre la población de trabajadores del sector del comercio se han considerado dos atributos, *A*, *preferencia respecto al horario de trabajo*, y *B*, *sexo de los trabajadores*. Cada uno de estos atributos posee dos modalidades, *jornada continua* y *jornada partida*, y *mujer* y *hombre*, respectivamente.

El enunciado proporciona las frecuencias condicionadas correspondientes a las distribuciones condicionadas del atributo *A*, para cada una de las modalidades del atributo *B*. Así, decir que el 70 por ciento de las trabajadoras desea jornada continua equivale a decir que la frecuencia relativa condicionada de la modalidad *jornada continua*, A_1, dentro de la modalidad *mujer*, B_1,

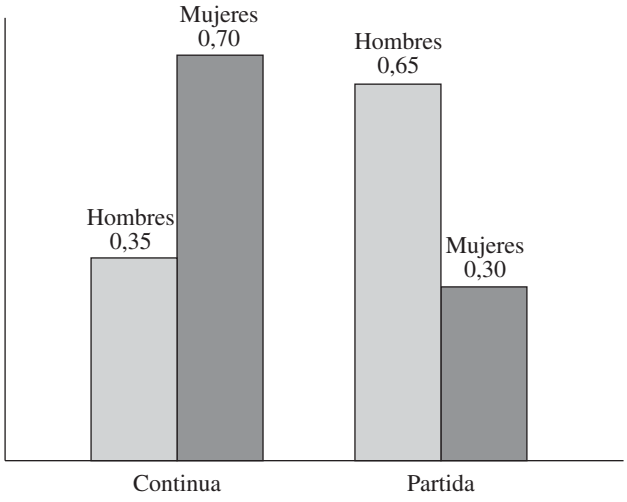
es 0,7, siendo, por tanto, 0,3 la frecuencia relativa de la modalidad *jornada partida*, A_2, dentro de la modalidad *mujer*, B_1.

A_i/B_1	$f_{i/1}$
Continua	0,7
Partida	0,3

Por un razonamiento análogo, 0,35 es la frecuencia relativa condicionada de la modalidad *jornada continua*, A_1, dentro de la modalidad *hombre*, B_2, y, consecuentemente, 0,65 se corresponde con la frecuencia relativa condicionada de la modalidad *jornada partida*, A_2, dentro de la modalidad *hombre*, B_2.

A_i/B_2	$f_{i/2}$
Continua	0,35
Partida	0,65

El siguiente diagrama de rectángulos recoge simultáneamente las dos distribuciones condicionadas anteriores.



b) El 60 por ciento de los que trabajan en el sector comercio son mujeres, es decir, la frecuencia relativa marginal de la modalidad B_1 es $f_{.1} = 0{,}6$.

La cuestión que se plantea en este apartado es el porcentaje de los que trabajan en este sector que prefieren jornada continua, esto es, la frecuencia relativa marginal de la modalidad A_1: $f_{1.}$.

Ahora bien, como, por un lado,

$$f_{1\cdot} = f_{11} + f_{12},$$

y, por otro lado, según se demostró en el capítulo 2 para variables, las frecuencias relativas conjuntas de la expresión anterior pueden calcularse como

$$f_{11} = f_{1/1} \cdot f_{\cdot 1}$$

y

$$f_{12} = f_{1/2} \cdot f_{\cdot 2},$$

entonces, con los datos del problema, se tiene que

$$f_{11} = 0{,}7 \cdot 0{,}6 = 0{,}42$$

y

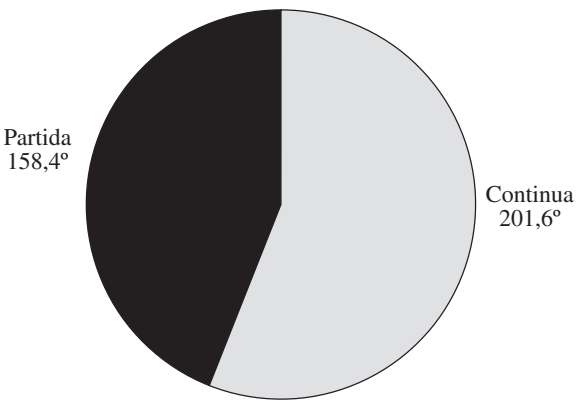
$$f_{12} = 0{,}35 \cdot 0{,}4 = 0{,}14,$$

y, en definitiva, la frecuencia pedida es

$$f_{1\cdot} = 0{,}42 + 0{,}14 = 0{,}56,$$

con lo cual, el 56 por ciento de los trabajadores del sector comercio prefiere jornada continua y el 44 por ciento jornada partida.

Puede comprobar el lector que la representación gráfica, mediante diagrama de sectores, de esta distribución de frecuencias unidimensional, distribución marginal del atributo *A*, es la que figura a continuación:



3.7 Se ha realizado encuesta sobre 1 000 personas para analizar, entre otros aspectos, la posible relación existente entre el medio de transporte utilizado habitualmente para

asistir al trabajo y la clase social a la que se pertenece. Los resultados obtenidos se recogen en la siguiente tabla:

Medio / Clase social	Tren	Autobús	Coche particular	A pie
Baja	150	200	50	40
Media	150	50	60	30
Alta	100	50	90	30

¿Son independientes estos dos atributos?

SOLUCIÓN

Dos atributos son independientes si, para cada modalidad i del primer atributo y cada modalidad j del segundo atributo, la frecuencia relativa conjunta es igual al producto de las correspondientes frecuencias relativas marginales; esto es, para que dos atributos sean independientes tendrá que cumplirse, para cualesquiera i y j, que

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$$

o, equivalentemente,

$$\frac{n_{ij}}{N} = \frac{n_{i\cdot}}{N} \cdot \frac{n_{\cdot j}}{N}.$$

Simplificando, la igualdad anterior se convierte en

$$n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

condición de independencia, que, cuando los atributos son independientes, se cumple para todos los pares (i, j).

Es interesante recordar que de esta condición se habló también en el capítulo anterior, pues es igualmente válida para el estudio de la independencia entre variables.

Para estos dos atributos observamos que, si tomamos, por ejemplo, la primera modalidad de cada uno de ellos, la frecuencia absoluta conjunta es

$$n_{11} = 150,$$

mientras que

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{400 \cdot 440}{1\,000} = 176,$$

no cumpliéndose, por tanto, la condición de independencia para, al menos, una pareja de modalidades, y pudiéndose concluir, en consecuencia, que los atributos no son independientes.

La dependencia entre los atributos permite el estudio del tipo de asociación entre las modalidades de cada uno de ellos. De este modo, si comparamos

$$n_{21} = 150$$

con

$$\frac{n_{2\cdot} \cdot n_{\cdot 1}}{N} = 116,$$

vemos que la relación

$$n_{21} > \frac{n_{2\cdot} \cdot n_{\cdot 1}}{N}$$

indica que entre las modalidades *clase social media* y *elegir el tren como medio de transporte* existe asociación positiva.

Sería un buen ejercicio para el lector el análisis del tipo de asociación del resto de pares de modalidades.

3.8 La propietaria del centro de estética Unisex sospecha que no existe relación entre el sexo de los clientes y los tratamientos que solicitan. La siguiente tabla refleja la clase de tratamientos realizados a los 200 clientes que, entre hombres y mujeres, han acudido el pasado mes.

Sexo / Tratamiento	Hombres	Mujeres
Peluquería	20	80
Tratamiento facial	14	56
Tratamiento corporal	6	24

¿Tiene razón la empresaria en su suposición?

SOLUCIÓN

El cumplimiento de la condición de independencia para todos los pares de modalidades de dos atributos es condición necesaria y suficiente para que éstos sean independientes. Ahora bien, al

igual que ocurría en el estudio de la independencia entre variables, este hecho es equivalente a que, para cualesquiera i y j, se cumpla que

$$f_{i/j} = f_{i\cdot}$$

o, lo que es igual, que

$$f_{j/i} = f_{\cdot j},$$

condiciones que equivalen, asimismo, a la proporcionalidad de filas y columnas de la tabla de contingencia.

A partir de los datos iniciales, obtenemos las frecuencias marginales que aparecen en las últimas fila y columna de la siguiente tabla:

Sexo / Tratamiento	Hombres	Mujeres	$n_{i\cdot}$
Peluquería	20	80	100
Tratamiento facial	14	56	70
Tratamiento corporal	6	24	30
$n_{\cdot j}$	40	160	200

Se comprueba, de modo inmediato, que

$$\frac{20}{40} = \frac{80}{160} = \frac{100}{200},$$

esto es,

$$f_{1/1} = f_{1/2} = f_{1\cdot},$$

que, además,

$$\frac{14}{40} = \frac{56}{160} = \frac{70}{200},$$

es decir,

$$f_{2/1} = f_{2/2} = f_{2\cdot},$$

y que, por último,

$$\frac{6}{40} = \frac{24}{160} = \frac{30}{200},$$

o, lo que es igual,

$$f_{3/1} = f_{3/2} = f_{3\cdot},$$

quedando, así, demostrada la independencia de ambos atributos.

Puede probarse, de igual forma, que también se cumplen las condiciones:

$$f_{1/1} = f_{1/2} = f_{1/3} = f_{\cdot 1}$$

y

$$f_{2/1} = f_{2/2} = f_{2/3} = f_{\cdot 2}.$$

3.9 Dadas dos modalidades A_i y B_j de una distribución de frecuencias bidimensional $(A_i, B_j; n_{ij})$, analícense las relaciones entre las frecuencias relativas condicionadas $f_{i/j}$ y $f_{j/i}$ y las frecuencias relativas marginales, $f_{i\cdot}$ y $f_{\cdot j}$, a partir del signo de

$$n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}.$$

SOLUCIÓN

Supongamos que

$$n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{N}$$

es, por ejemplo, una cantidad positiva, esto es, que

$$n_{ij} > \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

con lo cual, como es sabido, existe asociación positiva entre las modalidades A_i y B_j.

Ahora bien, la condición anterior es equivalente a

$$\frac{n_{ij}}{n_{i\cdot}} > \frac{n_{\cdot j}}{N},$$

o, lo que igual, a

$$f_{j/i} > f_{\cdot j},$$

lo cual significa que la proporción de unidades de la población que posee la modalidad B_j *dentro* de las que tienen la modalidad A_i es *mayor* que la proporción de unidades de la población

que poseen la modalidad B_j en el total, hecho que es coherente con que exista asociación positiva entre A_i y B_j.

De la condición

$$n_{ij} > \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

también se deduce que

$$\frac{n_{ij}}{n_{\cdot j}} > \frac{n_{i\cdot}}{N},$$

esto es, que

$$f_{i/j} > f_{i\cdot},$$

de lo cual se concluye que la proporción de unidades de la población que tiene la modalidad A_i *dentro* de las que tienen la modalidad B_j es *mayor* que la proporción de unidades que tienen la modalidad A_i en el total de la población.

Por un razonamiento análogo, si, por el contrario, se cumpliera que

$$n_{ij} < \frac{n_{i\cdot} \cdot n_{\cdot j}}{N},$$

se deduciría que $f_{j/i} > f_{\cdot j}$ y que $f_{i/j} > f_{i\cdot}$, dejando al lector la interpretación de estas desigualdades entre proporciones.

3.10 Demuéstrese que en una tabla de contingencia de dimensión 2×2 es suficiente probar la condición de independencia con una pareja cualquiera de modalidades para que los atributos sean independientes.

SOLUCIÓN

Supongamos que, partiendo de la tabla de contingencia,

A \ B	B_1	B_2	$n_{i\cdot}$
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	

comprobamos que se cumple la condición de independencia para las dos primeras modalidades de los atributos, es decir, que se verifica:

$$n_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

entonces, la frecuencia absoluta conjunta

$$n_{12} = n_{1\cdot} - n_{11},$$

puede escribirse como

$$n_{12} = n_{1\cdot} - n_{11} = n_{1\cdot} - \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

sin más que tener en cuenta la condición de independencia para las modalidades A_1 y B_1.

Operando en la igualdad anterior, resulta:

$$n_{12} = \frac{N \cdot n_{1\cdot} - n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{n_{1\cdot}\,(N - n_{\cdot 1})}{N}.$$

En definitiva, considerando que $n_{\cdot 1} + n_{\cdot 2} = N$, se obtiene que

$$n_{12} = \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N},$$

condición de independencia de las modalidades A_1 y B_2.

Puede comprobar el lector que igualmente se cumplen las condiciones de independencia para los dos pares de modalidades restantes.

3.11 Demuéstrese que el tipo de asociación entre dos modalidades en una tabla de contingencia de dimensión 2×2 determina el resto.

SOLUCIÓN

Supongamos que existe asociación positiva entre las dos primeras modalidades de cada atributo,

$$n_{11} > \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

y queremos analizar el tipo de asociación que hay entre el resto de pares de modalidades.

Si nos fijamos, por ejemplo, en la segunda modalidad del primer atributo, A_2, y en la primera del segundo, B_1, y observamos su frecuencia absoluta conjunta, comprobamos que

$$n_{21} = n_{\cdot 1} - n_{11} < n_{\cdot 1} - \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

donde la última desigualdad es el resultado de aplicar la condición de asociación positiva entre las dos primeras modalidades y de tener en cuenta que un cambio de signo en los dos miembros de una desigualdad implica un cambio en el sentido de la misma:

$$-n_{11} < -\frac{n_{1\cdot} \cdot n_{\cdot 1}}{N}.$$

Operando en la frecuencia absoluta conjunta, resulta:

$$n_{21} < \frac{N \cdot n_{\cdot 1} - n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{n_{\cdot 1}\,(N - n_{1\cdot})}{N},$$

esto es,

$$n_{21} < \frac{n_{2\cdot} \cdot n_{\cdot 1}}{N},$$

con lo cual existe asociación negativa entre las modalidades consideradas.

Proponemos que el lector compruebe, mediante procedimiento análogo, la existencia de asociación positiva entre las modalidades A_2 y B_2,

$$n_{22} > \frac{n_{2\cdot} \cdot n_{\cdot 2}}{N},$$

y de asociación negativa entre la primera modalidad del primer atributo, A_1, y la segunda modalidad del segundo atributo, B_2,

$$n_{12} < \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N}.$$

3.12 Una empresa con 1 000 trabajadores ha solicitado un estudio con objeto de conocer la relación existente entre el sexo y el poseer o no titulación superior. Los resultados obtenidos se recogen en la siguiente tabla:

Sexo / Titulación superior	Hombre	Mujer
Sí	200	300
No	150	350

a) ¿Son independientes los atributos considerados?

b) Estúdiese el tipo de relación que existe entre *ser mujer* y *no poseer titulación superior.*

SOLUCIÓN

a) Si tomamos, por ejemplo, la segunda modalidad de primer atributo, *no poseer titulación superior*, y la primera del segundo atributo, *ser hombre*, vemos que

$$n_{21} = 150 \neq 175 = \frac{n_{2\cdot} \cdot n_{\cdot 1}}{N},$$

es decir, no se cumple la condición de independencia para, al menos, una pareja de modalidades y, por tanto, podemos afirmar que los dos atributos no son independientes.

Obsérvese que, por tratarse de una tabla de contingencia de dimensión 2×2, si se hubiera cumplido la condición de independencia para las dos modalidades elegidas, este hecho sería suficiente para afirmar que los dos atributos habrían sido independientes.

b) Vemos que

$$n_{22} = 350$$

es mayor que

$$\frac{n_{2\cdot} \cdot n_{\cdot 2}}{N} = \frac{500 \cdot 650}{1\,000} = 325,$$

con lo cual, entre las modalidades *no poseer titulación superior*, segunda modalidad del primer atributo, y *ser mujer*, segunda modalidad del segundo atributo, existe asociación positiva. Esto supone que el porcentaje de mujeres que no posee titulación superior es mayor al que existiría en caso de que ambos atributos fueran independientes.

Se puede obtener la misma conclusión teniendo en cuenta que, según comprobamos en el apartado anterior,

$$n_{21} = \frac{n_{2\cdot} \cdot n_{\cdot 1}}{N},$$

por lo que entre las modalidades *no poseer titulación superior* y *ser hombre* existe asociación negativa, lo que implica, necesariamente, que hay asociación positiva entre las modalidades *no poseer titulación superior* y *ser mujer*.

3.13 Se ha analizado una población de 300 individuos y se han clasificado según el tipo de trabajo que realizan (manual o intelectual) y su ideología política (conservador o liberal).

El estudio se ha elaborado con muestras de 100 trabajadores (3 estratos diferentes): hombres mayores de 30 años, mujeres mayores de 30 años y jóvenes (hombres y mujeres) menores de 30 años.

- Los resultados obtenidos en la primera muestra, mujeres mayores de 30 años, han sido:

Trabajo / Ideología	Manual	Intelectual
Conservador	36	24
Liberal	24	16

- Los resultados obtenidos en la segunda muestra, hombres mayores de 30 años, han sido:

Trabajo / Ideología	Manual	Intelectual
Conservador	40	20
Liberal	20	20

- Por último, para la tercera muestra, jóvenes menores de 30 años, los datos han sido:

Trabajo / Ideología	Manual	Intelectual
Conservador	10	20
Liberal	70	0

Analícese la relación existente entre estos atributos para cada una de las muestras seleccionadas.

SOLUCIÓN

Por lo que se refiere a la primera muestra, si observamos, por ejemplo, las modalidades *manual* y *conservador*, vemos que se cumple la condición de independencia:

$$n_{11} = 36 = \frac{60 \cdot 60}{100} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N}.$$

En otros términos, la proporción de conservadores entre los que realizan trabajo manual, $n_{11}/n_{\cdot 1} = 36/60$, es la misma que la proporción en el total, $n_{1\cdot}/N = 60/100$, es decir, el 60 por ciento.

Además, por tratarse de una tabla de dimensión 2×2, el cumplimiento de la condición de independencia para una pareja cualquiera de modalidades es suficiente para afirmar que existe independencia entre estos dos caracteres. Podemos observar, en cualquier caso, que con el resto de parejas de modalidades ocurre lo mismo.

Sin embargo, para la segunda muestra, si nos fijamos en la misma pareja de modalidades, se tiene que

$$n_{11} = 40 \neq 36 = \frac{60 \cdot 60}{100} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

esto es, no se cumple la condición de independencia entre estas dos modalidades, pudiendo decirse, por tanto, que existe dependencia entre estos dos atributos.

Además, como

$$n_{11} > \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

existe asociación positiva entre las modalidades *manual* y *conservador*, con lo cual, el porcentaje de conservadores entre los que realizan trabajos manuales $n_{11}/n_{\cdot 1} = 40/60$, esto es, el 66,66 por ciento, es superior al porcentaje de conservadores entre el total que, como sabemos, es del 60 por ciento.

Por último, para la tercera muestra vemos que

$$n_{11} = 10 < 24 = \frac{30 \cdot 80}{100} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

por lo que existe dependencia entre los atributos, habiendo, además, asociación negativa entre las modalidades *manual* y *conservador* porque el porcentaje de conservadores entre los que realizan trabajos manuales, $n_{11}/n_{\cdot 1} = 10/80$, es decir, 12,5 por ciento, es menor que 30, porcentaje de conservadores en el total de esta muestra.

Este ejemplo pone de manifiesto que el hecho de que exista o no independencia entre atributos y el tipo de asociación entre las modalidades de los mismos en caso de que éstos sean dependientes, no es algo intrínseco ni a los atributos ni a sus modalidades, sino que es consecuencia exclusivamente de las correspondientes frecuencias.

Proponemos al lector que analice el tipo de asociación que hay entre todos los pares de modalidades de los atributos considerados en los tres casos, bien directamente, esto es, comparando frecuencias, bien teniendo en cuenta que se trata de tablas de dimensión 2×2 y aplicando el resultado **3.11**.

3.14 Un estudio sobre la ocupación hotelera durante el mes de agosto del pasado año en 10 comunidades autónomas y su relación con el número de días lluviosos en dicho mes arrojó los siguientes resultados:

Ocupación hotelera / Días lluviosos	$< 50\%$	$\geq 50\%$
< 10	2	4
≥ 10	3	1

Indíquese qué tipo de asociación existe entre las modalidades de los atributos considerados.

SOLUCIÓN

Puesto que estos atributos tienen dos modalidades cada uno de ellos, es decir, se trata de una tabla de contingencia de dimensión 2×2, para el estudio del tipo de asociación existente entre cada par de modalidades podemos utilizar el coeficiente de asociación H, tomando como referencia una pareja cualquiera de modalidades. Así, considerando, por ejemplo, la primera modalidad del primer atributo y la segunda modalidad del segundo, se obtiene un valor del coeficiente de asociación:

$$\mathrm{H} = n_{12} - \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N} = 4 - \frac{6 \cdot 5}{10} = 1$$

que, al ser positivo, permite concluir la existencia de asociación positiva entre las modalidades *menos de 10 días lluviosos* y *50 por ciento o más de ocupación hotelera*.

En consecuencia con lo anterior, podemos afirmar, también, que entre las modalidades *10 o más días lluviosos* y *menos del 50 por ciento de ocupación hotelera* existe, igualmente, asociación positiva; que entre *menos de 10 días lluviosos* y *menos del 50 por ciento de ocupación hotelera* hay asociación negativa y, por último, que *entre 10 o más días lluviosos* y *50 por ciento o más de ocupación hotelera* existe, también, asociación negativa.

Puede comprobar el lector que llegaríamos a idénticas conclusiones calculando el coeficiente de asociación H a partir de cualquier otra pareja de modalidades.

3.15 Se clasifica una población según su sexo y situación laboral, y se obtienen los siguientes resultados:

Situación laboral / Sexo	Ocupado	Parado
Hombre	a	15
Mujer	30	10

a) ¿Cuál tendría que ser el valor de la constante a para que los atributos considerados fuesen independientes?

b) ¿Qué condición debería cumplir la constante a para que existiera asociación positiva entre las modalidades *mujer* y *parado*?

SOLUCIÓN

a) Puesto que, como es sabido, en una tabla de dimensión 2×2 es suficiente comprobar la condición de independencia con una pareja cualquiera de modalidades, eligiendo, por ejemplo, las dos segundas modalidades de los dos atributos, éstos serán independientes, si se cumple la igualdad:

$$n_{22} = \frac{n_{2\cdot} \cdot n_{\cdot 2}}{N}.$$

Teniendo en cuenta que el total de individuos de la población coincide con la suma de los elementos de las cuatro casillas de la tabla,

$$a + 15 + 30 + 10 = a + 55,$$

la condición de independencia anterior se convierte en

$$10 = \frac{40 \cdot 25}{a + 55},$$

con lo cual, despejando, se obtiene un valor de a igual a 45 para que los atributos sean independientes.

b) Si queremos que exista asociación positiva entre las modalidades *mujer* y *parado* deberá cumplirse:

$$n_{22} > \frac{n_{2\cdot} \cdot n_{\cdot 2}}{N}.$$

Un desarrollo análogo al realizado en el apartado anterior lleva a la condición $a > 45$ para que las modalidades *mujer* y *parado* tengan asociación positiva.

3.16 Una empresa con 100 trabajadores estudia la posibilidad de instalar en sus dependencias una máquina cafetera. Ante la duda de colocarla en el área de descanso de hombres o mujeres, la dirección encarga a dos de sus empleados, S. Alonso y L. Martínez, un estudio que arroje información sobre cuál de los dos grupos de trabajadores es mayor consumidor de café durante la jornada de trabajo.

Las tablas A y B recogen los resultados obtenidos por S. Alonso y L. Martínez, respectivamente:

A		
	Hombre	Mujer
Toma café	25	50
No toma	15	10

B		
	Mujer	Hombre
Toma café	50	25
No toma	10	15

A la vista de los datos, S. Alonso dice que «existe asociación negativa entre el sexo y el tomar o no café», mientras que L. Martínez afirma lo contrario. ¿Qué opinión estadística merecen estas conclusiones?

SOLUCIÓN

Como puede observarse, los datos obtenidos por ambos empleados, y reflejados en sus correspondientes tablas, son idénticos —únicamente están cambiadas las columnas de orden—, aunque sus conclusiones sean opuestas y, por supuesto, ambas erróneas, ya que realizan afirmaciones sobre la existencia de *tipo* de asociación entre atributos, siendo posible únicamente analizar si dos atributos son o no independientes y, cuando sean dependientes, su grado de asociación; el tipo de asociación se estudia solamente entre las distintas modalidades de cada atributo, en el caso de que éstos no sean independientes.

El único matiz que añadimos a lo comentado en este párrafo se refiere, como veremos en problemas posteriores, al caso de atributos cuyas modalidades admiten una ordenación; en tal circunstancia es posible analizar, además del grado de asociación entre los atributos, si la asociación entre las intensidades de ambos es creciente (positiva) o decreciente (negativa).

En consecuencia, con los datos obtenidos por estos empleados deberíamos empezar por comprobar la existencia o no de independencia entre los dos atributos.

Así, teniendo en cuenta que se trata de una tabla de dimensión 2×2 y fijándonos, por ejemplo, en las dos primeras modalidades de cada atributo, la condición de independencia es

$$n_{11} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N}.$$

En este caso, tomando como referencia la tabla A, se tiene, por un lado,

$$n_{11} = 25,$$

y, por otro,

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{N} = \frac{75 \cdot 40}{100} = 30,$$

con lo cual los atributos no son independientes.

Una vez constatada la dependencia entre los atributos, cabe preguntarse por el tipo de asociación que existe entre las modalidades de los mismos.

Puesto que

$$n_{11} < \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

concluimos que entre las modalidades *hombre* y *tomar café* hay asociación negativa o discordancia, y que es positiva la asociación entre las modalidades *hombre* y *no tomar café*. Asimismo, entre las modalidades *mujer* y *tomar café* existe asociación positiva o concordancia y entre las modalidades *mujer* y *no tomar café*, discordancia o asociación negativa.

Se puede resolver este apartado utilizando el coeficiente de asociación H para otra pareja cualquiera de modalidades.

3.17 Una editorial desea promocionar, en unos grandes almacenes, una colección de libros escritos por mujeres. Con objeto de decidir sobre la ubicación de los folletos de propaganda en una sección de compradores mayoritariamente masculinos o femeninos, intenta analizar la posible relación entre el sexo y los hábitos de lectura de autores femeninos, realizando, para ello, una encuesta a 100 personas de edades comprendidas entre 30 y 45 años.

Los resultados obtenidos, referidos a las características sexo y número de libros escritos por mujeres adquiridos en un año, se recogen en la siguiente tabla.

N.º libros / Sexo	< 5	5-10	> 10
Mujer	8	5	37
Hombre	32	5	13

A partir de los datos se calculó el coeficiente de asociación,

$$\mathrm{H} = n_{12} - \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N} = 0,$$

y se concluyó que ambas características son independientes. ¿Es acertada esta afirmación?

SOLUCIÓN

Al calcular el coeficiente de asociación H, tomando como referencia la modalidades primera y segunda de cada atributo, respectivamente, se está comprobando implícitamente la condición

de independencia para $i = 1$ y $j = 2$, que, en caso de verificarse, esto es, en caso de que H fuera igual a cero, bastaría para concluir que los atributos son independientes, siempre y cuando estemos ante una tabla de dimensión 2×2. Puesto que en esta ocasión la tabla es de dimensión 2×3, el hecho de que H sea cero es condición *necesaria* pero no *suficiente* para afirmar que los atributos son independientes y, en consecuencia, la conclusión, en principio, y salvo comprobaciones adicionales, es errónea.

Hecha esta consideración, y aunque para verificar que entre estos atributos no existe independencia, basta con observar, por ejemplo, que

$$n_{13} = 37 \neq 25 = \frac{n_{1\cdot} \cdot n_{\cdot 3}}{N},$$

vamos a optar, sin embargo, por utilizar el coeficiente χ^2 de Pearson, medida del grado de asociación entre atributos:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{ij} - \dfrac{n_{i\cdot} \cdot n_{\cdot j}}{N}\right)^2}{\dfrac{n_{i\cdot} \cdot n_{\cdot j}}{N}}.$$

Se observa que, en realidad, con el numerador de cada uno de los sumandos estamos comprobando la condición de independencia para cada par de modalidades de los atributos.

El cálculo de las frecuencias teóricas para cada par de modalidades nos lleva la siguiente tabla de doble entrada:

N.º libros / Sexo	< 5	5-10	> 10
Mujer	20	5	25
Hombre	20	5	25

En consecuencia, el coeficiente de contingencia de Pearson, coeficiente que compara las frecuencias observadas con las frecuencias teóricas, es

$$\chi^2 = \frac{(20-8)^2}{20} + \frac{(5-5)^2}{5} + \frac{(25-37)^2}{25} + \frac{(20-32)^2}{20} + \frac{(5-5)^2}{5} + \frac{(25-13)^2}{25} = 25{,}92,$$

valor distinto de cero indicativo, como ya sabíamos, de que los atributos no son independientes.

El coeficiente χ^2 es, además, la base para el cálculo del coeficiente de contingencia:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}},$$

coeficiente acotado entre 0 y 1, hecho que facilita su interpretación. Para los datos de este problema, este coeficiente resulta ser

$$C = \sqrt{\frac{25{,}92}{100 + 25{,}92}} = 0{,}45,$$

que, al estar más próximo a 0 que a 1, nos indica que existe escasa relación entre los atributos.

3.18 Demuéstrese que el coeficiente χ^2 admite como expresión la siguiente:

$$\chi^2 = N\left(\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_{i\cdot}\cdot n_{j}} - 1\right).$$

SOLUCIÓN

El coeficiente χ^2 es igual a

$$\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{\left(n_{ij} - \dfrac{n_{i\cdot}\cdot n_{j}}{N}\right)^2}{\dfrac{n_{i\cdot}\cdot n_{j}}{N}}.$$

Desarrollando el binomio que aparece en el numerador de este coeficiente, se tiene que

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2 + \left(\dfrac{n_{i\cdot}\cdot n_{\cdot j}}{N}\right)^2 - 2\cdot n_{ij}\cdot\dfrac{n_{i\cdot}\cdot n_{j}}{N}}{\dfrac{n_{i\cdot}\cdot n_{\cdot j}}{N}}.$$

Descomponiendo el doble sumatorio anterior en tres sumandos y haciendo las oportunas simplificaciones, resulta:

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^2}{\dfrac{n_{i\cdot}\cdot n_{\cdot j}}{N}} + \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{i\cdot}\cdot n_{\cdot j}}{N} - 2\sum_{i=1}^{h}\sum_{j=1}^{k}n_{ij}.$$

Ahora bien, por un lado,

$$\sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = N,$$

y, por otro,

$$\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{i\cdot}\cdot n_{\cdot j}}{N} = \frac{1}{N}\sum_{i=1}^{h} n_{i\cdot}\sum_{j=1}^{k} n_{\cdot j} = \frac{1}{N}\sum_{i=1}^{h} n_{i\cdot}\cdot N = \frac{N}{N}\sum_{i=1}^{h} n_{i\cdot} = N,$$

con lo cual, sustituyendo en la expresión del coeficiente, se obtiene que

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{ij}^2}{\frac{n_{i\cdot}\cdot n_{\cdot j}}{N}} + N - 2\cdot N = N\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot}\cdot n_{\cdot j}} - N$$

o, equivalentemente, tras sacar factor común,

$$\chi^2 = N\left(\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot}\cdot n_{j}} - 1\right),$$

según queríamos demostrar.

3.19 El departamento de recursos humanos de un laboratorio farmacéutico se propone controlar el absentismo laboral. El jefe de personal opina que éste puede estar relacionado con el nivel educativo de los trabajadores.

En la siguiente tabla se recoge la información sobre estas dos características para los empleados del laboratorio.

Absentismo / Nivel educativo	Alto	Medio	Bajo
Primario	1	26	33
Secundario	2	4	4
Superior	12	10	8

Indíquese si las siguientes afirmaciones son verdaderas o falsas:

a) Existe independencia entre el absentismo laboral y el nivel educativo de los trabajadores.

b) Existe asociación positiva entre el absentismo laboral y el nivel educativo de los trabajadores.

c) Existe asociación negativa entre el absentismo laboral y el nivel educativo de los trabajadores.

d) La proporción de trabajadores con un absentismo laboral alto es la misma, sea cual sea su nivel educativo.

e) El porcentaje de absentismo laboral medio dentro de los trabajadores con estudios secundarios es el mismo que en el total de trabajadores.

SOLUCIÓN

Completamos la tabla de contingencia con las frecuencias absolutas marginales, añadiendo las últimas fila y columna de la tabla:

Absentismo / Nivel educativo	Alto	Medio	Bajo	$n_{i\cdot}$
Primario	1	26	33	60
Secundario	2	4	4	10
Superior	12	10	8	30
$n_{\cdot j}$	15	40	45	100

a) FALSO. Basta con comprobar, por ejemplo, que

$$n_{11} = 1 \neq 9 = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{N},$$

con lo cual no se cumple la condición de independencia para las dos primeras modalidades de los atributos, condición necesaria para que éstos sean independientes.

b) FALSO. No tiene sentido hablar del tipo de asociación existente entre atributos; esta clase de análisis sólo puede realizarse entre modalidades.

c) FALSO. Por idénticas razones a las expresadas en el apartado anterior.

d) FALSO. La proporción de absentismo laboral alto dentro del nivel educativo primario es igual a

$$f_{1/1} = \frac{n_{11}}{n_{1\cdot}} = \frac{1}{60} = 0{,}0167,$$

en el nivel educativo secundario es

$$f_{1/2} = \frac{n_{12}}{n_{2\cdot}} = \frac{2}{10} = 0{,}2$$

y, dentro del nivel educativo superior, la proporción de absentismo laboral alto es

$$f_{1/3} = \frac{n_{13}}{n_{3\cdot}} = \frac{12}{30} = 0{,}4.$$

e) VERDADERO. El porcentaje de absentismo laboral medio entre los trabajadores con estudios secundarios se obtiene a partir de la proporción de trabajadores con absentismo laboral medio dentro del nivel de estudios secundarios:

$$f_{2/2} = \frac{n_{22}}{n_{2\cdot}} = \frac{4}{10} = 0{,}4.$$

Por tanto, el referido porcentaje es del 40 por ciento, valor que coincide con el porcentaje que representa dicho nivel de absentismo entre el total de los trabajadores, ya que

$$f_{\cdot 2} = \frac{n_{\cdot 2}}{N} = \frac{40}{100} = 0{,}4.$$

3.20 Se realiza una encuesta con el fin de estudiar las preferencias en materia de vivienda de los habitantes de una ciudad, resultando que el 40 por ciento de ellos prefieren la zona centro frente a la zona residencial.

Dividida la población en estratos, se obtuvo, además, que el 90 por ciento de los jóvenes (entre 18 y 35 años) prefiere la zona centro, siendo estos porcentajes del 30 por ciento y del 50 por ciento para adultos (entre 35 y 65 años) y ancianos (más de 65 años), respectivamente.

a) ¿Existe relación entre la edad y las preferencias sobre vivienda?

b) ¿Qué tipo de asociación hay entre la población adulta y la preferencia por la zona residencial?

SOLUCIÓN

La información que proporciona el enunciado puede expresarse de forma más cómoda mediante la siguiente tabla:

Edad	Zona residencial	Zona centro
Jóvenes	10%	90%
Adultos	70%	30%
Ancianos	50%	50%

Las filas segunda, tercera y cuarta de esta tabla contienen las frecuencias relativas condicionadas, expresadas en porcentajes, de las tres distribuciones del atributo *preferencia en materia de vivienda*, A, condicionadas por cada modalidad, B_j, del atributo *edad*, B:

Edad	Zona residencial	Zona centro
Jóvenes	$f_{1/1}$	$f_{2/1}$
Adultos	$f_{1/2}$	$f_{2/2}$
Ancianos	$f_{1/3}$	$f_{2/3}$

Conviene tener en cuenta, además, que, formalmente, haciendo las oportunas sustituciones, se obtiene que

$$f_{1/1} + f_{2/1} = \frac{n_{11}}{n_{.1}} + \frac{n_{12}}{n_{.1}} = \frac{n_{11} + n_{12}}{n_{.1}} = \frac{n_{.1}}{n_{.1}} = 1,$$

esto es, las frecuencias relativas de la distribución del atributo A condicionada por las modalidad ser *joven*, del atributo B, suman, naturalmente, la unidad.

Proponemos al lector que compruebe este hecho para las dos filas restantes de la tabla anterior.

a) Como puede observarse en las casillas de la primera columna de la tabla, la proporción de jóvenes que prefieren la zona residencial, $f_{1/1} = 0{,}10$, no es la misma que la proporción de adultos, $f_{1/2} = 0{,}70$, ni que la proporción de ancianos que tienen tal preferencia, $f_{1/3} = 0{,}50$, con lo cual, puede afirmarse que existe dependencia entre los atributos *edad* y *preferencia en materia de vivienda*. Si no fuera así, es decir, si los atributos fueran independientes, se cumpliría que

$$f_{1/1} = f_{1/2} = f_{1/3} = f_{1\cdot},$$

esto es, las proporciones serían iguales y coincidirían, además, con la proporción total de individuos de la población —jóvenes, junto con adultos y ancianos— que prefieren la zona residencial, que es igual a 0,6, es decir, al 60 por ciento.

Habríamos llegado a idéntico resultado razonando con las casillas de la segunda columna de la tabla.

b) Por un lado, el 60 por ciento de los individuos de la población prefieren la zona residencial y, por otro lado, el porcentaje de adultos que prefieren la zona residencial es de un 70 por ciento —mayor que el referido 60 por ciento—, puede afirmarse, entonces, que entre las modalidades *adulto* y *zona residencial* existe asociación positiva.

En realidad, hemos comparado las frecuencias relativas $f_{1\cdot}$ y $f_{1/2}$, viendo que

$$f_{1\cdot} < f_{1/2},$$

lo cual, según comprobamos en el problema **3.9** es equivalente a

$$n_{12} > \frac{n_{1\cdot} \cdot n_{\cdot 2}}{N},$$

condición indicativa de que existe asociación positiva entre estas modalidades.

3.21 Para conocer la relación existente entre el sexo y la posesión del título de doctor en una universidad, se ha realizado un estudio sobre su profesorado, obteniéndose que el 30 por ciento no posee titulación de doctor.

De los resultados del estudio se obtuvo que el 15 por ciento de los hombres no son doctores.

¿Qué tipo de asociación hay entre ser hombre y estar en posesión de título de doctor?

SOLUCIÓN

Como los dos atributos considerados, *sexo* y *poseer o no titulación de doctor*, tienen únicamente dos modalidades cada uno, la información que proporciona el enunciado es suficiente para saber que el 70 por ciento, esto es, 100-30, de los profesores universitarios poseen titulación de doctor y que, además, este porcentaje aumenta hasta el 85 por ciento, es decir, 100-15, en el caso de los hombres: en definitiva, entre *hombre* y *estar en posesión del título de doctor* existe asociación positiva.

3.22 Un estudio sobre las «grandes superficies» que se distribuyen por todo el territorio nacional pretende conocer si el tamaño de éstas está relacionado con sus beneficios anuales. Sobre una muestra de 100 centros se han obtenido los siguientes datos relativos a su tamaño y a sus beneficios anuales, en millones de euros.

Beneficios / Tamaño	Menos de 1	Entre 1 y 5	Más de 5 y menos de 7	7 o más
Pequeño	8	20	8	2
Mediano	1	15	10	8
Grande	1	5	12	10

Estúdiese la relación existente entre estos dos atributos.

SOLUCIÓN

Definimos para el atributo *tamaño del centro comercial* la variable X, cuyos valores 1, 2 y 3 son el rango o número de orden de las modalidades *pequeño*, *mediano* y *grande*. De la misma forma, para la característica *beneficios anuales* se define una variable, Y, con valores 1, 2, 3 y 4, correspondientes a los estados de esta característica. Hay que tener en cuenta que este carácter es cuantitativo, es decir, se trata, en realidad, de una variable de la cual nos interesa, exclusivamente, la ordenación de sus estados a la hora de estudiar el grado de asociación con las intensidades del atributo *tamaño del centro comercial*.

Para este análisis se calcula el coeficiente de correlación entre las variables X e Y,

$$r = \frac{S}{S_X \cdot S_Y},$$

El diagrama siguiente, cuya estructura fue analizada con detalle en el capítulo 2, servirá de apoyo a la hora de hallar los momentos necesarios para la obtención de r.

X \ Y	1	2	3	4	$n_{i\cdot}$	$x_i \cdot n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$	$\sum_{j=1}^{k} y_j \cdot n_{ij}$	$x_i \sum_{j=1}^{k} y_j \cdot n_{ij}$
1	8	20	8	2	38	38	38	80	80
2	1	15	10	8	34	68	136	93	186
3	1	5	12	10	28	84	252	87	261
$n_{\cdot j}$	10	40	30	20	100	190	426	260	527
$y_j^2 \cdot n_{\cdot j}$	10	160	270	320	760				

Del diagrama anterior resultan los momentos no centrales:

$$\bar{x} = a_{1,0} = \frac{1}{N} \sum_{i=1}^{h} x_i \cdot n_{i\cdot} = 1{,}9,$$

$$\bar{y} = a_{0,1} = \frac{1}{N} \sum_{j=1}^{k} y_j \cdot n_{\cdot j} = 2{,}6,$$

$$a_{2,0} = \frac{1}{N} \sum_{i=1}^{h} x_i^2 \cdot n_{i\cdot} = 4{,}26,$$

$$a_{0,2} = \frac{1}{N} \sum_{i=1}^{h} y_j^2 \cdot n_{\cdot j} = 7{,}6$$

y

$$a_{1,1} = \frac{1}{N} \sum_{i=1}^{h} \sum_{j=1}^{k} x_i \cdot y_j \cdot n_{ij} = 5{,}27.$$

A partir de estos momentos, obtenemos los momentos centrales, covarianza y varianzas de X y de Y:

$$S = a_{1,1} - a_{1,0} \cdot a_{0,1} = 5{,}27 - 1{,}9 \cdot 2{,}6 = 0{,}33,$$

$$S_X^2 = a_{2,0} - a_{1,0}^2 = 4{,}26 - 1{,}9^2 = 0{,}65$$

y

$$S_Y^2 = a_{0,2} - a_{0,1}^2 = 7{,}6 - 2{,}6^2 = 0{,}84.$$

Por consiguiente, el coeficiente de correlación lineal entre X e Y es

$$r = \frac{0{,}33}{\sqrt{0{,}65 \cdot 0{,}84}} = 0{,}45,$$

indicativo de un cierto grado de asociación creciente entre las intensidades de los caracteres considerados.

3.23 Obténganse los valores máximo y mínimo de $\sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j)$ y justifíquese, a la vista de los resultados, la expresión del coeficiente τ de Kendall.

SOLUCIÓN

Supongamos una ordenación de las N unidades de la población, según el orden natural de los rangos del primer atributo, A, de manera que la primera unidad será la que tenga el rango 1 para el atributo A, la segunda unidad tendrá rango 2 respecto al atributo, A, etc. De este modo, tendremos parejas de observaciones $(1, y_1)$, ..., (i, y_i), ..., (N, y_N), donde la primera componente es el rango de cada unidad según el atributo A y la segunda el rango respecto al atributo B.

Cuanto más próxima esté la ordenación de los rangos del atributo B, y_1, ..., y_i, ..., y_N, al orden natural de los rangos del atributo A, 1, ..., i, ..., N, mayor será la concordancia, y cuanto más próxima dicha ordenación esté al orden inverso al natural, N, ..., i, ..., 1, mayor será la discordancia. Consecuentemente, medir el grado de concordancia entre ambas ordenaciones es equivalente a ver cuál es el desorden —entendido como diferencia con el orden natural— que hay en los rangos del atributo B.

Por tanto, el coeficiente τ de Kendall, medida del grado de desorden en los rangos del atributo B, tendrá que reflejar en sus valores máximo y mínimo las dos situaciones extremas, situaciones que habrán de recogerse necesariamente en los valores máximo y mínimo de $\sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j)$,

ya que el resto de los términos que aparecen en la expresión del coeficiente no están influidos por la ordenación de los rangos.

Esta expresión alcanza su máximo cuando el resultado de todas las comparaciones entre cada rango, y_i, y cada uno de los rangos siguientes, y_j, con $i < j$, da como resultado un 1, es decir, cuando se mantiene el orden natural entre todos los rangos que se comparan, hecho que se produce cuando cada valor es menor que todos los siguientes. En tal caso, los pares de observaciones serán (1, 1), ..., (N, N), existiendo *concordancia absoluta* entre las ordenaciones según los rangos de los dos atributos.

Ahora bien, ¿cuánto vale esa suma cuando todos los sumando son iguales a 1? Para saberlo, basta con calcular el número de comparaciones, cantidad que, por supuesto, coincide con el número de sumandos. Ahora bien, teniendo en cuenta que y_i se compara con los $N-1$ rangos siguientes, y_2 con los $N-2$ rangos siguientes y, así sucesivamente, hasta llegar al rango y_{N-1} que se compara exclusivamente con y_N, tendremos un total de comparaciones igual a

$$(N-1) + (N-2) + ... + 1 = \frac{N(N-1)}{2},$$

siendo el segundo miembro de esta igualdad el resultado de sumar los $N-1$ términos de la progresión aritmética anterior[2].

Por el contrario, $\sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j)$ toma su mínimo valor cuando todos sus sumandos son iguales a -1, esto es, cuando cada rango que se compara es menor que el siguiente, situación que se da cuando la ordenación según los rangos del atributo B es N, ..., 1. En ese caso, los pares de datos son (1, N), ..., (N, 1), estando en la situación de *discordancia perfecta* entre las dos ordenaciones. Este valor mínimo será

$$-\frac{N(N-1)}{2},$$

cantidad que se obtiene multiplicando por -1 el número de comparaciones.

En definitiva,

$$-\frac{N(N-1)}{2} < \sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j) < \frac{N(N-1)}{2}.$$

[2] Recordemos que la suma de los n primeros términos de una progresión aritmética es

$$S = \frac{a_1 + a_n}{2} \cdot n,$$

donde a_1 y a_n son el primero y el último término de dicha progresión, respectivamente.

Dividiendo los tres miembros de las desigualdades anteriores entre $N\,(N-1)/2$, resulta:

$$-1 < \frac{2\sum_{\substack{i=1\\ i<j}}^{N} \delta\,(y_i, y_j)}{N\,(N-1)} < 1,$$

esto es, el coeficiente τ de Kendall está acotado entre -1 y 1, tomando sus valores extremos en las situaciones de perfecta discordancia y perfecta concordancia, respectivamente.

3.24 En los últimos años se ha venido impartiendo un curso de iniciación en técnicas informáticas destinado a los empleados de una cierta empresa. Los profesores están convencidos de que existe un alto grado de concordancia entre la efectividad del curso y el número de años que el empleado lleva en la empresa. Para probarlo toman un grupo de cinco trabajadores, J. Fernández, M. Domínguez, L. Sáez, F. González y T. Pérez, cuyo orden, atendiendo al número de años que han dedicado a la empresa, es: 2, 4, 5, 3 y 1.

Tras finalizar el curso, se les somete a un examen y su clasificación, según las puntuaciones obtenidas, es la siguiente: 4, 2, 5, 3 y 1.

A la vista de ambas ordenaciones, ¿qué juicio merece la opinión de los profesores?

SOLUCIÓN

El enunciado presenta cinco individuos, o unidades de la población, ordenados según los rangos de dos características. De estas características, *número de años que han dedicado a la empresa* y *puntuación que han obtenido en un examen*, que, por naturaleza, son variables pues sus observaciones son numéricas, no nos interesa, sin embargo, su cuantificación, sino la ordenación que sobre los individuos inducen.

Para analizar la posible existencia de concordancia entre estas dos ordenaciones vamos a calcular el coeficiente τ de Kendall,

$$\tau = \frac{2\sum_{\substack{i=1\\ i<j}}^{N} \delta\,(y_i, y_j)}{N\,(N-1)},$$

donde $\delta\,(y_i, y_j)$ es un indicador de la existencia o no de orden natural entre los rangos y_i e y_j.

Para hallar el valor del coeficiente elegimos una de las dos características, por ejemplo, *número de años dedicados a la empresa*, y ordenamos a los trabajadores siguiendo el orden natural —empezando por 1 y terminando por 5—, adjudicándoles, después, el rango correspondiente

de la otra característica, y_i. El resultado de esta reordenación queda recogido en la siguiente tabla.

Trabajador	Rango antigüedad	Rango puntuaciones y_i
T. Pérez	1	1
J. Fernández	2	4
F. González	3	3
M. Domínguez	4	2
L. Sáez	5	5

En la ordenación, $y_1, ..., y_N$, comparamos cada rango con todos los siguientes, asignando un 1, si hay orden natural y un –1, en caso contrario, es decir, si hay inversión del orden natural. Así, por ejemplo, $\delta(y_1, y_2) = \delta(1,4) = 1$, ya que 1 es menor que 4 y, por tanto, hay orden natural; en cambio, $\delta(y_2, y_3) = \delta(4,3) = -1$, pues, en este caso, hay inversión del orden natural al ser 4 menor que 3.

Repitiendo el proceso para el resto de los rangos, se obtienen los valores de todos los indicadores:

$\delta(y_1, y_2) = 1$ $\quad \delta(y_2, y_3) = -1$ $\quad \delta(y_3, y_4) = -1$ $\quad \delta(y_4, y_5) = 1$

$\delta(y_1, y_3) = 1$ $\quad \delta(y_2, y_4) = -1$ $\quad \delta(y_3, y_5) = 1$

$\delta(y_1, y_4) = 1$ $\quad \delta(y_2, y_5) = 1$

$\delta(y_1, y_5) = 1$

Por tanto, la suma de los indicadores es

$$\sum_{\substack{i=1 \\ i<j}}^{N} \delta(y_i, y_j) = 1 + 1 + 1 + 1 - 1 - 1 + 1 - 1 + 1 + 1 = 4$$

y el coeficiente de Kendall toma el valor

$$\tau = \frac{2 \cdot 4}{5\,(5-1)} = 0{,}4,$$

reflejo de cierta concordancia entre las dos ordenaciones.

3.25 De un estudio realizado por el departamento de marketing del grupo editorial Omega se obtiene que 10 de las familias consultadas presentan los ingresos anuales, en miles de euros, que aparecen recogidos en la siguiente tabla:

Familia	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}
Ingresos	10	52	61	30	63	28	23	80	31	18

Atendiendo al gasto anual en suscripciones a cualquier tipo de revista, la ordenación (de menor a mayor gasto) de estas familias es: 1, 6, 7, 4, 8, 2, 3, 9, 5 y 10.

A la vista de los datos, ¿puede afirmarse que el gasto anual en suscripciones a revistas y los ingresos familiares están relacionados?

SOLUCIÓN

Responder a esta pregunta es equivalente a analizar si existe o no concordancia entre las ordenaciones de las familias según los rangos de las dos características consideradas. Por ello, calculamos el coeficiente τ de Kendall,

$$\tau = \frac{2 \sum_{\substack{i=1 \\ i<j}}^{N} \delta\,(y_i, y_j)}{N\,(N-1)},$$

para lo cual, elegimos una de las dos características, por ejemplo, los *ingresos familiares*, y reclasificamos a las familias por el orden natural, haciéndoles corresponder después el respectivo rango de la característica *gasto anual en suscripciones*. Este proceso se recoge en la siguiente tabla:

Familia	Rango en ingresos	Rango gasto en suscripciones y_i
F_1	1	1
F_{10}	2	10
F_7	3	3
F_6	4	2
F_4	5	4
F_9	6	5
F_2	7	6
F_3	8	7
F_5	9	8
F_8	10	9

En la nueva ordenación, y_1, ..., y_N, hemos de comparar cada rango con todos los siguientes, asignando un 1, si existe orden natural, y un -1, en caso contrario, es decir, si hay inversión.

Este modo de proceder conduce a los valores que se explicitan en la tabla adjunta y que puede resultar cómoda al lector a la hora de calcular estos coeficientes. Obsérvese que en la primera fila de la tabla se recogen los resultados de comparar el rango y_1 con todos los siguientes; la segunda surge de comparar y_2 con los siguientes y así, hasta la última fila de la tabla, donde se compara y_9 con y_{10}.

	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
y_1	1	1	1	1	1	1	1	1	1
y_2		−1	−1	−1	−1	−1	−1	−1	−1
y_3			−1	1	1	1	1	1	1
y_4				1	1	1	1	1	1
y_5					1	1	1	1	1
y_6						1	1	1	1
y_7							1	1	1
y_8								1	1
y_9									1

En consecuencia, la suma de los indicadores, esto es, la suma de todos los elementos de la tabla anterior, es

$$\sum_{\substack{i=1 \\ i<j}}^{N} \delta\,(y_i, y_j) = 27$$

y el coeficiente de Kendall resulta ser

$$\tau = \frac{2 \cdot 27}{10\,(10-1)} = 0{,}6,$$

de lo cual concluimos que existe cierta concordancia entre ambas ordenaciones.

3.26 Obténgase la expresión del coeficiente de rangos de Spearman.

SOLUCIÓN

Según se vio en el capítulo anterior, dada una distribución de frecuencias bidimensional $(x_i, y_j; n_{ij})$, el coeficiente de correlación entre las variables X e Y se define como

$$r = \frac{S}{S_X \cdot S_Y},$$

donde S, S_X y S_Y son, respectivamente, la covarianza y las varianzas de las variables.

En esta situación, tenemos una distribución de frecuencias bidimensional unitaria con una tabla de correlación formada por unos y ceros (véase problema **2.31**), pues cada valor de la variable X se corresponde con un valor y sólo uno de la variable Y y, además, las dos variables toman los valores, 1, ..., N. Estas razones conducen a una expresión del coeficiente de correlación adaptada al caso que nos ocupa.

Así,

$$\overline{y} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}(1 + ... + N) = \frac{1}{N}\cdot\frac{N(N+1)}{2} = \frac{N+1}{2},$$

pues $\sum_{i=1}^{N} x_i$, suma de los términos de una progresión aritmética[3], es igual a $N(N+1)/2$.

Además, los momentos no centrales de orden 2 de las variables son[4]

$$a_{2,0} = a_{0,2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2 = \frac{1}{N}(1^2 + ... + N^2) =$$

$$= \frac{1}{N}\cdot\frac{N(N+1)\cdot(2\cdot N+1)}{6} = \frac{(N+1)\cdot(2\cdot N+1)}{6},$$

con lo cual, las varianzas son

$$S_X^2 = S_Y^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2 = \frac{(N+1)\cdot(2\cdot N+1)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12},$$

sin más que realizar sencillas operaciones aritméticas.

Denotando por $d_i = x_i - y_i$, se tiene que

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (x_i - y_i)^2 = \sum_{i=1}^{N} [(x_i - \overline{x}) - (y_i - \overline{y})]^2,$$

siendo el último miembro consecuencia de sumar y restar la misma cantidad, pues $\overline{x}$ coincide con $\overline{y}$.

Desarrollando el binomio anterior y descomponiendo en tres sumandos el resultado, se obtiene que

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (x_i - \overline{x})^2 + \sum_{i=1}^{N} (y_i - \overline{y})^2 - 2\sum_{i=1}^{N} (x_i - \overline{x})\cdot(y_i - \overline{y}).$$

---

[3] Véase nota 2.

[4] Estos momentos se obtienen a partir de la suma de los cuadrados de los n primeros números naturales que, recordemos, es igual a

$$\frac{n(n+1)\cdot(2\cdot n+1)}{6}.$$

Dividiendo los dos miembros de esta igualdad por N y sustituyendo por los correspondientes momentos, resulta que

$$\frac{1}{N}\sum_{i=1}^{N} d_i^2 = S_X^2 + S_Y^2 - 2 \cdot S,$$

por lo que, despejando la covarianza y teniendo en cuenta que las varianzas de las variables son iguales, se obtiene:

$$S = S_X^2 - \frac{1}{2 \cdot N}\sum_{i=1}^{N} d_i^2.$$

Por último, sustituyendo los valores de la varianza calculada anteriormente, se tiene el valor de la covarianza,

$$S = \frac{N^2 - 1}{12} - \frac{1}{2 \cdot N}\sum_{i=1}^{N} d_i^2.$$

En definitiva, el coeficiente de correlación para esta distribución de frecuencias o coeficiente de correlación de rangos de Spearman es, tras efectuar las oportunas operaciones:

$$\rho = \frac{\dfrac{N^2 - 1}{12} - \dfrac{1}{2 \cdot N}\displaystyle\sum_{i=1}^{N} d_i^2}{\sqrt{\dfrac{N^2 - 1}{12} \cdot \dfrac{N^2 - 1}{12}}} = \frac{\dfrac{N^2 - 1}{12} - \dfrac{1}{2 \cdot N}\displaystyle\sum_{i=1}^{N} d_i^2}{\dfrac{N^2 - 1}{12}},$$

esto es,

$$\rho = 1 - \frac{6\displaystyle\sum_{i=1}^{N} d_i^2}{N\,(N^2 - 1)}.$$

3.27 El nivel de eficiencia de los servicios de protección contra incendios de 5 ciudades españolas se analizó mediante dos técnicas diferentes. Los resultados obtenidos con la primera técnica indican que la ciudad más eficiente es Getafe, seguida, por orden de eficiencia, por Marbella, Santander, Barcelona y Oviedo. La ordenación proporcionada por la segunda técnica es: Marbella, Getafe, Santander, Oviedo y Barcelona. ¿Puede decirse que ambas técnicas conducen a análogos resultados?

SOLUCIÓN

Para ver si ambas técnicas aportan resultados análogos, esto es, para analizar el grado de concordancia entre ambas ordenaciones, calcularemos, en esta ocasión, el coeficiente de correlación de rangos de Spearman:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N},$$

donde d_i es la diferencia genérica entre los rangos de los dos atributos, es decir, entre los rangos de las dos ordenaciones.

En las columnas de la siguiente tabla aparecen ambas ordenaciones, así como los valores de d_i.

Ciudad	Rango 1ª ordenación x_i	Rango 2ª ordenación y_i	$d_i = x_i - y_i$
Getafe	1	2	–1
Marbella	2	1	1
Santander	3	3	0
Barcelona	4	5	–1
Oviedo	5	4	1

Teniendo en cuenta que

$$\sum_{i=1}^{N} d_i^2 = (-1)^2 + 1^2 + 0^2 + (-1)^2 + 1^2 = 4$$

y sustituyendo en la expresión del coeficiente, resulta:

$$\rho = 1 - \frac{6 \cdot 4}{5^3 - 5} = 0{,}8,$$

valor próximo a 1, lo cual indica que hay concordancia entre las ordenaciones y, en definitiva, que hay semejanza entre los resultados proporcionados por ambas técnicas.

3.28 En el país de Malustiana se han celebrado recientemente elecciones generales en las que los ciudadanos votaron a sus representantes políticos. Cinco fueron los partidos y coaliciones que se presentaron a los comicios con las siguientes siglas: P.O.S.; C.C.A.; P.V.T.; A.S.P. y U.P.

La empresa Alfa Cuatro, dedicada a estudios de mercado, realizó una encuesta previa a las elecciones con los siguientes resultados:

Partido	P.O.S.	C.C.A.	P.V.T.	A.S.P.	U.P.
% votos	30%	20%	9%	40%	1%

Una vez finalizado el escrutinio de los votos se obtuvieron los siguientes porcentajes:

Partido	P.O.S.	C.C.A.	P.V.T.	A.S.P.	U.P.
% votos	15%	10%	30%	5%	40%

Analícese el grado de concordancia existente entre los resultados de las votaciones y los previstos por la encuesta.

SOLUCIÓN

Para obtener el coeficiente de correlación de rangos de Spearman,

$$\rho = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N^3 - N},$$

calculamos las correspondientes diferencias d_i entre los rangos de ambas ordenaciones, según se recoge en la última fila de la tabla adjunta:

Partido	P.O.S.	C.C.A.	P.V.T.	A.S.P.	U.P.
1ª ordenación (encuesta)	2	3	4	1	5
2ª ordenación (votaciones)	3	4	2	5	1
d_i	–1	–1	2	–4	4

Sustituyendo en la expresión de coeficiente resulta que

$$\rho = 1 - \frac{6 \cdot 38}{5^3 - 5} = 1 - \frac{228}{120} = -0{,}9,$$

valor indicativo de una fuerte discordancia entre las dos ordenaciones.

Capítulo 4

Números índices y tasas de variación

Principales conceptos y resultados

Con frecuencia interesa analizar la evolución de una magnitud en el tiempo o en el espacio; más concretamente, el objetivo puede ser comparar las observaciones de una variable obtenida a lo largo del tiempo o del espacio.

Así, un **número índice** es una medida estadística que sirve para estudiar las variaciones de una magnitud en distintas situaciones[1].

La observación que deseamos comparar pertenece al denominado **periodo actual** o **corriente** y se hace con respecto a una observación tomada en el **periodo base** o **de referencia**.

Si se trata de estudiar la evolución de una magnitud *simple* —variable estadística *unidimensional*—, utilizaremos **índices simples**. Si, por el contrario, el propósito es analizar la variación de una magnitud *compleja* —variable estadística *N-dimensional*—, trabajaremos con **índices complejos**. A su vez, los índices complejos pueden ser **no ponderados** o **ponderados**, según se considere, a la hora de realizar la comparación, que las componentes de la magnitud han de tener la misma importancia o no.

Sea una variable Y y sean y_0 e y_t las observaciones de dicha variable en los periodos base y actual, respectivamente. El número índice simple mide la variación de la variable entre los periodos considerados y se define como

$$I_0^t = \frac{y_t}{y_0}.$$

[1] Los ejercicios de este capítulo corresponden únicamente a comparaciones en el tiempo.

Los índices simples más frecuentes son los que resultan de considerar como variables el precio, la cantidad o el valor de un bien.

Los índices complejos más utilizados son el **índice de Laspeyres** y el **índice de Paasche**. Estos índices son índices complejos ponderados y están basados en la *media aritmética ponderada*[2].

Así, dada una variable Y, cuyas componentes son $Y_1, ..., Y_i, ..., Y_N$ y dadas y_{i0} e y_{it} observaciones de Y_i $(i = 1, ..., N)$ en los periodos base y actual, respectivamente, se define el **índice media aritmética ponderada** en el periodo t con base en el periodo 0 como

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde

$$I_0^t(i) = \frac{y_{it}}{y_{i0}}$$

es el índice simple de la componente i-ésima, w_i es el *coeficiente de ponderación* y $w_i / \sum_{i=1}^{N} w_i$ es el *peso* o *ponderación* de dicho índice simple.

Cuando en la expresión anterior consideramos como magnitud compleja el precio de N bienes, $P_1, ..., P_i, ..., P_N$, y como coeficiente de ponderación $w_i = p_{i0} \cdot q_{i0}$, obtenemos el índice de precios de Laspeyres[3]:

$$L_{p0}^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}.$$

Si, por el contrario, tomamos como coeficiente de ponderación $w_i = p_{i0} \cdot q_{it}$, tendremos el índice de precios de Paasche:

$$P_{p0}^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}}.$$

---

[2] Los índices complejos *resumen* en una sola medida la información proporcionada por los índices simples de cada componente, siendo, por tanto, razonable utilizar promedios a la hora de realizar dicha síntesis de información. En este sentido, podemos construir índices complejos basados en otras medidas de posición.

[3] El índice de precios al consumo, IPC, es un índice de Laspeyres.

Análogamente, se definen los índices de cantidades de Laspeyres y Paasche. Estos índices miden la evolución conjunta de las cantidades correspondientes a N bienes, Q_1, ..., Q_i, ..., Q_N.

Por tanto,

$$L_{q0}^{t} = \frac{\sum_{i=1}^{N} q_{it} \cdot p_{i0}}{\sum_{i=1}^{N} q_{i0} \cdot p_{i0}}$$

es el índice de cantidades de Laspeyres, en el que como coeficiente de ponderación se utiliza $w_i = q_{i0} \cdot p_{i0}$, y

$$P_{q0}^{t} = \frac{\sum_{i=1}^{N} q_{it} \cdot p_{it}}{\sum_{i=1}^{N} q_{i0} \cdot p_{it}}$$

es el índice de cantidades de Paasche, cuyo coeficiente de ponderación[4] es $w_i = q_{i0} \cdot p_{it}$.

Las propiedades *deseables* de un número índice son:

- *Existencia*: un número índice ha de ser una cantidad finita distinta de cero.
- *Identidad*:

$$I_0^0 = 1.$$

- *Circular*: al considerar los periodos 0, t' y t se cumple que

$$I_0^t = I_0^{t'} \cdot I_{t'}^{t}.$$

- *Inversión*:

$$I_t^0 = \frac{1}{I_0^t}.$$

- *Proporcionalidad*: dada una constante de proporcionalidad k, si

$$y_{t'} = (1 + k)\, y_t,$$

entonces,

$$I_0^{t'} = (1 + k)\, I_0^t.$$

[4] En los índices de producción se pondera por el *valor neto* o *valor añadido* del bien, esto es, se toma como precio correspondiente al año base la diferencia entre el precio de venta y el precio de coste.

• *Homogeneidad:* no debe estar afectado por cambios en las unidades de medida de la magnitud.

Estas propiedades se cumplen en los números índices simples y, en general, no se cumplen en los complejos.

El **cambio de base** en una serie de números índices simples se realiza aplicando la propiedad circular[5]. En efecto, dados los periodos 0, t' y t, se cumple que

$$I_{t'}^{t} = \frac{I_0^t}{I_0^{t'}},$$

expresión que permite referir la serie de índices al año t'.

En el estudio de la evolución en el tiempo del valor de una magnitud económica, surge el problema de la *depreciación* monetaria. Para poder comparar el valor de una magnitud económica en distintas situaciones hemos de considerar su **valor real** a **precios constantes**, esto es, a precios que rigen en el mercado en un periodo concreto. Con objeto de transformar el **valor nominal**, expresado a **precios corrientes** que rigen en cada periodo, en valor real, se utiliza un número índice denominado **deflactor**. El deflactor más utilizado es el índice de Laspeyres.

En general, la transformación de precios corrientes a precios constantes responde a la expresión:

$$\text{precios constantes año } t \text{ (base 0)} = \frac{\text{precios corrientes año } t}{D_0^t},$$

donde D_0^t es el número índice utilizado como deflactor.

Dada una serie temporal, y_1, ..., y_T, la **variación absoluta** entre los periodos $t - 1$ y t es

$$\Delta y_t = y_t - y_{t-1}.$$

La **variación relativa** o **proporcional** o **tasa de variación** entre los periodos $t - 1$ y t se define como[6]:

$$\dot{y}_t = \frac{\Delta y_t}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1.$$

Operando en la expresión de la tasa de variación relativa resulta la igualdad:

$$y_t = y_{t-1}\,(1 + \dot{y}_t),$$

[5] El hecho de que los índices complejos no cumplan, en general, la propiedad circular supone que el cambio de base en ellos se realice sólo de manera aproximada.

[6] La tasa de variación, $\dot{y}_t$, es igual al índice simple, I_{t-1}^{t}, menos la unidad.

es decir, el valor correspondiente al periodo t, y_t, es el resultado de incrementar el valor en el periodo anterior, y_{t-1}, en la tasa de variación, $\dot{y}_t$.

La **tasa media de variación** o **tasa media acumulativa**, tm, es un valor que, aplicado sucesivamente de periodo a periodo a las distintas observaciones de la serie, permite la obtención de la última observación, partiendo de la primera:

$$y_T = y_1 (1 + \text{tm})^{T-1}.$$

Despejando se tiene que

$$\text{tm} = \sqrt[T-1]{\frac{y_T}{y_1}} - 1,$$

expresión de la tasa media de variación en función de las observaciones inicial y final.

La tasa media de variación puede calcularse, también, mediante las tasas de variación:

$$\text{tm} = \sqrt[T-1]{(1 + \dot{y}_2) \ldots (1 + \dot{y}_T)} - 1,$$

interpretándose como la *media geométrica* de los **factores de variación unitarios**, $(1 + \dot{y}_2)$, ... $(1 + \dot{y}_T)$, de cada periodo menos la unidad.

De la aplicación de las definiciones de variación absoluta y relativa a un índice surgen los conceptos de repercusión y participación.

En efecto, dado un índice complejo ponderado de índices simples,

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t (i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

se denomina **repercusión absoluta** de la componente i-ésima sobre la *variación absoluta del índice* entre los periodos $t-1$ y t, $\Delta I_0^t = I_0^t - I_0^{t-1}$, al cociente:

$$R_i = \frac{\Delta I_0^t (i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde $\Delta I_0^t (i) = I_0^t (i) - I_0^{t-1} (i)$ es la *variación absoluta del índice simple de la componente i-ésima* entre los periodos $t - 1$ y t.

La variación absoluta del índice se obtiene como suma de las repercusiones absolutas de todos los bienes.

El cociente

$$r_i = \frac{\Delta I_0^t\,(i) \cdot w_i}{\sum_{i=1}^{N} I_0^{t-1}\,(i) \cdot w_i}$$

es la **repercusión relativa** de la componente *i*-ésima sobre *la variación relativa del índice* entre los periodos $t - 1$ y t, $\dot{I}_0^t = \Delta I_0^t / I_0^{t-1}$, y suele expresarse en porcentajes.

La suma de las repercusiones relativas de todos los bienes es igual a la tasa de variación del índice.

Se denomina **participación** de la componente *i*-ésima en la variación relativa del índice al cociente entre la repercusión relativa de la componente y la tasa de variación del índice:

$$P_i = \frac{r_i}{\dot{I}_0^t}.$$

La suma de las participaciones de las componentes de un índice complejo ponderado, expresadas en porcentajes, es igual a 100.

Los conceptos de participación y repercusión son habitualmente aplicados al índice de precios de Laspeyres.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

4.1 Dado el precio de un bien para el periodo 1999-2004, obténgase una serie de índices, tomando como año base el año 1999.

Años	Precios
1999	10
2000	15
2001	17
2002	20
2003	25
2004	30

SOLUCIÓN

Para obtener una serie de números índices con base en el año 1999, comparamos los precios del bien para cada año, y_t, con el precio de 1999, y_{99}, esto es,

$$I_{99}^{t} = \frac{y_t}{y_{99}},$$

con lo cual, resulta una serie de índices *fijos*:

$$I_{99}^{99} = \frac{y_{99}}{y_{99}} = \frac{10}{10} = 1$$

$$I_{99}^{00} = \frac{y_{00}}{y_{99}} = \frac{15}{10} = 1{,}5$$

$$I_{99}^{01} = \frac{y_{01}}{y_{99}} = \frac{17}{10} = 1{,}7$$

$$I_{99}^{02} = \frac{y_{02}}{y_{99}} = \frac{20}{10} = 2$$

$$I_{99}^{03} = \frac{y_{03}}{y_{99}} = \frac{25}{10} = 2{,}5$$

$$I_{99}^{04} = \frac{y_{04}}{y_{99}} = \frac{30}{10} = 3,$$

serie que pone de manifiesto el crecimiento del precio a lo largo del periodo, llegando a ser dicho aumento del 300 por ciento entre los años 1999 y 2004.

Es inmediato comprobar que el índice definido, como índice simple que es, cumple la propiedad de la *identidad*:

$$I_{99}^{99} = \frac{y_{99}}{y_{99}} = 1,$$

y, también, la propiedad de la *inversión*:

$$I_t^{99} = \frac{y_{99}}{y_t} = \frac{1}{\frac{y_t}{y_{99}}} = \frac{1}{I_{99}^t}.$$

Proponemos al lector que calcule otras series de índices simples con los datos del enunciado, cambiando el año de referencia.

Para obtener la serie de números índices *en cadena,* comparamos el precio en cada año con el precio del bien en el año inmediatamente anterior, según la relación

$$I_{t-1}^t = \frac{y_t}{y_{t-1}},$$

que, aplicada a los datos del problema, proporciona la serie de índices:

$$I_{99}^{00} = \frac{y_{00}}{y_{99}} = \frac{15}{10} = 1{,}5$$

$$I_{00}^{01} = \frac{y_{01}}{y_{00}} = \frac{17}{15} = 1{,}13$$

$$I_{01}^{02} = \frac{y_{02}}{y_{01}} = \frac{20}{17} = 1{,}17$$

$$I_{02}^{03} = \frac{y_{03}}{y_{02}} = \frac{25}{20} = 1{,}25$$

$$I_{03}^{04} = \frac{y_{04}}{y_{03}} = \frac{30}{25} = 1{,}2.$$

Se obtendría el mismo resultado con la propiedad *circular* de los índices simples que demostramos a continuación. Así, dado un índice simple de una magnitud en el año t con base en el año 1999,

$$I_{99}^{t} = \frac{y_t}{y_{99}},$$

multiplicando y dividiendo por y_{t-1} resulta que

$$I_{99}^{t} = \frac{y_t}{y_{t-1}} \cdot \frac{y_{t-1}}{y_{99}},$$

o, lo que es igual,

$$I_{99}^{t} = I_{t-1}^{t} \cdot I_{99}^{t-1},$$

propiedad circular de los índices simples.

Ahora bien, despejando en la igualdad anterior, se obtiene:

$$I_{t-1}^{t} = \frac{I_{99}^{t}}{I_{99}^{t-1}},$$

relación entre los índices en cadena y los índices fijos con base, en este caso, el año 1999.

Puede ser un buen ejercicio para que el lector se familiarice con las notaciones de este capítulo el comprobar que este camino conduce, efectivamente, a la misma solución.

4.2 Sean las variables Y, U y V, tales que $Y = U \cdot V$. Exprésese el índice simple de la variable Y en el año t con base en el año 0, I_0^t, a partir de U_0^t y de V_0^t, índices simples de las variables U y V en el año t con base en el año 0, respectivamente.

SOLUCIÓN

Por definición de índice simple,

$$I_0^t = \frac{y_t}{y_0}.$$

De la relación existente entre las variables Y, U y V se tiene, para los años t y 0, que

$$y_t = u_t \cdot v_t$$

e

$$y_0 = u_0 \cdot v_0.$$

Sustituyendo en la expresión del índice simple de la variable Y, resulta la siguiente relación entre los índices simples de las tres variables:

$$I_0^t = \frac{u_t \cdot v_t}{u_0 \cdot v_0} = U_0^t \cdot V_0^t$$

4.3 El precio de la bombona de butano aumentó entre 2002 y 2004 un 12 por ciento disminuyendo la cantidad vendida en un 5 por ciento. Obténgase el valor relativo de este artículo entre los años considerados.

SOLUCIÓN

La relación entre las variables valor, V, precio, P, y cantidad, Q, es $V = P \cdot Q$, por lo que aplicando el resultado del ejercicio anterior, se tiene que

$$V_{02}^{04} = P_{02}^{04} \cdot Q_{02}^{04},$$

donde V_{02}^{04}, P_{02}^{04} y Q_{02}^{04} son los correspondientes índices simples.

El enunciado del problema especifica que el precio de la bombona aumentó en un 12 por ciento entre 2002 y 2004, es decir, el precio en el año 2004, p_{04}, se relaciona con el precio en 2002, p_{02}, según la expresión:

$$p_{04} = p_{02} + 0{,}12 \cdot p_{02} = (1 + 0{,}12)\, p_{02} = 1{,}12 \cdot p_{02}.$$

Por tanto, el precio relativo es

$$P_{02}^{04} = \frac{p_{04}}{p_{02}} = \frac{1{,}12 \cdot p_{02}}{p_{02}} = 1{,}12.$$

Este resultado es coherente con la definición de índice simple, indicador de la variación de una magnitud entre dos periodos.

Puesto que, por otro lado, la cantidad disminuyó un 5 por ciento, la relación entre las cantidades en 2004, q_{04}, y en 2002, q_{02}, es

$$q_{04} = q_{02} - 0{,}05 \cdot q_{02} = (1 - 0{,}05)\, q_{02} = 0{,}95 \cdot q_{02},$$

con lo cual, la cantidad relativa es

$$Q_{02}^{04} = \frac{q_{04}}{q_{02}} = \frac{0{,}95 \cdot q_{02}}{q_{02}} = 0{,}95,$$

que refleja la disminución de la cantidad vendida en un 5 por ciento entre 2002 y 2004.

En definitiva, el valor relativo de este artículo entre los años considerados es

$$V_{02}^{04} = 1{,}12 \cdot 0{,}95 = 1.064,$$

que indica que entre 2002 y 2004 el valor del artículo de consumo aumentó un 6,4 por ciento.

4.4 El precio de un modelo de «deportivas» en 2004 es un 3 por ciento superior a su precio en 2003 y un 15 por ciento superior a su precio en 2000. Hállese el precio relativo entre 2000 y 2003.

SOLUCIÓN

Por definición de índice simple, y según lo visto en **4.3**, se tiene, por un lado, que el precio relativo de las «deportivas» en 2004 respecto a 2003 es

$$P_{03}^{04} = 1{,}03,$$

y, por otro lado, el precio relativo en 2004 respecto a 2000 es

$$P_{00}^{04} = 1{,}15.$$

Para calcular P_{00}^{03}, precio relativo en 2003 respecto a 2000, ha de aplicarse la propiedad circular:

$$P_{00}^{04} = P_{00}^{03} \cdot P_{03}^{04}.$$

Despejando en la igualdad anterior,

$$P_{00}^{03} = \frac{P_{00}^{04}}{P_{03}^{04}} = \frac{1{,}15}{1{,}03} = 1{,}116,$$

índice que expresa que entre 2000 y 2003 el precio de las «deportivas» se incrementó un 11,6 por ciento.

4.5 Los créditos obtenidos por un país en los mercados internacionales a medio y largo plazo, en miles de euros, en el periodo 2001-2004, han sido:

Años	2001	2002	2003	2004
Créditos	9 000	9 200	6 300	6 000

Determínese la tasa media de variación para dicho periodo.

SOLUCIÓN

El incremento medio anual o tasa media de los créditos se obtiene a partir de las observaciones inicial, y_1, y final, y_T, de la serie de créditos:

$$\text{tm} = \sqrt[T-1]{\frac{y_T}{y_1}} - 1 = \sqrt[4-1]{\frac{6\,000}{9\,000}} - 1 = -0{,}126.$$

Advierta el lector que son 4 los periodos (años) considerados, con lo cual, el orden de la raíz es 4-1.

4.6 Demuéstrese que $\text{tm} = \sqrt[T-1]{(1 + \dot{y}_2) \cdot (1 + \dot{y}_3) \ldots (1 + \dot{y}_T)} - 1$.

SOLUCIÓN

La tasa media de variación es, por definición,

$$\text{tm} = \sqrt[T-1]{\frac{y_T}{y_1}} - 1.$$

Multiplicando y dividiendo el cociente y_T/y_1 por $y_2, y_3, \ldots, y_{T-1}$ se tiene que

$$\frac{y_T}{y_1} = \frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \ldots \frac{y_T}{y_{T-1}}.$$

Ahora bien, de la definición de tasa de variación,

$$y_t = y_{t-1}\,(1 + \dot{y}_t),$$

resultan, de modo inmediato, las siguientes igualdades:

$$\frac{y_2}{y_1} = 1 + \dot{y}_2$$

$$\frac{y_3}{y_2} = 1 + \dot{y}_3$$

...

$$\frac{y_T}{y_{T-1}} = 1 + \dot{y}_T.$$

Con lo cual, sin más que sustituir en la definición de tasa media,

$$\text{tm} = \sqrt[T-1]{\frac{y_T}{y_1}} - 1 = \sqrt[T-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \ldots \frac{y_T}{y_{T-1}}} - 1 = \sqrt[T-1]{(1 + \dot{y}_2) \cdot (1 + \dot{y}_3) \ldots (1 + \dot{y}_T)} - 1,$$

queda demostrado el resultado que permite la obtención de la tasa media acumulativa a partir de las tasas de variación.

4.7 El precio medio, en miles euros, de las motos de una cierta cilindrada para el periodo 2000-2004 ha sido:

Años	2000	2001	2002	2003	2004
Créditos	800	850	900	950	1 000

a) Hállense los incrementos relativos de los precios en el periodo considerado.

b) Calcúlese la tasa media anual de los precios medios a partir de las tasas obtenidas en el apartado anterior.

SOLUCIÓN

a) Los incrementos relativos de los precios, esto es, las tasas de variación de la variable precio, se calculan según la expresión genérica:

$$\dot{p}_t = \frac{p_t}{p_{t-1}} - 1,$$

para $t = 2001, \ldots, 2004$.

Los resultados obtenidos de la aplicación de esta expresión para los años del periodo considerado se recogen en la tabla siguiente:

Años	Tasas de variación
2001	$\frac{850}{800} - 1 = 0{,}062$
2002	$\frac{900}{850} - 1 = 0{,}059$
2003	$\frac{950}{900} - 1 = 0{,}055$
2004	$\frac{1\,000}{950} - 1 = 0{,}053$

b) La tasa media de los precios para el periodo 2000-2004, calculada a partir de las tasas de variación, es

$$\text{tm} = \sqrt[5-1]{(1 + 0{,}062) \cdot (1 + 0{,}059) \cdot (1 + 0{,}055) \cdot (1 + 0{,}053)} - 1 = 0{,}057.$$

4.8 La siguiente tabla recoge el número de alumnos, en miles, en Educación infantil/ Preescolar y EGB/Primaria para un cierto periodo:

Años	1995	1996	1997	1998	1999
Públicos	47,63	46,44	44,80	43,25	41,73
Privados	34,47	33,57	32,47	30,67	28,98

Calcúlese la tasa media de variación del total de alumnos para el periodo considerado.

SOLUCIÓN

Puesto que hemos de calcular la tasa media de variación del total de alumnos y la información del enunciado está desagregada en centros públicos y privados, obtendremos el total en cada año sumando los totales de cada tipo de centro. Ahora bien, para el cálculo de la tasa media de variación sólo necesitamos dicho valor para el primer y último año, es decir, para 1995 y para 1999.

Así, el total de alumnos para el año 1995 es

$$y_{95} = 47{,}63 + 34{,}47 = 82{,}10$$

y el total de alumnos para 1999 resulta ser

$$y_{99} = 41{,}73 + 28{,}98 = 70{,}71.$$

En consecuencia, la tasa media de variación es

$$\text{tm} = \sqrt[5-1]{\frac{y_{99}}{y_{95}}} - 1 = \sqrt[4]{\frac{70{,}71}{82{,}10}} - 1 = -0{,}0366.$$

También podríamos haber llegado al mismo resultado partiendo de las tasas de variación calculadas de año en año,

$$\dot{y}_{96} = \frac{80{,}01}{82{,}10} - 1 = -0{,}0254$$

$$\dot{y}_{97} = \frac{77{,}27}{80{,}01} - 1 = -0{,}0342$$

$$\dot{y}_{98} = \frac{73{,}92}{77{,}27} - 1 = -0{,}0433$$

$$\dot{y}_{99} = \frac{70{,}71}{73{,}92} - 1 = -0{,}0434,$$

y, aplicando el resultado **4.6**:

$$\begin{aligned}\text{tm} &= \sqrt[4]{(1+\dot{y}_{96})\cdot(1+\dot{y}_{97})\cdot(1+\dot{y}_{98})\cdot(1+\dot{y}_{99})} - 1 = \\ &= \sqrt[4]{(1-0{,}0254)\cdot(1-0{,}0342)\cdot(1-0{,}0433)\cdot(1-0{,}0434)} - 1 = -0{,}0366.\end{aligned}$$

4.9 La siguiente tabla refleja la recaudación líquida por operaciones corrientes de un Ayuntamiento, en miles de euros, para el periodo 2001-2004.

Ingresos por capítulos	2001	2002	2003	2004
1 Impuestos directos	5 500	5 666	6 227	6 743
2 Impuestos indirectos	583	391	456	388
3 Tasas y otros ingresos	3 934	4 185	4 079	4 341
4 Transferencias corrientes	3 508	3 661	3 899	4 426
5 Ingresos patrimoniales	204	724	242	201
Total operaciones corrientes	13 729	14 627	14 903	16 099

a) Calcúlese el incremento relativo interanual de la recaudación líquida por operaciones corrientes para el periodo considerado.

b) ¿En qué capítulo se produjo un mayor incremento medio anual?

SOLUCIÓN

a) El incremento relativo interanual o tasa de variación entre t–1 y t es, por definición,

$$\dot{y}_t = \frac{y_t}{y_{t-1}} - 1.$$

Por tanto, las tasas de variación del periodo considerado son:

$$\dot{y}_{02} = \frac{14\,627}{13\,729} - 1 = 0{,}065$$

$$\dot{y}_{03} = \frac{14\,903}{14\,627} - 1 = 0{,}019$$

$$\dot{y}_{04} = \frac{16\,099}{14\,903} - 1 = 0{,}080.$$

b) Calculando la tasa media de variación de cada uno de los capítulos de ingresos, según la expresión:

$$\text{tm} = \sqrt[T-1]{\frac{y_T}{y_1}} - 1,$$

se obtienen las tasas que figuran en la tabla siguiente.

Ingresos por capítulo	Tasa media
1 Impuestos directos	0,070
2 Impuestos indirectos	−0,127
3 Tasas y otros ingresos	0,033
4 Transferencias corrientes	0,080
5 Ingresos patrimoniales	−0,005

Como puede observarse, la mayor tasa media, 0,080, corresponde al capítulo 4, transferencias corrientes.

4.10 Un empresario dedicado a la hostelería posee hoteles en varios puntos turísticos del país. Para paliar la necesidad de personal que se le plantea durante algunos periodos del año realiza contratos temporales, seleccionando estudiantes de las escuelas de hostelería y turismo.

El número de contratos temporales que realizó durante el periodo 2001-2004 se refleja en la siguiente tabla.

	2001	2002	2003	2004
1er Trimestre	10	15	20	25
2º Trimestre	100	120	130	200
3º Trimestre	50	55	70	100
4º Trimestre	15	16	20	30

Hállese una serie de números índices en cadena del número anual de contratos para el periodo considerado.

SOLUCIÓN

Sumando las casillas de cada columna, se obtiene el número de contratos de cada año:

$$y_{01} = 10 + 100 + 50 + 15 = 175$$

$$y_{02} = 15 + 120 + 55 + 16 = 206$$

$$y_{03} = 20 + 130 + 70 + 20 = 240$$

$$y_{04} = 25 + 200 + 100 + 30 = 355,$$

con lo cual, la serie de números índices en cadena resulta:

$$I_{01}^{02} = \frac{206}{175} = 1{,}177$$

$$I_{02}^{03} = \frac{240}{206} = 1{,}165$$

$$I_{03}^{04} = \frac{355}{240} = 1{,}479.$$

4.11 La cifra de ventas de discos en un comercio aumentó entre 2001 y 2002 en un 25 por ciento y entre 2002 y 2003 en un 30 por ciento. Sin embargo, la puesta en marcha de un centro comercial en la zona redujo las ventas del producto en un 10 por ciento entre 2003 y 2004.

a) Hállese la serie de índices simples de las ventas de discos para el periodo 2001-2004 con base en el año 2001.

b) ¿Cuáles han sido las tasas de variación de las ventas en el periodo considerado?

c) Obténgase la tasa media de variación correspondiente a dicho periodo.

SOLUCIÓN

a) El índice entre 2001 y 2002 de la cantidad (vendida), Q, esto es, Q_{01}^{02}, cantidad (vendida) relativa, recoge la variación de esta magnitud entre dichos años, con lo que es igual a 1,25.

De manera análoga, la cantidad (vendida) relativa entre 2002 y 2003, Q_{02}^{03}, es igual a 1,3 y Q_{03}^{04}, cantidad (vendida) relativa entre 2003 y 2004 es 0,9.

Con estos tres índices en cadena, Q_{01}^{02}, Q_{02}^{03} y Q_{03}^{04}, y aplicando la propiedad circular de los índices simples, se obtienen los índices con base en el año 2001:

$$Q_{01}^{03} = Q_{01}^{02} \cdot Q_{02}^{03} = 1{,}3 \cdot 1{,}25 = 1{,}625$$

y

$$Q_{01}^{04} = Q_{01}^{03} \cdot Q_{03}^{04} = 0{,}9 \cdot 1{,}625 = 1{,}4625.$$

b) De la serie de índices en cadena proporcionada por el enunciado resultan las tasas de variación, ya que ambos conceptos se relacionan según la expresión genérica:

$$\dot{y}_t = I_{t-1}^{t} - 1.$$

Por tanto, las tasas de variación de las ventas en el periodo considerado han sido:

$$\dot{y}_{02} = 1{,}25 - 1 = 0{,}25$$
$$\dot{y}_{03} = 1{,}3 - 1 = 0{,}3$$
$$\dot{y}_{04} = 0{,}9 - 1 = -0{,}1.$$

c) Aplicando la relación entre la tasa media y las correspondientes tasas de variación, se obtiene la tasa media para el periodo 2001-04:

$$\text{tm} = \sqrt[4-1]{(1 + \dot{y}_{02}) \cdot (1 + \dot{y}_{03}) \cdot (1 + \dot{y}_{04})} - 1 = \sqrt[3]{(1 + 0{,}25) \cdot (1 + 0{,}3) \cdot (1 - 0{,}1)} - 1 = 0{,}135.$$

4.12 La siguiente tabla recoge el número medio por entidad de suscripciones de fondos de inversión mobiliaria, clasificadas por grupos financieros en los años 2003 y 2004 de una región.

Grupo financiero	Año 2003	Año 2004
Cajas	50 616	28 612
Bancos	45 000	94 280
Sociedades de valores	1 560	4 120
Agencias de valores	212	324
Compañías de seguros	972	2 562
Cooperativas de crédito	1 000	1 832

Se sabe, además, que el número de entidades de cada grupo es 5, 46, 12, 4, 5 y 6, respectivamente.

a) Hállense los índices simples que midan la evolución del número medio de suscripciones de cada uno de los grupos financieros entre los años 2003 y 2004.

b) Hállese un índice complejo ponderado de la evolución del número medio de suscripciones para el periodo considerado.

SOLUCIÓN

a) Para obtener la serie de números índices que miden la variación de la magnitud en cada uno de los grupos financieros entre los años considerados, aplicamos la expresión general,

$$I_{03}^{04}(i) = \frac{y_{04}(i)}{y_{03}(i)},$$

donde $y_{03}(i)$ e $y_{04}(i)$ son, respectivamente, el número medio de fondos de inversión mobiliaria del año 2003 y del año 2004 suscritos por la entidad financiera i-ésima.

De este modo, con los datos del enunciado resulta la siguiente serie de índices simples:

$$I_{03}^{04}(1) = \frac{28\,612}{50\,616} = 0{,}565$$

$$I_{03}^{04}(2) = \frac{94\,280}{45\,000} = 2{,}095$$

$$I_{03}^{04}(3) = \frac{4\,120}{1\,560} = 2{,}641$$

$$I_{03}^{04}(4) = \frac{324}{212} = 1{,}528$$

$$I_{03}^{04}(5) = \frac{2\,562}{972} = 2{,}636$$

$$I_{03}^{04}(6) = \frac{1\,832}{1\,000} = 1{,}832.$$

b) Un índice complejo ponderado de la evolución del número medio de suscripciones para el periodo 2003-2004 es

$$I_{03}^{04} = \frac{\sum_{i=1}^{N} I_{03}^{04}(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde w_i es el coeficiente de ponderación del grupo i. En este caso, parece razonable tomar como coeficiente de ponderación de cada grupo el número de entidades que lo constituyen, con lo cual, el índice complejo ponderado es

$$I_{03}^{04} = \frac{0{,}565 \cdot 5 + 2{,}095 \cdot 46 + 2{,}641 \cdot 12 + 1{,}528 \cdot 4 + 2{,}636 \cdot 5 + 1{,}832 \cdot 6}{5 + 46 + 12 + 4 + 5 + 6} = 2{,}066.$$

4.13 Sean P_1, ..., P_N las variables que denotan los precios de N bienes en los periodos t y 0.

a) Demuéstrese que el índice de precios de Laspeyres en el periodo t con base en el periodo 0 es un índice complejo ponderado con coeficiente de ponderación $w_i = p_{i0} \cdot q_{i0}$.

b) Pruébese que el índice de precios de Paasche en el periodo t con base en el periodo 0 es, también, un índice complejo ponderado, siendo el coeficiente de ponderación, para este caso, $w_i = p_{i0} \cdot q_{it}$.

SOLUCIÓN

Cuando en la expresión genérica de un índice complejo ponderado,

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde

$$I_0^t(i) = \frac{y_{it}}{y_{i0}}$$

es el índice simple de la componente i-ésima y w_i es el coeficiente de ponderación de dicho índice simple, consideramos como magnitud compleja el precio de N bienes, P_1, ..., P_i, ..., P_N, en los periodos base y actual y como coeficiente de ponderación $w_i = p_{i0} \cdot q_{i0}$, simplificando, obtenemos

$$L_0^t = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}} = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}},$$

esto es, el índice de precios de Laspeyres.

Nótese que el coeficiente de ponderación es, en este caso, el valor de la cantidad correspondiente al bien *i*-ésimo en el periodo base a precios de dicho periodo.

b) Si, por el contrario, tomamos como coeficiente de ponderación $w_i = p_{i0} \cdot q_{it}$, esto es, el valor de la cantidad del bien *i*-ésimo en el periodo actual a precios del año base, tendremos el índice de precios de Paasche:

$$P_0^t = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}} = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}}.$$

Proponemos la realización de un ejercicio análogo con los índices cuánticos de Laspeyres y Paasche.

4.14 La dirección comercial de una empresa dedicada a la venta de combustibles desea construir un índice que refleje la evolución conjunta en los últimos tres años de los tres tipos de combustibles que tiene a la venta. El precio de venta, en euros, y la cantidad, en kilolitros, de combustible vendida en dicho periodo fueron los siguientes:

Combustible	2002		2003		2004	
	Precio	Ventas	Precio	Ventas	Precio	Ventas
Tipo A	1,25	200	1,37	210	1,75	350
Tipo B	1,30	250	1,43	265	1,80	320
Tipo C	1,50	300	1,65	301	2,00	370

Calcúlense los índices de precios de Laspeyres para 2003 y 2004, con base en el año 2002.

SOLUCIÓN

El índice de Laspeyres para los precios de combustible, con base en el año 2002, es

$$L_{p02}^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i02}}{\sum_{i=1}^{N} p_{i02} \cdot q_{i02}},$$

donde p_{i02} y q_{i02} son, respectivamente, el precio y la cantidad (vendida) del combustible *i*-ésimo ($i = 1, 2, 3$) en el año 2002.

Sustituyendo p_{it} por p_{i03} y p_{i04}, precios del combustible i-ésimo en los años 2003 y 2004, se obtienen, respectivamente, los índices de Laspeyres correspondientes a dichos años:

$$L_{p02}^{03} = \frac{1{,}37 \cdot 200 + 1{,}43 \cdot 250 + 1{,}65 \cdot 300}{1{,}25 \cdot 200 + 1{,}30 \cdot 250 + 1{,}50 \cdot 300} = 1{,}099$$

y

$$L_{p02}^{04} = \frac{1{,}75 \cdot 200 + 1{,}80 \cdot 250 + 2{,}00 \cdot 300}{1{,}25 \cdot 200 + 1{,}30 \cdot 250 + 1{,}50 \cdot 300} = 1{,}366.$$

Téngase en cuenta que para el cálculo de estos índices no son necesarias las columnas de ventas de los años 2003 y 2004.

4.15 Una panadería produce cuatro tipos de panes: de centeno, de dos cereales, de avena y de maíz. El coste de las materias primas para cada tipo de pan por kilogramo es de 1, 1,5, 2 y 0,75 euros. Las ventas en los últimos años, así como el precio de venta por kilogramo, se reflejan en la siguiente tabla.

Tipo	2002		2003		2004	
	Precio de venta	Kilos vendidos	Precio de venta	Kilos vendidos	Precio de venta	Kilos vendidos
Centeno	1,10	200	1,15	228	1,20	384
Dos cereales	1,80	300	1,90	480	2,00	858
Avena	3,00	510	5,10	605	5,30	700
Maíz	1,00	809	1,20	1 000	1,50	1 500

Calcúlense los índices de producción de Laspeyres para 2003 y 2004, tomando como base 2002.

SOLUCIÓN

El índice de producción de Laspeyres con base en el año 2002 es

$$L_{q02}^{t} = \frac{\sum_{i=1}^{N} q_{it} \cdot p_{i02}}{\sum_{i=1}^{N} q_{i02} \cdot p_{i02}},$$

donde q_{i02} es la cantidad (vendida) en 2002 del tipo de pan i-ésimo ($i = 1, ..., 4$) y p_{i02} es la diferencia entre el precio de venta y el precio de coste correspondiente al tipo i-ésimo en dicho año base.

Sustituyendo q_{it} por las cantidades vendidas del tipo de pan i-ésimo en los años 2003 y 2004, esto es, q_{i03} y q_{i04}, resultan:

$$L_{q02}^{03} = \frac{228 \cdot 0{,}1 + 480 \cdot 0{,}3 + 605 \cdot 1 + 1\,000 \cdot 0{,}25}{200 \cdot 0{,}1 + 300 \cdot 0{,}3 + 510 \cdot 1 + 809 \cdot 0{,}25} = 1{,}24$$

y

$$L_{q02}^{04} = \frac{384 \cdot 0{,}1 + 858 \cdot 0{,}3 + 700 \cdot 1 + 1\,500 \cdot 0{,}25}{200 \cdot 0{,}1 + 300 \cdot 0{,}3 + 510 \cdot 1 + 809 \cdot 0{,}25} = 1{,}67,$$

índices cuánticos de Laspeyres para 2003 y 2004 con base en el año 2002.

4.16 Demuéstrese que el índice de Laspeyres cumple la propiedad de la proporcionalidad.

SOLUCIÓN

Si entre el precio del bien i-ésimo ($i = 1, ..., N$) en el año t, p_{it}, y su precio en el año t', $p_{it'}$, existe la relación:

$$p_{it'} = (1 + k)\, p_{it},$$

con k valor constante, entonces, sustituyendo $p_{it'}$ por su valor en función de p_{it} en la expresión del índice de precios de Laspeyres del año t' con base en el año 0, se tiene que

$$L_0^{t'} = \frac{\sum_{i=1}^{N} p_{it'} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}} = \frac{\sum_{i=1}^{N} (1 + k)\, p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}.$$

Poniendo fuera del sumatorio la constante $(1 + k)$, resulta:

$$L_0^{t'} = (1 + k)\, \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}},$$

esto es,

$$L_0^{t'} = (1 + k)\, L_0^t,$$

donde L_0^t es el índice de precios de Laspeyres del año t con base en el año 0, quedando así, demostrado el resultado.

4.17 Una pequeña empresa se dedica a la venta de productos lácteos. El número de unidades vendidas, en miles, y el precio por unidad, en euros, durante el año 2003, de cuatro productos fueron:

Producto	Precio	Unidades
Leche	0,90	300
Mantequilla	1,20	100
Yogur	0,20	400
Queso	5,00	125

En 2004 se incrementaron los precios de todos los productos en un 1 por ciento y las cantidades vendidas disminuyeron en un 10 por ciento. Hállense los índices de precios de Laspeyres y de Paasche de 2004 con respecto a 2003, sin transformar los datos de la tabla.

SOLUCIÓN

Puesto que en 2004 se incrementan los precios de todos los productos en un 1 por ciento, la relación entre el precio del producto i-ésimo en 2003, p_{i03}, y el precio del mismo producto en 2004, p_{i04}, es, para $i = 1, 2, 3, 4$,

$$p_{i04} = 1{,}01 \cdot p_{i03}.$$

Sustituyendo la relación anterior en la expresión genérica del índice de precios de Laspeyres,

$$L_0^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}},$$

resulta el índice de precios del Laspeyres para el año 2004 con base 2003:

$$L_{03}^{04} = \frac{\sum_{i=1}^{N} p_{i04} \cdot q_{i03}}{\sum_{i=1}^{N} p_{i03} \cdot q_{i03}} = \frac{\sum_{i=1}^{N} 1{,}01 \cdot p_{i03} \cdot q_{i03}}{\sum_{i=1}^{N} p_{i03} \cdot q_{i03}} = \frac{1{,}01 \sum_{i=1}^{N} p_{i03} \cdot q_{i03}}{\sum_{i=1}^{N} p_{i03} \cdot q_{i03}} = 1{,}01.$$

Por lo que respecta al índice de precios de Paasche,

$$P_0^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}},$$

hay que considerar, además de la variación en los precios, la variación que se produce en las cantidades entre los años 2003 y 2004. Puesto que la cantidad vendida de cada bien en 2004, q_{i04}, disminuyó un 10 por ciento con respecto a la cantidad de 2003, q_{i03}, se tiene la relación:

$$q_{i04} = 0{,}9 \cdot q_{i03},$$

para $i = 1, 2, 3, 4$.

En consecuencia, el índice de precios de Paasche para el año 2004 con base 2003 es

$$P_{03}^{04} = \frac{\sum_{i=1}^{N} p_{i04} \cdot q_{i04}}{\sum_{i=1}^{N} p_{i03} \cdot q_{i04}} = \frac{\sum_{i=1}^{N} 1{,}01 \cdot p_{i03} \cdot 0{,}90 \cdot q_{i03}}{\sum_{i=1}^{N} p_{i03} \cdot 0{,}90 \cdot q_{i03}} = \frac{1{,}01 \cdot 0{,}90 \sum_{i=1}^{N} p_{i03} \cdot q_{i03}}{0{,}90 \sum_{i=1}^{N} p_{i03} \cdot q_{i03}} = 1{,}01.$$

4.18 ¿Cuál es el índice de precios más adecuado para realizar una deflación?

SOLUCIÓN

Supongamos que la serie que se desea deflactar, expresada, por tanto, en unidades monetarias corrientes, es una serie de valores, esto es, puede descomponerse en producto de precios por cantidades. Así, el *valor* de N bienes a precios corrientes del año t-ésimo es

$$\sum_{i=1}^{N} p_{it} \cdot q_{it}.$$

Para que el valor de estos N bienes esté expresado en unidades monetarias constantes es necesario tomar como precios los correspondientes al año base, esto es, el valor para el año t-ésimo debería ser

$$\sum_{i=1}^{N} p_{i0} \cdot q_{it}.$$

El paso de una serie de valores expresados a precios corrientes a una serie de valores constantes se lleva a cabo con un índice, D_0^t, denominado deflactor, mediante la transformación:

$$\text{valor constante año } t \text{ (base 0)} = \frac{\text{valor corriente año } t}{D_0^t},$$

con lo cual, ha de buscarse el índice D_0^t para que se cumpla la igualdad:

$$\sum_{i=1}^{N} p_{i0} \cdot q_{it} = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{D_0^t}.$$

Como puede comprobar el lector fácilmente, esta relación se verifica utilizando como deflactor el índice de precios de Paasche,

$$D_0^t = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}}.$$

Aunque desde el punto de vista teórico el índice de precios de Paasche es el deflactor más adecuado, en la práctica suele utilizarse el IPC que es un índice de precios construido con la metodología de un índice de Laspeyres.

4.19 El precio medio, en euros, de los alquileres de viviendas en la zona residencial de una ciudad, así como la correspondiente serie de índices de precios en el periodo 2000-2004, ha sido:

Años	2000	2001	2002	2003	2004
Precios	800	850	900	950	1 000
Índices	100	110	112	120	120

a) ¿Cuál es el año de referencia considerado?

b) Analícese la evolución del precio medio de los alquileres en términos reales.

SOLUCIÓN

a) El año de referencia considerado es el año 2000 puesto que el índice de precios para dicho año es igual a 100.

b) Para expresar el precio del alquiler de cada año en *términos reales*, transformando precios corrientes en precios constantes, se necesita un deflactor. Ahora bien, como en el enunciado del problema no se especifica cuál ha de ser el año de referencia que debe considerarse, lo más sencillo es tomar como base el año 2000 —año base de la serie de índices—, y deflactar los precios del periodo 2000-2004, convirtiéndolos en precios constantes del año 2000. Para ello, habrá que dividir el precio de cada año por el correspondiente índice que desempeñará, entonces, el papel de deflactor. En definitiva,

$$\text{precios constantes año } t \text{ (base 2000)} = \frac{\text{precios corrientes año } t}{D^t_{00}}.$$

Aplicando la expresión anterior a la serie de precios corrientes, se tienen los precios en términos reales de cada uno de los años que aparecen en la tabla siguiente.

Años	Precios (en términos reales)
2000	800/1 = 800,00
2001	850/1,10 = 772,73
2002	900/1,12 = 803,57
2003	950/1,15 = 826,09
2004	1 000/1,2 = 833,33

Obsérvese que la expresión de conversión de precios corrientes a precios constantes requiere que el índice sea una proporción y no un porcentaje para que los precios sigan estando en las mismas unidades.

4.20 Los precios medios de una mercancía, en euros, así como los índices de precios del periodo 1999-2004, son:

Años	Precios	Índices (1999 = 100)	Índices (2002 = 100)
1999	4	100	
2000	4,5	112	
2001	5,2	115	
2002	6,3	122	100
2003	6,4		110
2004	8,1		120

Obténgase la serie de precios en términos reales con base en el año 2002.

SOLUCIÓN

Para poder expresar los datos originales a precios constantes del año 2002, mediante la relación

$$\text{precios constantes año } t \text{ (base 2002)} = \frac{\text{precios corrientes año } t}{D_{02}^{t}},$$

hay que utilizar como deflactor un índice con base en dicho año y que bien pudiera ser el índice de precios que proporciona el enunciado.

Ahora bien, explícitamente sólo se dispone de índices con base en 2002 para el periodo 2002-2004, estando los índices del periodo 1999-2002 referidos al año 1999. No obstante, para el año 2002 disponemos de dos índices, lo cual permitirá *enlazar* ambas series, refiriendo todos ellos a la misma base. De este modo, dividiendo los índices correspondientes al periodo 1999-2001, I_{99}^{t}, entre el índice del año 2002 con base en 1999, es decir, entre el *enlace técnico*, I_{99}^{02}, resulta:

$$D_{02}^{99} = \frac{I_{99}^{99}}{I_{99}^{02}} = \frac{100}{122} = 0{,}8196$$

$$D_{02}^{00} = \frac{I_{99}^{00}}{I_{99}^{02}} = \frac{112}{122} = 0{,}9180$$

$$D_{02}^{01} = \frac{I_{99}^{01}}{I_{99}^{02}} = \frac{115}{122} = 0{,}9426.$$

Hay que tener en cuenta, por un lado, que los índices hallados son proporciones, y, por otro, que para realizar el enlace entre ambas series hemos aplicado la propiedad circular a índices complejos.

Se obtiene, de este modo, una serie completa de índices con base en 2002, D_{02}^{t}, para el periodo 1999-2004: 0,8196, 0,9180, 0,9426, 1, 1,1 y 1,2.

Dividiendo los precios del periodo considerado, precios corrientes, entre los correspondientes índices, deflactores, se tiene la serie de precios en términos reales con base en el año 2002 que se recoge en la siguiente tabla.

Años	1999	2000	2001	2002	2003	2004
Precios constantes	$\frac{4}{0{,}8196} = 4{,}88$	$\frac{4{,}5}{0{,}918} = 4{,}9$	$\frac{5{,}2}{0{,}9426} = 5{,}52$	$\frac{6{,}3}{1} = 6{,}3$	$\frac{6{,}4}{1{,}1} = 5{,}82$	$\frac{8{,}1}{1{,}2} = 6{,}75$

Obsérvese que en el año 2002 los precios corrientes y constantes coinciden puesto que se trata del año base y el deflactor, que, para dicho año, es igual a 1.

4.21 Los precios por unidad de un jabón ecológico, en euros, para el periodo 1998-2004 son:

Años	Precios corrientes	Precios constantes
1998	9	7
1999	11	8
2000	13	10
2001	14	14
2002	16	19
2003	17	26
2004	20	29

a) ¿Cuál es el incremento medio anual de los precios en dicho periodo?

b) Obténgase una serie de índices con base en el año 2003.

SOLUCIÓN

a) El incremento medio anual de los precios, es decir, la tasa media de variación, entre 1998 y 2004 es

$$\text{tm} = \sqrt[7-1]{\frac{20}{9}} - 1 = 0{,}142.$$

b) La serie de precios constantes está referida al año 2001 puesto que en ese año los valores nominal y real son iguales. En consecuencia, a partir de la relación:

$$\text{precios constantes año } t \text{ (base 01)} = \frac{\text{precios corrientes año } t}{D_{01}^{t}},$$

resulta, despejando, que

$$D_{01}^{t} = \frac{\text{precios corrientes año } t}{\text{precios constantes año } t \text{ (base 01)}}.$$

Aplicando esta relación a los datos del periodo considerado, se tiene la siguiente serie de índices con base en el año 2001, D_{01}^{t}:

$$D_{01}^{98} = \frac{9}{7} = 1{,}286$$

$$D_{01}^{99} = \frac{11}{8} = 1{,}375$$

$$D_{01}^{00} = \frac{13}{10} = 1{,}3$$

$$D_{01}^{01} = \frac{14}{14} = 1$$

$$D_{01}^{02} = \frac{16}{19} = 0{,}842$$

$$D_{01}^{03} = \frac{17}{26} = 0{,}654$$

$$D_{01}^{04} = \frac{20}{29} = 0{,}689.$$

Para calcular la serie de números índices con base en el año 2003, es decir, para efectuar un cambio de base en la serie de índices anteriores, no hay más que dividir cada uno de los índices entre el enlace, que, en este caso, es el índice D_{01}^{03}:

$$I_{03}^{98} = \frac{D_{01}^{98}}{D_{01}^{03}} = \frac{1{,}286}{0{,}654} = 1{,}966$$

$$I_{03}^{99} = \frac{D_{01}^{99}}{D_{01}^{03}} = \frac{1{,}375}{0{,}654} = 2{,}102$$

$$I_{03}^{00} = \frac{D_{01}^{00}}{D_{01}^{03}} = \frac{1{,}3}{0{,}654} = 1{,}988$$

$$I_{03}^{01} = \frac{D_{01}^{01}}{D_{01}^{03}} = \frac{1}{0{,}654} = 1{,}529$$

$$I_{03}^{02} = \frac{D_{01}^{02}}{D_{01}^{03}} = \frac{0{,}842}{0{,}654} = 1{,}287$$

$$I_{03}^{03} = \frac{D_{01}^{03}}{D_{01}^{03}} = \frac{0{,}654}{0{,}654} = 1$$

$$I_{03}^{04} = \frac{D_{01}^{04}}{D_{01}^{03}} = \frac{0{,}689}{0{,}654} = 1{,}053.$$

Se obtiene, así, una serie de números índices *simples* que miden la evolución, entre cada año de la serie y el año 2003, de los índices de precios que han sido utilizados como deflactores.

4.22 La población de un país alcanzó en 1998 la cifra de 15 millones de habitantes. En la siguiente tabla figura, además, la población del país en el periodo 1999-2004 expresada como proporción de la correspondiente al año anterior.

Años	Población año t/Población año $t-1$
1999	0,90
2000	0,95
2001	0,96
2002	1,10
2003	1,05
2004	1,06

a) Calcúlese las tasas de variación del periodo considerado.

b) Determínese una serie de índices para dicho periodo con base en el año 1998.

c) Hállese el número de habitantes del país para cada año del periodo 1999-2004.

SOLUCIÓN

a) Como datos del problema se dan los cocientes:

$$\frac{POB_t}{POB_{t-1}},$$

por lo que, para hallar la tasa de variación de la población entre los periodos $t-1$ y t basta con restar una unidad a la cantidad anterior. Así,

$$\dot{POB}_t = \frac{POB_t}{POB_{t-1}} - 1.$$

Realizando esta operación para cada año, resultan las tasas de variación del periodo 1999-2004 que figuran en la tabla siguiente:

Años	1999	2000	2001	2002	2003	2004
Tasas	−0,10	−0,05	−0,04	0,10	0,05	0,06

b) A partir de la serie de índices en cadena de la población entre $t-1$ y t que proporciona el enunciado,

$$I_{t-1}^{t} = \frac{POB_t}{POB_{t-1}},$$

ha de aplicarse la propiedad circular para obtener una serie de índices fijos con base en el año 1998:

$$I_{98}^{99} = 0{,}90$$

$$I_{98}^{00} = I_{98}^{99} \cdot I_{99}^{00} = 0{,}9 \cdot 0{,}95 = 0{,}855$$

$$I_{98}^{01} = I_{98}^{00} \cdot I_{00}^{01} = 0{,}855 \cdot 0{,}96 = 0{,}821$$

$$I_{98}^{02} = I_{98}^{01} \cdot I_{01}^{02} = 0{,}821 \cdot 1{,}1 = 0{,}903$$

$$I_{98}^{03} = I_{98}^{02} \cdot I_{02}^{03} = 0{,}903 \cdot 1{,}05 = 0{,}948$$

$$I_{98}^{04} = I_{98}^{03} \cdot I_{03}^{04} = 0{,}948 \cdot 1{,}06 = 1{,}005.$$

c) Puesto que un índice expresa la variación, en este caso de la variable población, entre los años considerados, para obtener el número de habitantes del año 1999, POB_{99}, habrá que multiplicar el número de habitantes del año 1998, POB_{98}, por la población relativa (índice) entre 1998 y 1999, I_{98}^{99}. Así,

$$POB_{99} = POB_{98} \cdot I_{98}^{99} = 15 \cdot 0{,}9 = 13{,}5.$$

Operando de igual modo para el resto de los años, se completan los datos referentes a la población en el periodo 1999-2004, en millones de habitantes:

$$POB_{00} = POB_{99} \cdot I_{99}^{00} = 13{,}5 \cdot 0{,}95 = 12{,}825$$

$$POB_{01} = POB_{00} \cdot I_{00}^{01} = 12{,}825 \cdot 0{,}96 = 12{,}312$$

$$POB_{02} = POB_{01} \cdot I_{01}^{02} = 12{,}312 \cdot 1{,}1 = 13{,}543$$

$$POB_{03} = POB_{02} \cdot I_{02}^{03} = 13{,}543 \cdot 1{,}05 = 14{,}220$$

$$POB_{04} = POB_{03} \cdot I_{03}^{04} = 14{,}220 \cdot 1{,}06 = 15{,}073.$$

4.23 El precio de un bien, en euros, así como una serie de índices de precios para el periodo 2000-2004 son:

Años	2000	2001	2002	2003	2004
Precios	10	15	20	35	51
Índices	95,5	100	110,3	120	122,4

Indíquese si son verdaderas o falsas las siguientes afirmaciones:

a) El precio del bien en 2004, a precios constantes del año 2004, es 51 euros.

b) El año base de la serie de índices anterior es 2000.

c) El precio del bien en 2004, en términos reales base 2001, es de 26 euros.

d) El precio del bien en 2004, a precios constantes base 2003, es de 50 euros.

SOLUCIÓN

a) Por definición de precios constantes se cumple que el precio de un bien en un año, considerado en términos reales de ese mismo año, coincide con el precio del bien. En efecto,

$$\text{precios constantes año 04 (base 04)} = \frac{\text{precios corrientes año 04}}{D_{04}^{04}}.$$

Puesto que D_{04}^{04}, deflactor que mide la evolución de los precios entre 2004 y 2004, es igual a la unidad, se tiene que

$$\text{precios constantes año 04 (base 04)} = \frac{51}{1} = 51 \text{ euros}.$$

En definitiva, esta afirmación es *verdadera*.

b) Para que el año 2000 fuese el año base de esta serie de índices, debería cumplirse que el índice de dicho año fuera igual a 100. Como el índice del año 2000 es 95,5, la afirmación es *falsa*.

Obsérvese que el índice de 2001 es 100, por lo que éste es el año base al que está referida la serie de índices de precios.

c) Utilizando como deflactor el índice de precios del año 2004 que, según hemos comentado en el apartado anterior, tiene como año base 2001, se tiene que

$$\text{precios constantes año 04 (base 01)} = \frac{\text{precios corrientes año 04}}{D_{01}^{04}} = \frac{51}{1{,}224} = 41{,}67 \text{ euros},$$

cantidad que no coincide con 26, siendo *falsa* esta afirmación.

d) En este caso es necesario calcular:

$$\text{precios constantes año 04 (base 03)} = \frac{\text{precios corrientes año 04}}{D_{03}^{04}}.$$

Con los datos del problema el deflactor D_{03}^{04} se obtiene del cambio de base:

$$D_{03}^{04} = \frac{I_{01}^{04}}{I_{01}^{03}} = \frac{122{,}4}{120} = 1{,}02.$$

El índice simple así hallado mide la variación de los índices de precios entre 2003 y 2004.

En consecuencia,

$$\text{precios constantes año 04 (base 03)} = \frac{51}{1{,}02} = 50,$$

con lo cual, esta última afirmación es *verdadera*.

4.24 El precio de un bien, en euros, así como la serie de índices de precios para el periodo 1999-2004, ha sido:

Años	1999	2000	2001	2002	2003	2004
Precios	10	15	25	30	45	50
Índices	100	105	109,9	115,1	119,2	121,3

Indíquese cuál de las siguientes afirmaciones es cierta:

a) El año base considerado es anterior a 1999 y el precio de dicho bien en 2004, en euros de 2003, es de 50,88 euros.

b) El año base considerado es anterior a 1999 y el precio de dicho bien en 2004, en euros de 2003, es de 50 euros.

c) El año base considerado es 1999 y el precio de dicho bien en 2004, en euros de 2003, es de 49,135 euros.

d) El año base considerado es 1999 y el precio de dicho bien en 2004, en euros de 2003, es de 50 euros.

SOLUCIÓN

El año base considerado es 1999 porque es el año para el cual el índice de precios es igual a 100. A partir de aquí sólo pueden ser correctas las respuestas ***c)*** o ***d)***.

El cálculo del precio del bien en 2004, en euros de 2003, resulta de aplicar la expresión:

$$\text{precios constantes año 04 (base 03)} = \frac{\text{precios corrientes año 04}}{D_{03}^{04}}.$$

Utilizando como deflactor el índice

$$D_{03}^{04} = \frac{I_{99}^{04}}{I_{99}^{03}} = \frac{121{,}3}{119{,}2} = 1{,}0176,$$

se tiene que

$$\text{precios constantes año 04 (base 03)} = \frac{50}{1{,}0176} = 49{,}135,$$

siendo, en consecuencia, cierta la afirmación ***c)***.

4.25 En el año 2002 un ayuntamiento estableció una tarifa de 40 euros para el impuesto sobre vehículos de tracción mecánica de una cierta cilindrada. Cada año se ha revisado este impuesto en base al incremento del IPC, obteniéndose los siguientes importes:

Años	2002	2003	2004
Importes	40	44	47,08

Sabiendo que el IPC del año 2002 es 110, calcúlense los valores de dicho índice para los años 2003 y 2004.

SOLUCIÓN

Puesto que el impuesto en el año 2003, y_{03}, se ha incrementado respecto al correspondiente al año 2002, y_{02}, en función del incremento del IPC entre ambos años, se tiene la relación:

$$\frac{y_{03}}{y_{02}} = \frac{\text{IPC}_0^{03}}{\text{IPC}_0^{02}},$$

que, despejando, conduce a

$$\text{IPC}_0^{03} = \text{IPC}_0^{02} \cdot \frac{y_{03}}{y_{02}}.$$

Como, por un lado,

$$\frac{y_{03}}{y_{02}} = \frac{44}{40} = 1{,}1,$$

es decir, el incremento ha sido del 10 por ciento, y, por otro lado, el índice de precios al consumo del año 2002 es

$$\mathrm{IPC}_0^{02} = 1{,}1,$$

sustituyendo, resulta que

$$\mathrm{IPC}_0^{03} = 1{,}1 \cdot 1{,}1 = 1{,}21.$$

Además, de la relación

$$\frac{y_{03}}{y_{02}} = \frac{\mathrm{IPC}_0^{03}}{\mathrm{IPC}_0^{02}},$$

se tiene, por las propiedades de las operaciones con fracciones, que

$$\frac{y_{02}}{\mathrm{IPC}_0^{02}} = \frac{y_{03}}{\mathrm{IPC}_0^{03}},$$

lo cual equivale a

$$\text{precios constantes año 02 (base 0)} = \text{precios constantes año 03 (base 0)},$$

igualdad cuyo significado es que el hecho de que la tarifa del impuesto se haya revisado en base al incremento del IPC es equivalente a decir que *no ha sufrido variación en términos reales.*

Análogamente, el incremento del impuesto entre 2003 y 2004 es igual al incremento del IPC entre ambos años, con lo que

$$\frac{y_{04}}{y_{03}} = \frac{\mathrm{IPC}_0^{04}}{\mathrm{IPC}_0^{03}},$$

y, por consiguiente,

$$\mathrm{IPC}_0^{04} = \mathrm{IPC}_0^{03} \cdot \frac{y_{04}}{y_{03}}.$$

Puesto que, por los datos del enunciado, se obtiene que el incremento ha sido del 7 por ciento:

$$\frac{y_{04}}{y_{03}} = \frac{47{,}08}{44} = 1{,}07,$$

y, además,

$$\mathrm{IPC}_0^{03} = 1{,}21,$$

sustituyendo, resulta que el índice de precios al consumo en 2004 es

$$\text{IPC}_0^{04} = 1{,}21 \cdot 1{,}07 = 1{,}2947.$$

4.26 Sabiendo que el precio de tasación del metro cuadrado de suelo urbanizable en una zona residencial en 2004 fue de 5 000 euros y los índices de precios al consumo del país para los años 2004 y 2005 fueron 110 y 115, ¿cuál ha sido el precio de tasación en 2005, si no ha experimentado variación en términos reales?

SOLUCIÓN

El hecho de que entre 2004 y 2005 el precio de tasación no sufriera variación en términos reales significa, según hemos visto en el problema anterior, que se cumple la igualdad:

$$\text{precios constantes año 04 (base 0)} = \text{precios constantes año 05 (base 0)}.$$

Suponiendo que el deflactor utilizado ha sido, en ambos casos, el índice de precios al consumo, la relación anterior puede expresarse como

$$\frac{\text{precios corrientes año 04}}{\text{IPC}_0^{04}} = \frac{\text{precios corrientes año 05}}{\text{IPC}_0^{05}}.$$

Despejando, se tiene que

$$\text{precios corrientes año 05} = \text{precios corrientes año 04} \cdot \frac{\text{IPC}_0^{05}}{\text{IPC}_0^{04}},$$

esto es,

$$\text{precios corrientes año 05} = 5\,000 \cdot \frac{1{,}15}{1{,}1} = 5\,227{,}27 \text{ euros}.$$

Advierta el lector que, según vimos también en el problema anterior, el que el precio de tasación no haya sufrido variación en términos reales significa que se ha obtenido en base al incremento del índice de precios al consumo.

4.27 El salario medio anual de los trabajadores de un país en 2003 fue de 25 mil euros, siendo el IPC igual a 150.

a) Sabiendo que en 2004 el convenio entre trabajadores y patronal contempló un aumento salarial basado exclusivamente en el incremento del IPC y que éste fue de

un 10 por ciento respecto al del año anterior, ¿qué ingresos medios anuales percibieron los trabajadores en 2004?

b) Obténgase el valor anterior en términos reales.

SOLUCIÓN

a) Dado que el IPC del año 2004 se ha incrementado un 10 por ciento con respecto al del año 2003, y puesto que el salario de los trabajadores en 2004, y_{04}, resulta de aplicar dicho incremento al salario de 2003, y_{03}, se tiene:

$$y_{04} = y_{03} + 0{,}1 \cdot y_{03} = y_{03}\,(1 + 0{,}1) = 25 \cdot 1{,}1 = 27{,}5 \text{ miles de euros.}$$

b) Para poder expresar el salario de los trabajadores en el año 2004 a precios constantes del año 2003 se necesita un deflactor, es decir, un índice que mida la variación de los salarios entre estos dos años.

Ahora bien, la relación entre el IPC del año 2004 y el IPC del año 2003 permite escribir:

$$\text{IPC}_0^{04} = \text{IPC}_0^{03} + 0{,}1 \cdot \text{IPC}_0^{03} = 1{,}1 \cdot \text{IPC}_0^{03},$$

con lo cual, el deflactor que se utilizará es

$$\frac{\text{IPC}_0^{04}}{\text{IPC}_0^{03}} = 1{,}1,$$

índice que expresa la variación del IPC entre los años 2003 y 2004.

Así,

$$\text{precios constantes año 04 (base 03)} = \frac{\text{precios corrientes año 04}}{1{,}1},$$

es decir, el salario real de los trabajadores en 2004 con base 2003 es

$$\frac{27{,}5}{1{,}1} = 25 \text{ mil euros.}$$

El resultado obtenido es obvio puesto que el convenio entre trabajadores y patronal contempla una subida salarial basada *exclusivamente* en el incremento del IPC: eliminando el efecto de la inflación —dividiendo por 1,1—, el salario del año 2004 seguirá siendo igual al de 2003.

Obsérvese, también, que el enunciado no especifica el año base para el cálculo del salario del año 2004 en términos reales. Por ello, una solución igualmente válida pasaría por calcular:

$$\text{precios constantes año 04 (base 0)} = \frac{\text{precios corrientes año 04}}{\text{IPC}_0^{04}},$$

para lo cual sería necesario conocer el valor del IPC del año 04. Ahora bien, según se ha visto,

$$\text{IPC}_0^{04} = 1{,}1 \cdot \text{IPC}_0^{03} = 1{,}1 \cdot 150 = 165,$$

con lo que

$$\text{precios constantes año 04 (base 0)} = \frac{27{,}5}{1{,}65} = 16{,}67 \text{ miles de euros.}$$

Como el salario de los trabajadores no experimentó evolución en términos reales, entonces,

$$\text{precios constantes año 04 (base 0)} = \text{precios constantes año 03 (base 0)},$$

y el salario en términos reales en 2003 fue, también, de 16,67 miles de euros.

4.28 En 2003 el precio del billete de autobús de una ciudad era de 0,90 euros. Sabiendo que los índices de precios al consumo para los años 2003 y 2004 fueron 115 y 120, respectivamente, ¿cuánto costó el billete en 2004, si su precio en términos reales aumentó un 10 por ciento?

SOLUCIÓN

Si el precio del billete sufrió entre 2003 y 2004 un incremento del 10 por ciento en términos reales, se cumple que

$$\text{precios constantes año 04 (base 0)} = 1{,}1 \cdot \text{precios constantes año 03 (base 0)},$$

expresión que, suponiendo que el deflactor utilizado ha sido el índice de precios al consumo, puede escribirse como

$$\frac{\text{precios corrientes año 04}}{\text{IPC}_0^{04}} = 1{,}1 \cdot \frac{\text{precios corrientes año 03}}{\text{IPC}_0^{03}}.$$

Despejando de la igualdad anterior el precio del billete en 2004, se tiene que

$$\text{precios corrientes año 04} = 1{,}1 \cdot \text{precios corrientes año 03} \cdot \frac{\text{IPC}_0^{04}}{\text{IPC}_0^{03}},$$

por lo que, sustituyendo los datos del problema, resulta el precio del billete de autobús en 2004:

$$\text{precios corrientes año 04} = 1{,}1 \cdot 0{,}90 \cdot \frac{1{,}2}{1{,}15} = 1{,}033 \text{ euros.}$$

De la relación entre los precios corrientes y los índices de precios al consumo de los años 2003 y 2004 se obtiene:

$$\frac{\text{precios corrientes año 04}}{\text{precios corrientes año 03}} = 1{,}1 \cdot \frac{\text{IPC}_0^{04}}{\text{IPC}_0^{03}},$$

lo cual indica que un aumento de un 10 por ciento en términos reales es lo mismo que una revisión del precio en base a un incremento del 10 por ciento en la variación del índice de precios al consumo.

4.29 Para elaborar una serie de números índices de precios, un analista dispone de la siguiente información: en 2001 el índice de precios fue de 102; en 2002 de 104; en 2003 el incremento del índice fue de un 10 por ciento respecto al del año anterior y en 2004 dicho incremento fue de un 7 por ciento, también respecto al año anterior.

a) Calcúlese una serie de índices con base en el año 2004.

b) Hállese la tasa media de variación de la serie de índices de precios elaborada en el apartado anterior.

c) ¿Cuánto debe ser el salario de un individuo en 2004 para no perder poder adquisitivo, si en 2003 percibió 30 000 euros?

SOLUCIÓN

El enunciado del problema permite completar la serie de índices con base en el año 0 de los años 2003,

$$I_0^{03} = I_0^{02} + 0{,}1 \cdot I_0^{02} = I_0^{02} \cdot 1{,}1 = 114{,}4,$$

y 2004,

$$I_0^{04} = I_0^{03} + 0{,}07 \cdot I_0^{03} = I_0^{03} \cdot 1{,}07 = 122{,}41.$$

En resumen, la serie de índices es la que figura en la tabla siguiente:

Años	2001	2002	2003	2004
Índices base 0	102	104	114,4	122,41

a) Dividiendo cada uno de los índices anteriores, I_0^t, por el índice correspondiente al año 2004, I_0^{04},

$$\frac{I_0^t}{I_0^{04}} = \frac{I_0^t}{122{,}41},$$

y multiplicando por 100 el resultado, se obtiene una serie de índices con base en el año 2004, expresados en porcentajes:

Años	2001	2002	2003	2004
Índices base 2004	83,33	84,96	93,46	100

b) La tasa media de variación de la serie de índices es

$$\text{tm} = \sqrt[4-1]{\frac{100}{83{,}33}} - 1 = 0{,}063.$$

c) Para que no pierda poder adquisitivo, el salario del individuo en 2004 debe incrementarse en el mismo porcentaje en que se incrementan los precios, esto es, en un 7 por ciento, con lo cual, tendrá que ser

$$30\,000 \cdot 1{,}07 = 32\,100 \text{ euros}.$$

4.30 Los fondos destinados por las distintas administraciones públicas para la formación de funcionarios durante el periodo 1995-1999, expresados en millones de unidades monetarias, se reflejan en la siguiente tabla:

Distribución de fondos	1995	1996	1997	1998	1999
Admón. General de Estado	1 250	1 759	1 871	2 063	2 308
Comunidades Autónomas	1 250	1 759	1 871	2 105	2 331
Corporaciones Locales	1 250	1 759	1 871	1 765	1 797
Centrales Sindicales	750	1 523	1 470	1 577	1 711

Se sabe que los índices de precios para los años del periodo 1995-1999 han sido: 116,7, 120,5, 122,9, 124,7 y 128.

a) Calcúlese el importe de la suma total de fondos destinados a formación, en términos reales con base en el año 1999.

b) Hállese la tasa media de variación del total de fondos destinados a formación durante el periodo considerado, en términos reales con base en el año 1999.

SOLUCIÓN

a) El total de fondos destinados a formación se obtiene sumando, para cada año, los fondos de cada una de las Administraciones, es decir, Estado, Comunidades Autónomas, Corporaciones Locales y Centrales Sindicales. Así, por ejemplo, para el año 1995, dicha cantidad se calcula como

$$1\,250 + 1\,250 + 1\,250 + 750 = 4\,500.$$

Para pasar de unidades monetarias corrientes a unidades monetarias constantes, con base en el año1999, según la fórmula de conversión:

$$\text{precios constantes año } t \text{ (base 99)} = \frac{\text{precios corrientes año } t}{D_{99}^{t}},$$

puede emplearse como deflactor el índice que mide la variación entre cada uno de los índices que proporciona el enunciado, I_0^t, y el correspondiente al año 1999, I_0^{99}, es decir,

$$D_{99}^{t} = \frac{I_0^t}{I_0^{99}}.$$

Así, para cada uno de los años del periodo considerado, resulta la siguiente serie de deflactores:

$$D_{99}^{95} = \frac{I_0^{95}}{I_0^{99}} = \frac{116{,}7}{128{,}3} = 0{,}910$$

$$D_{99}^{96} = \frac{I_0^{96}}{I_0^{99}} = \frac{120{,}5}{128{,}3} = 0{,}939$$

$$D_{99}^{97} = \frac{I_0^{97}}{I_0^{99}} = \frac{122{,}9}{128{,}3} = 0{,}958$$

$$D_{99}^{98} = \frac{I_0^{98}}{I_0^{99}} = \frac{124{,}7}{128{,}3} = 0{,}972$$

$$D_{99}^{99} = \frac{I_0^{99}}{I_0^{99}} = \frac{128{,}3}{128{,}3} = 1.$$

En esta tabla se recoge, para cada año, el importe total de fondos, el deflactor utilizado y el total de fondos en términos reales con base en el año 1999. Obsérvese que las cantidades de la última fila se obtienen dividiendo las correspondientes de las dos filas anteriores.

Años	1995	1996	1997	1998	1999
Importe fondos	4 500	6 800	7 083	7 510	8 147
Deflactor (base 1999)	0,910	0,939	0,958	0,972	1
Importe fondos precios constantes de 1999	4 945,05	7 241,75	7 393,53	7 726,34	8 147

b) La tasa media de variación es

$$\text{tm} = \sqrt[5-1]{\frac{8\,147}{4\,945{,}05}} - 1 = 0{,}133.$$

4.31 El salario mínimo interprofesional de la república de Arasua durante el periodo 2002-2004, en euros, ha sido de 400, 620 y 700, para cada uno de los años del periodo. Sabiendo que la inflación es sistemática y regular, tal que el nivel de precios de un año es un 10 por ciento superior al del año anterior, calcúlese el salario mínimo interprofesional corregido de la depreciación monetaria para los años de dicho periodo.

SOLUCIÓN

La información proporcionada corresponde al salario mínimo interprofesional para los años 2002 a 2004, a precios corrientes, o, lo que es lo mismo, al salario en términos nominales. Para el cálculo del salario mínimo corregido por la depreciación monetaria, esto es, del salario en términos reales, ha de deflactarse la serie de salarios corrientes mediante la expresión:

$$\text{precios constantes año } t \text{ (base 0)} = \frac{\text{precios corrientes año } t}{D_0^t},$$

para lo cual se requiere utilizar un deflactor, D_0^t. Ahora bien, puesto que la inflación ha sido sistemática y regular, de manera que el nivel de precios de un año ha sido un 10 por ciento superior al del año anterior, tomando como año base 2002, se tienen las siguientes relaciones:

$$I_{02}^{02} = 1$$

$$I_{02}^{03} = 1{,}1 \cdot I_{02}^{02} = 1{,}1 \cdot 1 = 1{,}1$$

$$I_{02}^{04} = 1{,}1 \cdot I_{02}^{03} = 1{,}1 \cdot 1{,}1 = 1{,}21.$$

Utilizando esta serie de índices de precios como serie de deflactores puede calcularse la serie de salarios a precios constantes para este periodo, con base en el año 2002. El resultado de aplicar la fórmula de conversión de precios corrientes a constantes para los diferentes años, queda recogido en la siguiente tabla:

Años	Precios constantes (base 2002)
2002	400/1,0 = 400
2003	620/1,1 = 563,64
2004	700/1,21 = 578,51

4.32 El precio de un modelo de teléfono móvil, en euros, así como la serie de índices de precios para el periodo 1999-2004, ha sido:

Años	1999	2000	2001	2002	2003	2004
Precios	100	150	250	300	450	500
Índices	100	105	109,9	115,1	119,2	121,3

Calcúlense:

a) Las tasas de variación de los precios de dicho bien.

b) La tasa media de variación de los precios, expresados éstos en términos constantes.

SOLUCIÓN

a) Aplicando la expresión de tasa de variación entre los periodos $t-1$ y t,

$$\dot{y}_t = \frac{y_t}{y_{t-1}} - 1,$$

a cada uno de los años del periodo considerado resultan las siguientes tasas:

$$\dot{y}_{00} = \frac{150}{100} - 1 = 0{,}50$$

$$\dot{y}_{01} = \frac{250}{150} - 1 = 0{,}67$$

$$\dot{y}_{02} = \frac{300}{250} - 1 = 0{,}20$$

$$\dot{y}_{03} = \frac{450}{300} - 1 = 0{,}50$$

$$\dot{y}_{04} = \frac{500}{450} - 1 = 0{,}11.$$

b) Para obtener la tasa media de variación,

$$\text{tm} = \sqrt[6-1]{\frac{y'_{04}}{y'_{99}}} - 1,$$

es necesario conocer y'_{04} e y'_{99}, precios en términos constantes. Puesto que no se especifica el año base respecto al cual referir los precios, lo más sencillo es utilizar como deflactor el índice de precios que proporciona el enunciado y cuya base es el año 1999.

Ahora bien, y'_{99} es igual a 100, precio del bien en 1999, ya que éste es el año base considerado.

Por otro lado,

$$\text{precios constantes año 04 (base 99)} = \frac{\text{precios corrientes año 04}}{D_{99}^{04}},$$

esto es,

$$y'_{04} = \frac{500}{1{,}213} = 412{,}2.$$

En definitiva, la tasa media de variación de los precios en términos constantes con base en el año 1999 es

$$\text{tm} = \sqrt[5]{\frac{412{,}2}{100}} - 1 = 0{,}327.$$

4.33 Dada la tasa de variación de la variable Y entre los periodos $t-1$ y t, $\dot{y}_t$, hállese la tasa de variación de la variable

$$Z = \frac{1}{Y},$$

en función de $\dot{y}_t$.

SOLUCIÓN

La tasa de variación de la variable Z entre los periodos $t-1$ y t es, por definición,

$$\dot{z}_t = \frac{z_t}{z_{t-1}} - 1.$$

Ahora bien, como

$$z_t = \frac{1}{y_t}$$

y

$$z_{t-1} = \frac{1}{y_{t-1}},$$

sustituyendo las relaciones anteriores en la expresión de la tasa de variación de la variable Z, se tiene que

$$\dot{z}_t = \frac{1/y_t}{1/y_{t-1}} - 1 = \frac{y_{t-1}}{y_t} - 1,$$

esto es,

$$1 + \dot{z}_t = \frac{y_{t-1}}{y_t}.$$

Por otro lado, a partir de la expresión genérica de la tasa de variación,

$$\dot{y}_t = \frac{y_t}{y_{t-1}} - 1,$$

puede escribirse:

$$1 + \dot{y}_t = \frac{y_t}{y_{t-1}},$$

es decir,

$$\frac{1}{1 + \dot{y}_t} = \frac{y_{t-1}}{y_t}.$$

En consecuencia, comparando igualdades, resulta que

$$1 + \dot{z}_t = \frac{1}{1 + \dot{y}_t},$$

y, por tanto, la tasa de variación de Z en función de la tasa de variación de Y resulta ser:

$$\dot{z}_t = \frac{1}{1 + \dot{y}_t} - 1.$$

4.34 Dadas las variables Y, U y V, tales que

$$Y = U \cdot V,$$

obténgase la tasa de variación de la variable Y entre los periodos $t-1$ y t, $\dot{y}_t$, en función de $\dot{u}_t$ y $\dot{v}_t$, tasas de variación de U y V entre los periodo $t-1$ y t, respectivamente.

SOLUCIÓN

De la definición de tasa de variación de una variable entre dos periodos consecutivos resultan las siguientes expresiones para las variables Y, U y V:

$$y_t = (1 + \dot{y}_t)\, y_{t-1},$$

$$u_t = (1 + \dot{u}_t)\, u_{t-1}$$

y

$$v_t = (1 + \dot{v}_t)\, v_{t-1}.$$

Considerando las igualdades anteriores, la relación

$$y_t = u_t \cdot v_t,$$

es equivalente, sustituyendo, a

$$(1 + \dot{y}_t)\, y_{t-1} = (1 + \dot{u}_t)\, u_{t-1}\, (1 + \dot{v}_t) v_{t-1},$$

esto es, a

$$1 + \dot{y}_t = (1 + \dot{u}_t) \cdot (1 + \dot{v}_t),$$

ya que

$$y_{t-1} = u_{t-1} \cdot v_{t-1}.$$

En definitiva, la tasa de variación de Y en función de las tasas de variación de U y V es

$$\dot{y}_t = (1 + \dot{u}_t) \cdot (1 + \dot{v}_t) - 1.$$

4.35 Dadas las variables Y, U y V, tales que

$$Y = \frac{U}{V},$$

obténgase la tasa de variación de Y entre los periodos $t-1$ y t, $\dot{y}_t$, en función de las tasas de variación de U y V, $\dot{u}_t$ y $\dot{v}_t$, respectivamente.

SOLUCIÓN

Haciendo el cambio de variable

$$Z = \frac{1}{V},$$

se tiene que

$$Y = \frac{U}{V} = U \cdot Z.$$

Aplicando la solución del problema anterior, resulta:

$$\dot{y}_t = (1 + \dot{u}_t) \cdot (1 + \dot{z}_t) - 1.$$

Ahora bien, según demostramos en el problema **4.33**,

$$\dot{z}_t = \frac{1}{1 + \dot{v}_t} - 1,$$

con lo cual, sustituyendo en la igualdad anterior,

$$\dot{y}_t = (1 + \dot{u}_t) \cdot \left(1 + \frac{1}{1 + \dot{v}_t} - 1\right) - 1 = \frac{1 + \dot{u}_t}{1 + \dot{v}_t} - 1,$$

o, equivalentemente,

$$\dot{y}_t = \frac{1 + \dot{u}_t - 1 - \dot{v}_t}{1 + \dot{v}_t} = \frac{\dot{u}_t - \dot{v}_t}{1 + \dot{v}_t}.$$

4.36 Dada la tasa media de variación de la variable Y en un cierto periodo, $\text{tm}(Y)$, hállese la tasa media de variación de la variable

$$Z = \frac{1}{Y}.$$

SOLUCIÓN

La tasa media de variación de la variable Z es, por definición,

$$\text{tm}(Z) = \sqrt[T-1]{\frac{z_T}{z_1}} - 1,$$

expresión que, tras sencillas operaciones, se convierte en

$$[\text{tm}(Z) + 1]^{T-1} = \frac{z_T}{z_1}.$$

Sustituyendo las observaciones de la variable Z en función de las observaciones de la variable Y, se tiene que

$$[\text{tm}(Z) + 1]^{T-1} = \frac{1/y_T}{1/y_1} = \frac{y_1}{y_T}.$$

Teniendo en cuenta que la tasa media de variación de la variable Y también verifica la relación:

$$[\text{tm}(Y) + 1]^{T-1} = \frac{y_T}{y_1},$$

entonces, se cumple la igualdad:

$$[\text{tm}(Z) + 1]^{T-1} = \frac{1}{[\text{tm}(Y) + 1]^{T-1}}.$$

En consecuencia, extrayendo la raíz $T-1$-ésima de ambos miembros, se tiene que

$$\text{tm}(Z) + 1 = \frac{1}{\text{tm}(Y) + 1},$$

esto es,

$$\text{tm}(Z) = \frac{1}{\text{tm}(Y) + 1} - 1.$$

4.37 Obténgase la tasa media de variación para un cierto periodo de la variable

$$Y = U \cdot V,$$

en función de $\text{tm}(U)$ y $\text{tm}(V)$, tasas medias de variación de las variables U y V para el mismo periodo, respectivamente.

SOLUCIÓN

Puesto que la variable Y es el producto de las variables U y V, la relación

$$[\text{tm}(Y) + 1]^{T-1} = \frac{y_T}{y_1}$$

puede escribirse como

$$[\text{tm}(Y) + 1]^{T-1} = \frac{u_T \cdot v_T}{u_1 \cdot v_1} = \frac{u_T}{u_1} \cdot \frac{v_T}{v_1}.$$

Aplicando la definición de tasa media de variación a las variables U y V, resulta:

$$[\text{tm}(Y) + 1]^{T-1} = [\text{tm}(U) + 1]^{T-1} \cdot [\text{tm}(V) + 1]^{T-1},$$

ya que

$$[\text{tm}(U) + 1]^{T-1} = \frac{u_T}{u_1}$$

y

$$[\text{tm}(V) + 1]^{T-1} = \frac{v_T}{v_1}.$$

Por último, extrayendo la raíz $T-1$-ésima, se tiene que

$$[\text{tm}(Y) + 1] = [\text{tm}(U) + 1] \cdot [\text{tm}(V) + 1],$$

es decir,

$$\text{tm}(Y) = [\text{tm}(U) + 1] \cdot [\text{tm}(V) + 1] - 1.$$

4.38 Hállese la tasa media de variación para un cierto periodo de la variable

$$Y = \frac{U}{V},$$

a partir de las tasas medias de variación de U y V, para el mismo periodo.

SOLUCIÓN

Este ejercicio es consecuencia inmediata de los problemas **4.36** y **4.37**. En efecto, haciendo el cambio de variable

$$Z = \frac{1}{V},$$

se tiene que

$$Y = \frac{U}{V} = U \cdot Z,$$

por lo que, considerando el resultado del problema **4.37**, resulta:

$$\text{tm}(Y) = [\text{tm}(U) + 1] \cdot [\text{tm}(Z) + 1] - 1.$$

En cuanto a la tasa media de variación de la variable Z, por aplicación del problema **4.36** se tiene que

$$\text{tm}(Z) = \frac{1}{\text{tm}(V) + 1} - 1.$$

Sustituyendo en la expresión de $\text{tm}(Y)$ antes hallada resulta:

$$\text{tm}(Y) = [\text{tm}(U) + 1] \cdot \left(\frac{1}{\text{tm}(V) + 1} - 1 + 1\right) - 1 = \frac{\text{tm}(U) + 1}{\text{tm}(V) + 1} - 1,$$

o, lo que es igual,

$$\text{tm}(Y) = \frac{\text{tm}(U) - \text{tm}(V)}{\text{tm}(V) + 1}.$$

4.39 Las tasas anuales de variación del PIB y de la población de un país en el periodo 1998-2004 han sido:

Años	Tasa de crecimiento PIB	Tasa de crecimiento población
1999	2,4	0,3
2000	2,5	0,2
2001	2,3	0,1
2002	2,5	0
2003	2,6	0,1
2004	2,7	0,1

a) Calcúlense las tasas medias anuales de variación del PIB y de la población para el periodo considerado.

b) Hállense las tasas de variación del PIB per cápita para dicho periodo.

c) Obténgase la tasa media de variación del PIB per cápita para el periodo 1993-1999.

SOLUCIÓN

a) La expresión que relaciona las tasas de variación con la tasa media permite obtener ésta para las variables *PIB* y población, *POB*. Así, teniendo en cuenta que el periodo de cálculo

de las tasas medias es 1998-2004 y que, por tanto, el orden de las raíces es $7 - 1 = 6$, resulta que

$$\text{tm}(PIB) = \sqrt[7-1]{(1 + 2{,}4) \cdot (1 + 2{,}5) \cdot (1 + 2{,}3) \cdot (1 + 2{,}5) \cdot (1 + 2{,}6) \cdot (1 + 2{,}7)} - 1 = 2{,}498$$

es el incremento medio del PIB para el periodo 1993-1999 y

$$\text{tm}(POB) = \sqrt[7-1]{(1 + 0{,}3) \cdot (1 + 0{,}2) \cdot (1 + 0{,}1) \cdot (1 + 0) \cdot (1 + 0{,}1) \cdot (1 + 0{,}1)} - 1 = 0{,}129$$

es la tasa media anual de la población para el mismo periodo.

b) La variable PIB per cápita, *PIBC*, es el cociente entre las variables *PIB* y población, *POB,*

$$PIBC = \frac{PIB}{POB},$$

Por ello, según se demuestra en el problema **4.35**, la tasa de variación de esta variable entre los periodos $t - 1$ y t, $\dot{PIBC}_t$, se calcula como

$$\dot{PIBC}_t = \frac{\dot{PIB}_t - \dot{POB}_t}{1 + \dot{POB}_t},$$

donde $\dot{PIB}_t$ y $\dot{POB}_t$ son, respectivamente, las tasas de variación del PIB y de la población entre los periodos $t - 1$ y t.

Los resultados obtenidos de la aplicación de la expresión anterior para los años del periodo considerado aparecen en la tabla siguiente.

Años	Tasa de crecimiento PIB per cápita
1999	1,615
2000	1,917
2001	2,000
2002	2,500
2003	2,273
2004	2,364

donde, por ejemplo,

$$\dot{PIBC}_{02} = \frac{\dot{PIB}_{02} - \dot{POB}_{02}}{1 + \dot{POB}_{02}} = \frac{2{,}5 - 0}{1 + 0} = 2{,}5.$$

c) La tasa media de variación del PIB per cápita, *PIBC*, se obtiene de las tasas de variación anteriores:

$$\text{tm}(PIBC) = \sqrt[7-1]{(1 + 1{,}615) \cdot (1 + 1{,}917) \cdot (1 + 2) \cdot (1 + 2{,}5) \cdot (1 + 2{,}273) \cdot (1 + 2{,}364)} - 1 = 2{,}097.$$

Se llega al mismo resultado utilizando las tasas medias anuales de variación del PIB y de la población. En efecto, mediante lo demostrado en el problema **4.38**:

$$\text{tm}(PIBC) = \frac{\text{tm}(PIB) - \text{tm}(POB)}{\text{tm}(POB) + 1},$$

y, con los datos calculados en el apartado ***a),*** se obtiene idéntico valor para la tasa media.

4.40 Se dispone de la siguiente serie de índices de precios de un modelo de coche para el periodo 1998-2004.

Años	Índices (1999 = 100)	Índices (2001 = 100)
1998	95	
1999	100	
2000	105	
2001	108	100
2002		105
2003		110
2004		112

Se sabe que la tasa media anual del precio en el periodo 1998-2004 es 0,23. Obténgase la tasa media anual del precio en términos reales con base en el año 2003.

SOLUCIÓN

La variable precio real, P_R, es el cociente entre la variable precio nominal, P_N, y el deflactor, D_{03}, esto es, el índice con base, en este caso, en el año 2003[1]:

$$P_R = \frac{P_N}{D_{03}}.$$

[1] Se han eliminado los subíndices correspondientes a los años corriente y de referencia para evitar complicaciones en la notación.

Esta relación permite aplicar el resultado del problema **4.38** para calcular la tasa media de la variable P_R. Así,

$$\mathrm{tm}(P_R) = \frac{\mathrm{tm}(P_N) - \mathrm{tm}(D_{03})}{\mathrm{tm}(D_{03}) + 1}.$$

Puesto que el enunciado proporciona la tasa media del precio, hay que calcular la correspondiente a la variable deflactor:

$$\mathrm{tm}(D_{03}) = \sqrt[7-1]{\frac{D_{03}^{04}}{D_{03}^{98}}} - 1.$$

Ahora bien, por un lado, la serie de índices con base en 2001 permite hallar el deflactor:

$$D_{03}^{04} = \frac{I_{01}^{04}}{I_{01}^{03}} = \frac{112}{110} = 1{,}0182,$$

y, por otro lado, para obtener

$$D_{03}^{98} = \frac{I_{01}^{98}}{I_{01}^{03}},$$

se empleará el índice I_{99}^{98} correspondiente a la primera serie de índices y el enlace técnico, I_{99}^{01}. De este modo, el numerador de la expresión anterior[2] es

$$I_{01}^{98} = \frac{I_{99}^{98}}{I_{99}^{01}} = \frac{95}{108} = 0{,}8796,$$

con lo cual,

$$D_{03}^{98} = \frac{87{,}96}{110} = 0{,}7996.$$

Puede calcularse, entonces, la tasa media de la variable deflactor:

$$\mathrm{tm}(D_{03}) = \sqrt[6]{\frac{1{,}0182}{0{,}7996}} - 1 = 0{,}041.$$

En definitiva, la tasa media anual del precio en términos reales con base en el año 2003 es

$$\mathrm{tm}(P_R) = \frac{0{,}23 - 0{,}041}{0{,}041 + 1} = 0{,}181.$$

[2] Como habrá advertido el lector, este índice, obtenido por aplicación a índices complejos de la propiedad circular, no debería denotarse de igual modo que éstos; sin embargo, nos permitimos esta licencia para que la notación no resulta engorrosa.

4.41 Demuéstrese que la suma de las repercusiones absolutas de las componentes de un índice complejo ponderado,

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

es igual a la variación absoluta del índice.

SOLUCIÓN

La variación absoluta de un índice complejo ponderado entre los periodos $t-1$ y t es, por definición, la diferencia entre los valores del índice en dichos periodos:

$$\Delta I_0^t = I_0^t - I_0^{t-1},$$

es decir,

$$\Delta I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i} - \frac{\sum_{i=1}^{N} I_0^{t-1}(i) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} [I_0^t(i) - I_0^{t-1}(i)]\, w_i}{\sum_{i=1}^{N} w_i}.$$

Ahora bien, $I_0^t(i) - I_0^{t-1}(i)$ es la variación absoluta de la componente i-ésima del índice complejo ponderado, $\Delta I_0^t(i)$, con lo cual,

$$\Delta I_0^t = \frac{\sum_{i=1}^{N} \Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i} = \sum_{i=1}^{N} \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i}.$$

Como puede observarse, cada sumando del último sumatorio es, por definición, la repercusión absoluta de cada una de las componentes, por lo que

$$\Delta I_0^t = \sum_{i=1}^{N} R_i,$$

según quería demostrarse.

4.42 Demuéstrese que la repercusión absoluta de la componente i-ésima sobre la variación absoluta de un índice complejo ponderado entre $t-1$ y t, dividida entre el valor del índice en $t-1$ es igual a la repercusión relativa sobre la variación relativa de dicho índice entre $t-1$ y t.

SOLUCIÓN

Dado el índice complejo ponderado,

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

la repercusión absoluta de la componente i-ésima sobre la variación absoluta del índice entre $t-1$ y t es, por definición,

$$R_i = \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i}.$$

Dividiendo R_i por el valor del índice en el periodo $t-1$ se tiene:

$$\frac{R_i}{I_0^{t-1}} = \frac{\dfrac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i}}{\dfrac{\sum_{i=1}^{N} I_0^{t-1}(i) \cdot w_i}{\sum_{i=1}^{N} w_i}} = \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} I_0^{t-1}(i) \cdot w_i},$$

expresión que se corresponde, en efecto, con la repercusión relativa de la componente i-ésima sobre la variación relativa del índice, r_i.

4.43 Demuéstrese que la suma de las repercusiones relativas del índice complejo ponderado

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i}$$

es igual a su tasa de variación.

SOLUCIÓN

La tasa de variación del índice I_0^t entre los periodos $t-1$ y t es, por definición,

$$\dot{I}_0^t = \frac{\Delta I_0^t}{I_0^{t-1}} = \frac{I_0^t - I_0^{t-1}}{I_0^{t-1}},$$

con lo cual,

$$\dot{I}_0^t = \frac{\dfrac{\sum_{i=1}^{N} I_0^t(i)\cdot w_i}{\sum_{i=1}^{N} w_i} - \dfrac{\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i}{\sum_{i=1}^{N} w_i}}{\dfrac{\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i}{\sum_{i=1}^{N} w_i}} = \frac{\sum_{i=1}^{N} [I_0^t(i) - I_0^{t-1}(i)]\, w_i}{\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i} = \frac{\sum_{i=1}^{N} \Delta I_0^t(i)\cdot w_i}{\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i},$$

es decir,

$$\dot{I}_0^t = \sum_{i=1}^{N} \frac{\Delta I_0^t(i)\cdot w_i}{\sum_{i=1}^{N} I_0^{t-1}(i)]\, w_i} = \sum_{i=1}^{N} r_i,$$

según queríamos demostrar.

También puede realizarse la comprobación teniendo en cuenta los resultados obtenidos en los dos ejercicios anteriores. Así, por un lado, la suma de las repercusiones relativas es

$$\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} \frac{R_i}{I_0^{t-1}} = \frac{1}{I_0^{t-1}} \sum_{i=1}^{N} R_i.$$

Pero, por otro lado, la suma de las repercusiones absolutas, $\sum_{i=1}^{N} R_i$, es igual a la variación absoluta del índice, ΔI_0^t, con lo que la expresión anterior se convierte en

$$\sum_{i=1}^{N} r_i = \frac{\Delta I_0^t}{I_0^{t-1}} = \dot{I}_0^t,$$

quedando, así, demostrado el resultado.

4.44 Dado un índice complejo ponderado, demuéstrese que la participación de la componente i-ésima sobre la variación del índice entre los periodos $t-1$ y t es igual al

cociente entre la repercusión absoluta de dicha componente y la variación absoluta del índice entre los periodos $t-1$ y t.

SOLUCIÓN

Por definición, la participación de la componente i-ésima sobre la variación de un índice complejo ponderado es el cociente entre su repercusión relativa y la tasa de variación del índice:

$$P_i = \frac{r_i}{I_0^t} = \frac{\dfrac{\Delta I_0^t(i)\cdot w_i}{\displaystyle\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i}}{\dfrac{\displaystyle\sum_{i=1}^{N}\Delta I_0^t(i)\cdot w_i}{\displaystyle\sum_{i=1}^{N} I_0^{t-1}(i)\cdot w_i}} = \frac{\Delta I_0^t(i)\cdot w_i}{\displaystyle\sum_{i=1}^{N}\Delta I_0^t(i)\cdot w_i}.$$

Ahora bien, dividiendo numerador y denominador de la expresión anterior por $\displaystyle\sum_{i=1}^{N} w_i$, se obtiene que la participación de la componente i-ésima es

$$P_i = \frac{\dfrac{\Delta I_0^t(i)\cdot w_i}{\displaystyle\sum_{i=1}^{N} w_i}}{\dfrac{\displaystyle\sum_{i=1}^{N}\Delta I_0^t(i)\cdot w_i}{\displaystyle\sum_{i=1}^{N} w_i}} = \frac{R_i}{\Delta I_0^t},$$

esto es, el cociente entre la repercusión absoluta y la variación absoluta del índice.

4.45 Demuéstrese que la suma de las participaciones de las componentes de un índice complejo ponderado, expresadas en porcentajes, es igual a 100.

SOLUCIÓN

El resultado probado en el ejercicio anterior permite expresar la suma de las participaciones, en tanto por ciento, como

$$\sum_{i=1}^{N} P_i = \sum_{i=1}^{N}\frac{R_i}{\Delta I_0^t}\cdot 100 = \frac{1}{\Delta I_0^t}\sum_{i=1}^{N} R_i\cdot 100.$$

Ahora bien, puesto que la suma de las repercusiones absolutas es igual a la variación absoluta del índice,

$$\Delta I_0^t = \sum_{i=1}^{N} R_i,$$

se deduce de manera inmediata el resultado.

Téngase en cuenta que se habría llegado a la misma conclusión, considerando que, según vimos también en **4.44**, la participación de la componente *i*-ésima puede expresarse como

$$P_i = \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} \Delta I_0^t(i) \cdot w_i}.$$

4.46 Obténganse la participación y las repercusiones absoluta y relativa del precio del bien *i*-ésimo en la variación del índice de precios de Laspeyres entre los periodos $t-1$ y t.

SOLUCIÓN

El índice de precios de Laspeyres es un índice complejo ponderado de la forma

$$L_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde $I_0^t(i) = p_{it}/p_{i0}$ y $w_i = p_{i0} \cdot q_{i0}$.

Por definición, la repercusión absoluta del bien *i*-ésimo sobre la variación absoluta entre los periodos $t-1$ y t de un índice complejo ponderado es

$$R_i = \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{[I_0^t(i) - I_0^{t-1}(i)]\, w_i}{\sum_{i=1}^{N} w_i},$$

con lo cual, para el índice de precios de Laspeyres, se tiene que

$$R_i = \frac{\left(\frac{p_{it}}{p_{i0}} - \frac{p_{it-1}}{p_{i0}}\right) p_{i0} \cdot q_{i0}}{\sum_{i=1}^{N} p_i \cdot q_{i0}} = \frac{(p_{it} - p_{it-1}) q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}} = \frac{\Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}.$$

Análogamente, el cociente

$$r_i = \frac{\Delta I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} I_0^{t-1}(i) \cdot w_i}$$

es la repercusión relativa del bien i-ésimo sobre la variación relativa del índice complejo ponderado entre los periodos $t-1$ y t. Aplicando esta definición al índice de precios de Laspeyres, se obtiene:

$$r_i = \frac{\left(\frac{p_{it}}{p_{i0}} - \frac{p_{it-1}}{p_{i0}}\right) p_{i0} \cdot q_{i0}}{\sum_{i=1}^{N} \frac{p_{it-1}}{p_{i0}} \cdot p_{i0} \cdot q_{i0}} = \frac{(p_{it} - p_{it-1}) q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}} = \frac{\Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}}.$$

Por último, la participación del bien i-ésimo en la variación del índice es el cociente entre la repercusión anterior y la tasa de variación del índice de Laspeyres. Ahora bien, la tasa de variación de índice de precios de Laspeyre entre los periodos $t-1$ y t es

$$\dot{L}_0^t = \frac{L_0^t - L_0^{t-1}}{L_0^{t-1}} = \frac{\frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}} - \frac{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}}{\frac{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}} = \frac{\sum_{i=1}^{N} (p_{it} - p_{it-1}) q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}} = \frac{\sum_{i=1}^{N} \Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}},$$

con lo cual, la participación del bien i-ésimo es

$$P_i = \frac{\frac{\Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}}}{\frac{\sum_{i=1}^{N} \Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{it-1} \cdot q_{i0}}} = \frac{\Delta p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} \Delta p_{it} \cdot q_{i0}}.$$

4.47 El 50 por ciento de los trabajadores de una cadena de montaje para la fabricación de piezas de automóvil pertenece al turno de mañana, el 35 por ciento al turno de tarde y el resto al turno de noche. Se dispone de información sobre el número medio de unidades producidas por hora y por trabajador para los años 2002, 2003 y 2004, en cada uno de los turnos.

Años	Turno mañana	Turno tarde	Turno noche
2002	79	75	54
2003	80	77	56
2004	84	78	58

a) Hállense los índices complejos ponderados de 2003 y 2004, con base en el año 2002, que midan la evolución del número medio de unidades producidas por hora y trabajador en toda la cadena de montaje.

b) Obténgase la repercusión absoluta de cada uno de los turnos en la variación del índice entre los años 2003 y 2004, así como su participación porcentual en la misma.

c) ¿Cuál es la repercusión relativa de cada turno de trabajadores en la variación del índice entre 2003 y 2004?

SOLUCIÓN

a) El índice complejo ponderado del año t con base en el año 2002 es

$$I_{02}^{t} = \frac{\sum_{i=1}^{N} I_{02}^{t}(i) \cdot w_i}{\sum_{i=1}^{N} w_i} = \sum_{i=1}^{N} I_{02}^{t}(i) \cdot \frac{w_i}{\sum_{i=1}^{N} w_i},$$

donde $I_{02}^{t}(i)$ es el índice simple que mide la variación entre los años 2002 y t del número medio de unidades producidas por hora y por trabajador en el turno i-ésimo y $w_i / \sum_{i=1}^{N} w_i$ es la ponderación de dicho turno en tanto por uno.

En la siguiente tabla se recogen los índices simples de los años 2003 y 2004, con base en el año 2002, expresados en porcentajes, así como la ponderación correspondiente a cada turno, esto es, la proporción que representa sobre el total.

Turnos	Turno de mañana	Turno de tarde	Turno de noche
Índice simples 2003	(80/79)100 = 101,3	(77/75)100 = 102,7	(56/54)100 = 103,7
Índices simples 2004	(84/79)100 = 106,3	(78/75)100 = 104	(58/54)100 = 107,4
Ponderaciones	0,5	0,35	0,15

A partir de los datos de la tabla anterior, calculamos los índices *compuestos* para los años 2003 y 2004, con base en el año 2002:

$$I_{02}^{03} = 101{,}3 \cdot 0{,}5 + 102{,}7 \cdot 0{,}35 + 103{,}7 \cdot 0{,}15 = 102{,}15$$

e

$$I_{02}^{04} = 106{,}3 \cdot 0{,}5 + 104 \cdot 0{,}35 + 107{,}4 \cdot 0{,}15 = 105{,}66.$$

b) La variación absoluta del índice que mide la evolución del número medio de unidades producidas por hora y trabajador en toda la cadena de montaje entre los años 2003 y 2004, con base en el año 2002, es

$$\Delta I_{02}^{04} = I_{02}^{04} - I_{02}^{03} = 105{,}66 - 102{,}15 = 3{,}51.$$

¿Cuál es la variación absoluta del índice simple de cada turno entre 2003 y 2004? La respuesta es sencilla: la diferencia entre los índices simples de cada uno de dichos años, esto es,

$$\Delta I_{02}^{04}(i) = I_{02}^{04}(i) - I_{02}^{03}(i).$$

Los resultados de aplicar esta expresión a cada uno de los turnos de trabajo se recogen en la última fila de la tabla:

Turnos	Turno de mañana	Turno de tarde	Turno de noche
Ponderaciones	0,5	0,35	0,15
Variación absoluta índices simples	106,3 − 101,3 = 5	104 − 102,7 = 1,3	107,4 − 103,7 = 3,7

Multiplicando la variación absoluta del índice simple de cada turno por su respectiva ponderación, en tanto por uno, resulta la *repercusión absoluta de cada turno sobre la variación absoluta del índice complejo entre los años 2003 y 2004*, según la expresión genérica:

$$R_i = \frac{\Delta I_{02}^{04}(i) \cdot w_i}{\sum_{i=1}^{N} w_i} = \Delta I_{02}^{04}(i) \cdot \frac{w_i}{\sum_{i=1}^{N} w_i},$$

con lo cual,

$$R_1 = \Delta I_{02}^{04}(1) \cdot \frac{w_1}{\sum_{i=1}^{N} w_i} = 5 \cdot 0{,}5 = 2{,}5,$$

$$R_2 = \Delta I_{02}^{04}(2) \cdot \frac{w_2}{\sum_{i=1}^{N} w_i} = 1{,}3 \cdot 0{,}35 = 0{,}455$$

y

$$R_3 = \Delta I_{02}^{04}(3) \cdot \frac{w_3}{\sum_{i=1}^{N} w_i} = 3{,}7 \cdot 0{,}15 = 0{,}555,$$

cantidades cuya suma es igual a la variación absoluta del índice complejo.

La *participación relativa porcentual de cada turno sobre la variación del índice complejo* responde a la expresión general

$$P_i = \frac{R_i}{\Delta I_{02}^{04}} \cdot 100,$$

con lo cual,

$$P_1 = \frac{R_1}{\Delta I_{02}^{04}} \cdot 100 = \frac{2{,}5}{3{,}51} \cdot 100 = 71{,}225,$$

$$P_2 = \frac{R_2}{\Delta I_{02}^{04}} \cdot 100 = \frac{0{,}455}{3{,}51} \cdot 100 = 12{,}963$$

y

$$P_3 = \frac{R_3}{\Delta I_{02}^{04}} \cdot 100 = \frac{0{,}555}{3{,}51} \cdot 100 = 15{,}812,$$

siendo la suma de las participaciones relativas igual a 100.

En la siguiente tabla figuran las repercusiones absolutas y las participaciones, ambas en porcentajes, de cada turno en la variación absoluta del índice:

Turnos	Turno de mañana	Turno de tarde	Turno de noche
Repercusiones absolutas	2,5	0,455	0,555
Participaciones	71,225	12,963	15,812

Concluimos, así, que el índice complejo que mide la evolución del número medio de unidades producidas por hora y trabajador en toda la cadena de montaje ha sufrido una variación de un 3,51 por ciento, teniendo el turno de mañana una repercusión en este aumento igual a 2,5 por ciento, lo que supone una participación en términos relativos del 71,225 por ciento; asimismo, la repercusión del turno de tarde en el aumento del índice ha sido de 0,455 por ciento, que constituye una participación relativa del 12,963 por ciento; finalmente, el turno de noche ha repercutido con un 0,555 por ciento en la variación absoluta del índice complejo, esto es, con una participación porcentual de 15,812.

c) Según vimos en **4.42**, dividiendo la repercusión absoluta de cada turno por el índice complejo del año 2003 y multiplicando por 100 el resultado, se obtiene la *repercusión relativa de cada turno sobre la variación del índice*, expresada en porcentajes,

$$r_i = \frac{R_i}{I_{02}^{03}} \cdot 100,$$

esto es,

$$r_1 = \frac{R_1}{I_{02}^{03}} \cdot 100 = \frac{2{,}5}{102{,}15} \cdot 100 = 2{,}45,$$

$$r_2 = \frac{R_2}{I_{02}^{03}} \cdot 100 = \frac{0{,}455}{102{,}15} \cdot 100 = 0{,}44$$

y

$$r_3 = \frac{R_3}{I_{02}^{03}} \cdot 100 = \frac{0{,}555}{102{,}15} \cdot 100 = 0{,}54.$$

Por otro lado, la variación relativa del índice, o tasa de variación, entre los años 2003 y 2004, en porcentajes, es

$$\dot{I}_{02}^{04} = \frac{\Delta I_{02}^{04}}{I_{02}^{03}} \cdot 100 = \frac{3{,}51}{102{,}15} \cdot 100 = 3{,}43,$$

que, como puede comprobar el lector, es igual a la suma de las repercusiones relativas de cada turno en la variación del índice.

4.48 En la tabla siguiente figuran los valores del índice de precios del consumo de un país para los años 1999 y 2000 con base en el año 1993, para cada uno de los doce grupos que lo constituyen, así como los coeficientes de ponderación, con base en el mismo año, correspondientes a cada uno.

Grupo	Coeficiente de ponderación	Índice 1999	Índice 2000
Alimentos y bebidas no alcohólicas	20	110	112
Bebidas alcohólicas y tabaco	3	180	181
Vestido y calzado	11	101	104
Vivienda	12	140	142
Medicina	7	143	145
Menaje	2,7	122	125
Transporte	14,2	127	123
Comunicaciones	2	169	172
Ocio y cultura	8	113	110
Enseñanza	0,8	133	134
Hoteles, cafés y restaurantes	9,5	119	121
Otros	9,8	167	168
Total	100		

a) Hállese el índice de precios para los años 1999 y 2000. ¿Cuál es la variación absoluta del índice entre estos dos años? ¿Y el incremento relativo?

b) Obténgase las participaciones y repercusiones de cada grupo en la variación experimentada por el índice entre los años 1999 y 2000.

SOLUCIÓN

a) Un índice de precios al consumo es un índice complejo ponderado que responde a la expresión general:

$$I_0^t = \frac{\sum_{i=1}^{N} I_0^t(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde $I_0^t(i)$ y w_i son, respectivamente, el índice de precios y el coeficiente de ponderación del grupo i-ésimo.

Teniendo en cuenta que, en este caso, son doce los grupos que constituyen el índice, que para cada uno de ellos se dispone del coeficiente de ponderación correspondiente a los años 1999 y

2000 y que la suma de los coeficientes de ponderación es igual a 100, los índices de precios de dichos años con base en 1993 son:

$$I_{93}^{99} = \frac{110 \cdot 20 + 180 \cdot 3 + 101 \cdot 11 + 140 \cdot 12 + 143 \cdot 7 + 122 \cdot 2{,}7 + 127 \cdot 14{,}2}{100} +$$

$$+ \frac{169 \cdot 2 + 113 \cdot 8 + 133 \cdot 0{,}8 + 119 \cdot 9{,}5 + 167 \cdot 9{,}8}{100} = 127{,}803$$

e

$$I_{93}^{00} = \frac{112 \cdot 20 + 181 \cdot 3 + 104 \cdot 11 + 142 \cdot 12 + 145 \cdot 7 + 125 \cdot 2{,}7 + 123 \cdot 14{,}2}{100} +$$

$$+ \frac{172 \cdot 2 + 110 \cdot 8 + 134 \cdot 0{,}8 + 121 \cdot 9{,}5 + 168 \cdot 9{,}8}{100} = 128{,}572.$$

La variación absoluta del índice con base en 1993 entre los años 1999 y 2000 es

$$\Delta I_{93}^{00} = I_{93}^{00} - I_{93}^{99} = 128{,}572 - 127{,}803 = 0{,}769.$$

En cuanto a la variación relativa, su cálculo es inmediato, pues

$$\dot{I}_{93}^{00} = \frac{\Delta I_{93}^{00}}{I_{93}^{99}},$$

con lo cual,

$$\dot{I}_{93}^{00} = \frac{0{,}769}{127{,}803} = 0{,}00601,$$

siendo, por tanto, del 0,601 por ciento el incremento relativo del índice, con base en 1993, entre los años 1999 y 2000.

b) La repercusión absoluta del grupo *i*-ésimo en la variación absoluta del índice entre los años 1999 y 2000 es

$$R_i = \frac{\Delta I_{93}^{00}(i) \cdot w_i}{\sum_{i=1}^{N} w_i}.$$

Aplicando esta expresión a cada uno de los doce grupos y considerando que la suma de los coeficientes de ponderación es igual a 100, se obtienen las repercusiones absolutas de cada uno de los grupos sobre la variación absoluta del índice:

$$R_1 = \frac{\Delta I_{93}^{00}(1) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(112 - 110)\ 20}{100} = 0{,}4$$

$$R_2 = \frac{\Delta I_{93}^{00}(2) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(181 - 180)\ 3}{100} = 0{,}03$$

$$R_3 = \frac{\Delta I_{93}^{00}(3) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(104 - 101)\ 11}{100} = 0{,}33$$

$$R_4 = \frac{\Delta I_{93}^{00}(4) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(142 - 140)\ 12}{100} = 0{,}24$$

$$R_5 = \frac{\Delta I_{93}^{00}(5) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(145 - 143)\ 7}{100} = 0{,}14$$

$$R_6 = \frac{\Delta I_{93}^{00}(6) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(125 - 122)\ 2{,}7}{100} = 0{,}081$$

$$R_7 = \frac{\Delta I_{93}^{00}(7) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(123 - 127)\ 14{,}2}{100} = -0{,}568$$

$$R_8 = \frac{\Delta I_{93}^{00}(8) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(172 - 169)\ 2}{100} = 0{,}06$$

$$R_9 = \frac{\Delta I_{93}^{00}(9) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(110 - 113)\,8}{100} = -0{,}24$$

$$R_{10} = \frac{\Delta I_{93}^{00}(10) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(134 - 133)\,0{,}8}{100} = 0{,}008$$

$$R_{11} = \frac{\Delta I_{93}^{00}(11) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(121 - 119)\,9{,}5}{100} = 0{,}19$$

$$R_{12} = \frac{\Delta I_{93}^{00}(12) \cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{(168 - 167)\,9{,}8}{100} = 0{,}098.$$

Como se puede comprobar, la variación absoluta del índice entre los años 1999 y 2000, esto es, ΔI_{93}^{00}, calculada en el apartado anterior, es igual a la suma de las repercusiones absolutas:

$$\sum_{i=1}^{N} R_i = 0{,}769.$$

Dividiendo las repercusiones absolutas entre el valor del índice para 1999 resultan, según se demostró en el problema **4.42**, las repercusiones relativas de cada grupo. Así, las repercusiones relativas, en porcentajes, son las siguientes:

$$r_1 = \frac{R_1}{I_{93}^{99}} \cdot 100 = \frac{0{,}4}{127{,}803} \cdot 100 = 0{,}313$$

$$r_2 = \frac{R_2}{I_{93}^{99}} \cdot 100 = \frac{0{,}03}{127{,}803} \cdot 100 = 0{,}023$$

$$r_3 = \frac{R_3}{I_{93}^{99}} \cdot 100 = \frac{0{,}33}{127{,}803} \cdot 100 = 0{,}258$$

$$r_4 = \frac{R_4}{I_{93}^{99}} \cdot 100 = \frac{0{,}24}{127{,}803} \cdot 100 = 0{,}188$$

$$r_5 = \frac{R_5}{I_{93}^{99}} \cdot 100 = \frac{0{,}14}{127{,}803} \cdot 100 = 0{,}109$$

$$r_6 = \frac{R_6}{I_{93}^{99}} \cdot 100 = \frac{0{,}081}{127{,}803} \cdot 100 = 0{,}063$$

$$r_7 = \frac{R_7}{I_{93}^{99}} \cdot 100 = \frac{-0{,}568}{127{,}803} \cdot 100 = -0{,}444$$

$$r_8 = \frac{R_8}{I_{93}^{99}} \cdot 100 = \frac{0{,}06}{127{,}803} \cdot 100 = 0{,}047$$

$$r_9 = \frac{R_9}{I_{93}^{99}} \cdot 100 = \frac{-0{,}24}{127{,}803} \cdot 100 = -0{,}188$$

$$r_{10} = \frac{R_{10}}{I_{93}^{99}} \cdot 100 = \frac{0{,}008}{127{,}803} \cdot 100 = 0{,}006$$

$$r_{11} = \frac{R_{11}}{I_{93}^{99}} \cdot 100 = \frac{0{,}19}{127{,}803} \cdot 100 = 0{,}149$$

$$r_{12} = \frac{R_{12}}{I_{93}^{99}} \cdot 100 = \frac{0{,}098}{127{,}803} \cdot 100 = 0{,}077.$$

Se comprueba que la suma de las repercusiones relativas,

$$\sum_{i=1}^{N} r_i = 0{,}601,$$

coincide con la variación relativa del índice, expresada en porcentajes, es decir, con la tasa de variación del índice, $\dot{I}_{93}^{00}$, hallada en el apartado ***a)***.

La participación de la componente *i*-ésima del índice en la variación del mismo es, según se demostró en el problema **4.44**, igual a

$$P_i = \frac{R_i}{\Delta I_{93}^{00}}.$$

En consecuencia, la participación, en porcentajes, de cada grupo de bienes es

$$P_1 = \frac{R_1}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}4}{0{,}769} \cdot 100 = 52{,}02$$

$$P_2 = \frac{R_2}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}03}{0{,}769} \cdot 100 = 3{,}9$$

$$P_3 = \frac{R_3}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}33}{0{,}769} \cdot 100 = 42{,}91$$

$$P_4 = \frac{R_4}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}24}{0{,}769} \cdot 100 = 31{,}21$$

$$P_5 = \frac{R_5}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}14}{0{,}769} \cdot 100 = 18{,}21$$

$$P_6 = \frac{R_6}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}081}{0{,}769} \cdot 100 = 10{,}53$$

$$P_7 = \frac{R_7}{\Delta I_{93}^{00}} \cdot 100 = \frac{-0{,}568}{0{,}769} \cdot 100 = -73{,}86$$

$$P_8 = \frac{R_8}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}06}{0{,}769} \cdot 100 = 7{,}8$$

$$P_9 = \frac{R_9}{\Delta I_{93}^{00}} \cdot 100 = \frac{-0{,}24}{0{,}769} \cdot 100 = -31{,}21$$

$$P_{10} = \frac{R_{10}}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}008}{0{,}769} \cdot 100 = 1{,}04$$

$$P_{11} = \frac{R_{11}}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}19}{0{,}769} \cdot 100 = 24{,}71$$

$$P_{12} = \frac{R_{12}}{\Delta I_{93}^{00}} \cdot 100 = \frac{0{,}098}{0{,}769} \cdot 100 = 12{,}74.$$

Los grupos que más han influido en la variación del índice entre los años 1999 y 2000 han sido el primer grupo (alimentos y bebidas no alcohólicas) y el séptimo (transporte).

El primer grupo es el que más ha contribuido al incremento del índice, ya que su repercusión del 0,4 por ciento supone una participación del 52,01 por ciento sobre la variación total. En sentido opuesto, el séptimo grupo es el que más ha influido en la disminución del índice, pues su repercusión de $-0{,}568$ por ciento supone, en sentido negativo, una participación del 73,86 por ciento en la variación total.

4.49 En la siguiente tabla figuran los coeficientes de ponderación con base en el año 1994 del índice de precios al consumo de un país, así como el valor del índice en el mes de febrero del año 2001 y su tasa de variación anual, en porcentajes, para cada uno de los ocho grupos de bienes utilizados para la elaboración del índice.

Grupo	Coeficiente de ponderación	Índice febrero 2001	Tasa variación anual
1	19,05	124,0	4,7
2	12,32	134,9	3,3
3	11,25	156,8	2,6
4	9,76	122,7	4,4
5	14,02	111,1	2,8
6	8,68	147,3	2,2
7	9,03	123,6	1,4
8	15,89	148,5	1,1

a) Hállese el índice de precios al consumo del mes de febrero del año 2001 con base en 1994.

b) Obténgase, para cada grupo, el índice del mes de febrero del año 2000.

c) Determínense las repercusiones de cada grupo en la variación experimentada por el índice entre febrero de 2000 y febrero de 2001. Calcúlese la participación del grupo 5 en dicha variación.

d) Calcúlese la tasa de variación anual del índice de precios al consumo para el mes de febrero del año 2001.

SOLUCIÓN

a) El índice de precios al consumo es un índice complejo ponderado de los índices simples de precios de cada grupo de bienes:

$$I_{94}^{01} = \frac{\sum_{i=1}^{N} I_{94}^{01}(i) \cdot w_i}{\sum_{i=1}^{N} w_i}.$$

Sustituyendo por los datos de la tabla y teniendo en cuenta, como puede comprobarse, que la suma de los coeficientes de ponderación es igual a 100, resulta:

$$I_{94}^{01} = \frac{19{,}05 \cdot 124 + 12{,}32 \cdot 134{,}9 + 11{,}25 \cdot 156{,}8 + 9{,}76 \cdot 122{,}7}{100} +$$

$$+ \frac{14{,}02 \cdot 111{,}1 + 8{,}68 \cdot 147{,}3 + 9{,}03 \cdot 123{,}6 + 15{,}89 \cdot 148{,}5}{100} = 132{,}98.$$

b) Considerando que, por definición, la tasa de variación anual entre febrero del año 2000 y febrero del año 2001, expresada en tanto por uno, del índice del grupo i-ésimo es

$$\dot{I}_{94}^{01}(i) = \frac{I_{94}^{01}(i)}{I_{94}^{00}(i)} - 1,$$

se obtiene, despejando, el índice del mes de febrero de 2000 de dicho grupo:

$$I_{94}^{00}(i) = \frac{I_{94}^{01}(i)}{\dot{I}_{94}^{01}(i) + 1}.$$

Sustituyendo por los datos de la tabla resulta, entonces, el índice del mes de febrero para cada grupo:

$$I_{94}^{00}(1) = \frac{I_{94}^{01}(1)}{\dot{I}_{94}^{01}(1) + 1} = \frac{124}{0{,}047 + 1} = 118{,}4$$

$$I_{94}^{00}(2) = \frac{I_{94}^{01}(2)}{\dot{I}_{94}^{01}(2) + 1} = \frac{134{,}9}{0{,}033 + 1} = 130{,}6$$

$$I_{94}^{00}(3) = \frac{I_{94}^{01}(3)}{\dot{I}_{94}^{01}(3) + 1} = \frac{156{,}8}{0{,}026 + 1} = 152{,}8$$

$$I_{94}^{00}(4) = \frac{I_{94}^{01}(4)}{\dot{I}_{94}^{01}(4) + 1} = \frac{122{,}7}{0{,}044 + 1} = 117{,}5$$

$$I_{94}^{00}(5) = \frac{I_{94}^{01}(5)}{\dot{I}_{94}^{01}(5) + 1} = \frac{111{,}1}{0{,}028 + 1} = 108{,}1$$

$$I_{94}^{00}(6) = \frac{I_{94}^{01}(6)}{\dot{I}_{94}^{01}(6) + 1} = \frac{147{,}3}{0{,}022 + 1} = 144{,}1$$

$$I_{94}^{00}(7) = \frac{I_{94}^{01}(7)}{\dot{I}_{94}^{01}(7) + 1} = \frac{123{,}6}{0{,}014 + 1} = 121{,}9$$

$$I_{94}^{00}(8) = \frac{I_{94}^{01}(8)}{\dot{I}_{94}^{01}(8) + 1} = \frac{148{,}5}{0{,}011 + 1} = 146{,}9.$$

c) La repercusión absoluta del grupo i-ésimo en la variación absoluta del índice entre los meses de febrero de los años 2000 y 2001 responde a la expresión:

$$R_i = \frac{\Delta I_{94}^{01}(i) \cdot w_i}{\sum_{i=1}^{N} w_i},$$

donde $\Delta I_{94}^{01}(i)$ es la variación absoluta del índice del grupo i-ésimo entre los dos periodos contemplados, esto es,

$$\Delta I_{94}^{01}(i) = I_{94}^{01}(i) - I_{94}^{00}(i).$$

En la tabla siguiente figuran los índices de precios para el mes de febrero de los años considerados, así como la variación absoluta y la repercusión absoluta de cada grupo, calculadas según las expresiones anteriores.

Grupo	Índice febrero 2000	Índice febrero 2001	Variación absoluta	Repercusión absoluta
1	118,4	124,0	5,6	1,067
2	130,6	134,9	4,3	0,530
3	152,8	156,8	4	0,450
4	117,5	122,7	5,2	0,507
5	108,1	111,1	3	0,421
6	144,1	147,3	3,2	0,278
7	121,9	123,6	1,7	0,153
8	146,9	148,5	1,6	0,254

Por lo que respecta a las repercusiones relativas, hay que dividir cada una de las absolutas entre el valor del índice en el mes de febrero de 2000, I_{94}^{00}, índice del que no se dispone explícitamente pero que es posible obtener.

Como recordará el lector, la suma de las repercusiones absolutas es igual a la variación absoluta del índice, con lo cual, en este caso,

$$\sum_{i=1}^{N} R_i = 3{,}66 = \Delta I_{94}^{01},$$

y, a su vez, por definición de variación absoluta,

$$\Delta I_{94}^{01} = I_{94}^{01} - I_{94}^{00},$$

por lo que, conocidos los valores de la variación absoluta y del índice del mes de febrero del año 2001, se tiene que

$$I_{94}^{00} = I_{94}^{01} - \Delta I_{94}^{01} = 132{,}98 - 3{,}66 = 129{,}32.$$

En definitiva, las repercusiones relativas de cada componente, en porcentajes, son:

$$r_1 = \frac{R_1}{I_{94}^{00}} \cdot 100 = \frac{1{,}067}{129{,}32} \cdot 100 = 0{,}825$$

$$r_2 = \frac{R_2}{I_{94}^{00}} \cdot 100 = \frac{0{,}53}{129{,}32} \cdot 100 = 0{,}41$$

$$r_3 = \frac{R_3}{I_{94}^{00}} \cdot 100 = \frac{0{,}45}{129{,}32} \cdot 100 = 0{,}348$$

$$r_4 = \frac{R_4}{I_{94}^{00}} \cdot 100 = \frac{0{,}507}{129{,}32} \cdot 100 = 0{,}392$$

$$r_5 = \frac{R_5}{I_{94}^{00}} \cdot 100 = \frac{0{,}421}{129{,}32} \cdot 100 = 0{,}326$$

$$r_6 = \frac{R_6}{I_{94}^{00}} \cdot 100 = \frac{0{,}278}{129{,}32} \cdot 100 = 0{,}215$$

$$r_7 = \frac{R_7}{I_{94}^{00}} \cdot 100 = \frac{0{,}153}{129{,}32} \cdot 100 = 0{,}118$$

$$r_8 = \frac{R_8}{I_{94}^{00}} \cdot 100 = \frac{0{,}254}{129{,}32} \cdot 100 = 0{,}197.$$

Por último, la participación, en porcentaje, del quinto grupo de bienes en la variación del índice entre los periodos considerados es

$$P_5 = \frac{R_5}{\Delta I_{94}^{01}} \cdot 100 = \frac{0{,}421}{3{,}66} \cdot 100 = 11{,}50,$$

con lo cual, el grupo 5 ha influido en la variación del índice con una repercusión del 0,421 por ciento, que supone una participación del 11,50 por ciento.

d) Por definición, la tasa de variación anual del índice de precios al consumo entre el mes de febrero de 2000 y el mismo mes de 2001, $\dot{I}_{94}^{01}$, es

$$\dot{I}_{94}^{01} = \frac{I_{94}^{01}}{I_{94}^{00}} - 1.$$

Sustituyendo, entonces, los valores del índice para el mes de febrero de los años 2000 y 2001, hallados en los apartados ***a)*** y ***c),*** respectivamente, resulta:

$$\dot{I}_{94}^{01} = \frac{132{,}98}{129{,}32} - 1 = 0{,}0283,$$

es decir, una tasa de variación anual del 2,83 por ciento.

Puede comprobar el lector que la suma de las repercusiones relativas obtenidas en el apartado ***c)*** es igual a la tasa de variación anual del índice.

Capítulo 5

Análisis clásico de series de tiempo

Principales conceptos y resultados

Una **serie temporal** es el conjunto de observaciones de una variable en diferentes periodos de tiempo. En el análisis de una serie de tiempo la variable se explica exclusivamente por su historia, es decir, cada dato está determinado por *el simple paso del tiempo*.

La teoría clásica en el análisis de las series temporales se basa en que cada observación de la variable es el resultado de la acción conjunta de cuatro componentes[1]:

- **Tendencia** o componente a largo plazo de la serie, T.
- **Ciclo** o componente a medio plazo, c.
- **Variaciones estacionales** de periodicidad corta, e.
- **Componente accidental** o **residual** sin periodicidad reconocida, a.

Considerando N periodos de tiempo divididos en k subperiodos, la observación genérica de la variable Y, y_{ij}, $(i = 1, ..., N; j = 1, ..., k)$, se obtiene como

$$y_{ij} = T_{ij} + c_{ij} + e_{ij} + a_{ij},$$

si optamos por un esquema aditivo.

[1] El análisis de las series de tiempo ha experimentado un gran avance en las últimas décadas, a partir de los trabajos de Box y Jenkins basados en el concepto de proceso estocástico. Con el análisis clásico de las series temporales se pretende una aproximación al conocimiento de la evolución en el tiempo de una variable, mediante la descripción de las componentes de la serie.

Por el contrario, con un esquema multiplicativo tendremos que las cuatro componentes se relacionan según el modelo[2]:

$$y_{ij} = T_{ij} \cdot c_{ij} \cdot e_{ij} + a_{ij}.$$

Cuando las observaciones de la variables están referidas a periodos, sin consideración de subperiodos dentro de los mismos, denotaremos por[3] y_t al dato del periodo t.

El estudio descriptivo de una serie de tiempo pretende aislar las distintas componentes de la misma[4].

El **método mecánico de las medias móviles** es un procedimiento para aislar la tendencia de la serie, suavizándola al eliminar oscilaciones[5], mediante el cálculo de las denominadas **medias móviles**[6]. La media móvil de orden $2h + 1$ de la observación y_t responde a la expresión:

$$\overline{y}_t = \frac{y_{t-h} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+h}}{2 \cdot h + 1}.$$

Si el número de observaciones consideradas es impar, las medias móviles están *centradas*. Si, por el contrario, promediamos un número par de observaciones, las medias móviles aparecerán *descentradas*, siendo necesario centrarlas, promediando, esta vez de dos en dos, los valores de la serie de medias móviles descentradas.

El **método analítico de los mínimos cuadrados** consiste en *estimar* la tendencia mediante la regresión mínimo-cuadrática de la variable con respecto al tiempo. Puesto que lo más habitual es contar con más de una observación por periodo, lo correcto es realizar la regresión de *los valores medios de cada periodo* de la variable sobre el tiempo *expresado en periodos*, obvian-

---

[2] La elección de uno u otro esquema se basa en métodos empíricos aplicados sobre la componente estacional. Uno de los métodos más utilizados consiste en calcular los valores medios y las desviaciones típicas de las observaciones correspondientes a cada periodo y obtener el ajuste lineal entre estas nuevas variables; una línea *prácticamente* paralela al eje horizontal es indicativa de esquema aditivo.

En cualquier caso, las series correspondientes a magnitudes económicas suelen regirse por el esquema multiplicativo.

Nótese, también, que, por su naturaleza errática, la componente accidental no ha de tener relación con el resto de las componentes, con lo cual, su influencia sobre la observación debe ser aditiva.

[3] Al no disponer de observaciones de la variable en los subperiodos, no es posible estudiar la componente estacional cuya periodicidad es inferior a un periodo.

[4] En este capítulo estudiamos los métodos descriptivos habitualmente utilizados para aislar la tendencia y el ciclo conjuntamente, esto es, la denominada *componente extraestacional*, también llamada componente a largo plazo, así como la componente estacional. El aislamiento del ciclo y de la componente accidental requiere el empleo de técnicas más complejas que superan los objetivos de esta obra.

[5] Cuando la serie tiene componente estacional, con este modo de actuar eliminamos, además de esta componente, parte de la componente accidental, quedándonos con la componente extraestacional. En cualquier caso, para series desprovistas de componente estacional, también emplearemos el método de las medias móviles como intento de desprendernos de la componente cíclica.

[6] Al promediar observaciones contiguas, y dado que las componentes estacional y residual tienen signos opuestos de unos periodos a otros debido a su corta duración, conseguimos eliminar estos tipos de fluctuaciones.

do en el análisis la existencia de componente estacional. Resultará, así, la ecuación de la recta[7] de tendencia,

$$\bar{y}_{i\cdot} = a + b \cdot i,$$

donde $\bar{y}_{i\cdot}$ es el valor medio del periodo *i*-ésimo.

Cuando en la estimación de la recta de tendencia se obtiene $\bar{y}_{i\cdot} = a$, esto es, una recta paralela al eje horizontal, el modelo se denomina **estacionario** o de media constante.

El **método de las relaciones (razón o diferencia**, **según que el esquema sea multiplicativo o aditivo) a la media móvil**, que permite la obtención de la componente estacional, consta de las siguientes fases:

- Se halla la serie de tendencia[8] mediante el procedimiento de cálculo de las medias móviles con el objeto de eliminar distorsiones debidas a la componente estacional.
- Si el esquema es multiplicativo (aditivo) se dividen (restan) las observaciones de la serie original, y_{ij}, por las correspondientes observaciones de la serie de tendencia, obteniéndose, así, una nueva serie de observaciones, y'_{ij}, que corresponden solamente a las componentes estacional y residual.
- Para eliminar la componente residual de la nueva serie se calcula la media aritmética de cada uno de los subperiodos, esto es, para todo *j*,

$$\bar{y}'_{\cdot j} = \frac{1}{N} \sum_{i=1}^{N} y'_{ij}.$$

- Se obtiene la media *global*:

$$\bar{y}' = \frac{1}{k} \sum_{j=1}^{k} \bar{y}'_{\cdot j}.$$

- Si el esquema es aditivo, la componente estacional, expresada en las mismas unidades que la variable[9], del subperiodo *j*-ésimo se obtiene restando, según la expresión genérica:

$$e_{\cdot j} = \bar{y}'_{\cdot j} - \bar{y}'.$$

- Si el esquema es multiplicativo, la componente estacional genérica, coeficiente adimensional[10], es

$$e_{\cdot j} = \frac{\bar{y}'_{\cdot j}}{\bar{y}'}.$$

---

[7] La representación gráfica de la serie de tiempo puede sugerir una tendencia no lineal, siendo en tales casos más adecuado el ajuste funcional que mejor refleje dicha tendencia. Es frecuente, sin embargo, que las series económicas presenten tendencia lineal.

[8] En realidad, la componente a largo plazo, constituida por tendencia y ciclo. Véase nota 4.

[9] Cuando el esquema es aditivo, todas las componentes están expresadas en las mismas unidades que la variable.

[10] En un esquema multiplicativo las componentes estacional, cíclica y accidental son adimensionales.

Es frecuente, en este caso, expresar la componente estacional en tanto por ciento, obteniéndose el denominado **índice de variación estacional**, cuya expresión genérica es

$$I_j = e_{\cdot j} \cdot 100.$$

El **método de las relaciones de las medias de cada subperiodo con respecto a la tendencia** permite, también, la identificación de la componente estacional y consta de las siguientes fases:

- Se calculan las medias de cada subperiodo, según la expresión genérica:

$$\bar{y}_{\cdot j} = \frac{1}{N} \sum_{i=1}^{N} y_{ij}.$$

- Se corrigen las medias anteriores por la tendencia, eliminando la variación debida únicamente al paso del tiempo. Para ello, y una vez estimada la tendencia por el método de los mínimos cuadrados, se resta de cada media la proporción del incremento de todo el periodo que corresponde a cada subperiodo que ha transcurrido; la *media corregida* del subperiodo *j*-ésimo se calcula como

$$\bar{y}'_{\cdot j} = \bar{y}_{\cdot j} - \frac{b}{k}(j - 1),$$

  donde *b/k* es la variación que se produce en el valor medio del subperiodo por el paso de un subperiodo.

- Se halla la media *global* corregida, esto es, el promedio de las medias corregidas:

$$\bar{y}' = \frac{1}{k} \sum_{j=1}^{k} \bar{y}'_{\cdot j}.$$

- Si el esquema es aditivo, la componente estacional del subperiodo *j*-ésimo se obtiene por diferencia[11],

$$e_{\cdot j} = \bar{y}'_{\cdot j} - \bar{y}'.$$

- Si el esquema es multiplicativo, la componente estacional del subperiodo *j*-ésimo es

$$e_{\cdot j} = \frac{\bar{y}'_{\cdot j}}{\bar{y}'},$$

  proporción de variación sobre el valor global medio de una observación por pertenecer a un subperiodo determinado.

  Al igual que en el método de las relaciones a la media móvil, expresando la componente estacional en porcentajes, se obtiene el índice de variación estacional.

[11] Nótese que estamos considerando, por un lado, la tendencia y la componente cíclica conjuntamente y, por otro, estamos ignorando la posible existencia de componente accidental.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

5.1 Un consultorio psicopedagógico desea estudiar la evolución que, desde su inauguración, se ha producido en el número de clientes que han acudido al mismo.

La siguiente tabla recoge el número de personas, en miles, que anualmente han acudido a la consulta desde 1990.

Años	N.º pacientes
1990	7,30
1991	7,50
1992	8,40
1993	8,80
1994	9,12
1995	9,80
1996	10,22
1997	10,95
1998	11,31
1999	11,70
2000	12,04
2001	12,77
2002	13,50
2003	14,60
2004	17,20

a) Represéntese gráficamente la serie de tiempo.

b) Calcúlese la tendencia, utilizando el método de los mínimos cuadrados ordinarios.

c) Si no se produce un cambio en la estructura de la serie, ¿qué número de pacientes se prevé que acuda en el año 2006?

d) Analícese la fiabilidad del resultado obtenido.

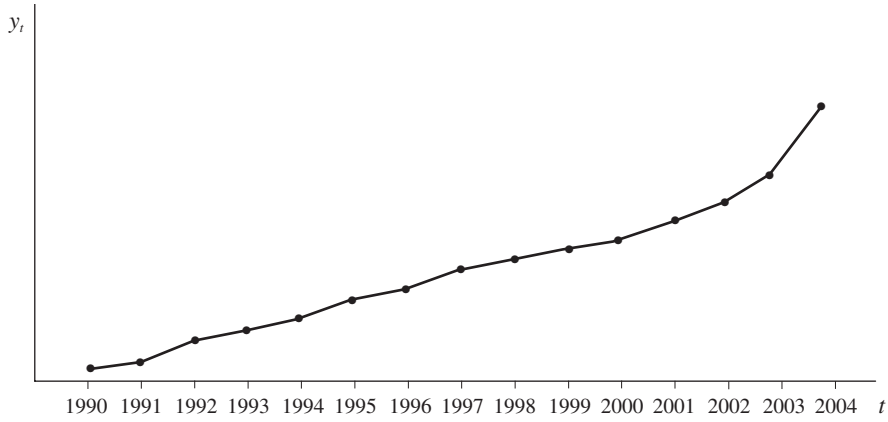
SOLUCIÓN

a) Tenemos un conjunto de observaciones de la variable *Y*, número de personas, en miles, que anualmente han acudido a un consultorio psicopedagógico en diferentes periodos de tiem-

po, como son los últimos quince años; se trata, por tanto, de una serie temporal de la variable *Y*.

La representación gráfica de la serie temporal está formada por pares de puntos, (t, y_t), correspondiendo la primera componente al periodo de tiempo —en este caso el año— y la segunda a la observación de la variable en dicho año.

Dibujamos, por tanto, unos ejes de coordenadas, donde el eje de abscisas es para los periodos de tiempo y el eje de ordenadas para los valores de la variable.



Si, en lugar de datos anuales, es decir, observaciones de la variable referidas a un periodo de tiempo, hubiéramos dispuesto de datos correspondientes a subperiodos (semestres, cuatrimestres, trimestres o meses), la serie de tiempo mostraría oscilaciones de periodicidad inferior al año debidas a la componente estacional. Con la información disponible no sería posible, por tanto, realizar un estudio de dicha componente para la variable considerada.

b) La existencia de una tendencia lineal sugerida por la representación gráfica de la serie de tiempo conduce a la estimación de la recta de regresión de la variable *Y* con respecto al tiempo,

$$y_t = a + b \cdot t,$$

donde los parámetros *a* y *b* se hallan mediante aplicación del criterio de los mínimos cuadrados; es necesario comentar que, aunque el procedimiento de regresión es el mismo que hemos seguido con dos variables cualesquiera en el capítulo 2, en esta ocasión no es nuestra intención *explicar* la variable *Y a partir* del tiempo, sino únicamente en función de su propio comportamiento *en* el tiempo.

Los valores teóricos resultantes de este ajuste lineal son los valores de tendencia de la serie de tiempo.

Para la obtención de la recta de regresión se transforman las observaciones de la variable tiempo, pasando de *t* a $t - o = t'$, donde *o* es el periodo que ocupa la posición central; de este modo,

se logra que la media de la variable transformada sea cero, con la consiguiente simplificación de las operaciones. Obtendremos, así, los parámetros de la recta de regresión,

$$y_t = a + b \cdot t',$$

con $t' = t - 1997$, en este caso.

En la siguiente tabla aparecen los cálculos intermedios que permitirán la obtención de los momentos no centrales y centrales necesarios para realizar el ajuste: además de las dos primeras columnas correspondientes a las dos variables de partida, la tercera columna contiene las observaciones de la variable transformada que, como puede verse en la última casilla de dicha columna —suma de los valores de la misma—, tiene media cero; la cuarta y quinta columna contienen, respectivamente, los cuadrados de las observaciones de la variable transformada y de la variable *Y*; finalmente, en la sexta columna figuran los productos de las observaciones de ambas variables.

t	y_t	t'	t'^2	y_t^2	$y_t \cdot t'$
1990	7,30	−7	49	53,2900	−51,10
1991	7,50	−6	36	56,2500	−45,00
1992	8,40	−5	25	70,5600	−42,00
1993	8,80	−4	16	77,4400	−35,20
1994	9,12	−3	9	83,1744	−27,36
1995	9,80	−2	4	96,0400	−19,60
1996	10,22	−1	1	104,4484	−10,22
1997	10,95	0	0	119,9025	0,00
1998	11,31	1	1	127,9161	11,31
1999	11,70	2	4	136,8900	23,40
2000	12,04	3	9	144,9616	36,12
2001	12,77	4	16	163,0729	51,08
2002	13,50	5	25	182,2500	67,50
2003	14,60	6	36	213,1600	87,60
2004	17,20	7	49	295,8400	120,40
	165,21	**0**	**280**	**1 925,1959**	**166,93**

Las expresiones de las estimaciones de los parámetros de la recta de regresión estudiadas en el capítulo 2 se adaptan al contexto de las series de tiempo, de modo que el coeficiente de regresión de la recta es

$$b = \frac{S_{t',y_t}}{S_{t'}^2},$$

y el término independiente:

$$a = \bar{y} - b \cdot \bar{t}'.$$

La elección de la variable transformada t' hace que[1] el cálculo, tanto de su varianza como de la covarianza de las variables, se simplifique. Así, si N es el número de periodos,

$$S_{t',y_t} = \frac{1}{N}\sum t' \cdot y_t - \bar{t}' \cdot \bar{y}_t = \frac{1}{N}\sum t' \cdot y_t - 0 \cdot \bar{y} = \frac{1}{N}\sum t' \cdot y_t$$

y

$$S^2_{t'} = \frac{1}{N}\sum t'^2 - \bar{t}'^2 = \frac{1}{N}\sum t'^2 - 0 = \frac{1}{N}\sum t'^2.$$

En definitiva, los coeficientes de la recta de tendencia son:

$$b = \frac{\frac{1}{N}\sum t' \cdot y_t}{\frac{1}{N}\sum t'^2} = \frac{\sum t' \cdot y_t}{\sum t'^2} = \frac{166{,}93}{280} = 0{,}596$$

y

$$a = \bar{y} = \frac{165{,}21}{15} = 11{,}014.$$

Por tanto, la ecuación de la recta de tendencia resulta:

$$y_t = 11{,}014 + 0{,}596\ (t - 1997).$$

Habríamos llegado a idéntico resultado partiendo del *sistema de ecuaciones normales* que, como puede comprobar el lector, adaptando las notaciones del capítulo 2 en la regresión de la variable Y sobre la variable transformada t', es

$$\sum y_t = N \cdot a + b \sum t'$$

$$\sum t' \cdot y_t = a \sum t' + b \sum t'^2.$$

[1] Prescindiremos de los índices de los sumatorios para evitar complicar las notaciones; en cualquier caso, estos sumatorios se extienden a todas las observaciones de cada variable. Con idéntico fin hemos empleado minúsculas para designar las variables en las expresiones de los momentos.

Puesto que $\sum t' = 0$, el sistema anterior queda reducido a

$$\sum y_t = N \cdot a$$

$$\sum t' \cdot y_t = b \sum t'^2,$$

de donde se despejan lo valores de a y de b.

c) Si suponemos que la tendencia se mantiene a lo largo del tiempo, la previsión del número de pacientes para 2006 será, sustituyendo en la recta de tendencia el valor de dicho año,

$$y^*_{06} = 11{,}014 + 0{,}596\ (2006 - 1997) = 16{,}378 \text{ miles de pacientes.}$$

d) Para analizar la fiabilidad del resultado anterior, calculamos el coeficiente de determinación lineal, medida de la bondad del ajuste en la regresión lineal efectuada:

$$r^2 = \frac{S^2_{t',y_t}}{S^2_{t'} \cdot S^2_{y_t}}$$

Puesto que la covarianza entre las variables es

$$S_{t',y_t} = \frac{1}{N} \sum t' \cdot y_t = \frac{166{,}93}{15} = 11{,}128,$$

y las varianzas

$$S^2_{t'} = \frac{1}{N} \sum t'^2 = \frac{280}{15} = 18{,}67$$

y

$$S^2_{y_t} = \frac{1}{N} \sum y_t^2 - \overline{y}^2 = \frac{1\,925{,}1959}{15} - 11{,}014^2 = 7{,}04,$$

el coeficiente de determinación lineal resulta ser

$$r^2 = \frac{11{,}128^2}{18{,}67 \cdot 7{,}04} = 0{,}942,$$

de lo cual se concluye que la predicción es razonablemente fiable.

5.2 Un estudio destinado a analizar la evolución experimentada por el número de viviendas nuevas construidas en una ciudad, cuyo desarrollo urbanístico se ha producido

fundamentalmente en los últimos veinte años, arroja, entre otros, los datos que figuran en la siguiente tabla.

Año	N.º viviendas nuevas
1985	1 100
1986	2 000
1987	800
1988	500
1989	1 800
1990	450
1991	800
1992	1 500
1993	920
1994	1 400
1995	800
1996	1 400
1997	850
1998	1 500
1999	1 225
2000	1 600
2001	700
2002	1 800
2003	1 350
2004	2 000

Obténgase la serie de tendencia aplicando el método de las medias móviles, bajo el supuesto de que esta variable tiene oscilaciones cíclicas de periodo igual a 5 años.

SOLUCIÓN

El método de las medias móviles consiste en *suavizar* la serie quitando las oscilaciones; en este caso, al tratarse de datos anuales y no existir, por tanto, componente estacional en esta serie, nuestro objetivo será eliminar las oscilaciones cíclicas que presenta la variable, para lo cual hallaremos medias móviles de orden 5.

La obtención de una media móvil de una observación de la variable consiste en sumar dicha observación con las inmediatamente anteriores y posteriores, dividiendo el resultado por el número de observaciones consideradas; se trata, por tanto, de una media aritmética de observaciones. La expresión genérica de la media móvil de orden 5 de la observación y_t, es

$$y_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5},$$

siendo la primera media móvil de la serie la correspondiente a la tercera observación de la variable y la última media móvil a la antepenúltima observación.

La aplicación de la expresión anterior a los datos de la variable *Y*, número de viviendas nuevas construidas en una ciudad, conduce a la obtención de la serie de medias móviles que aparece en la siguiente tabla:

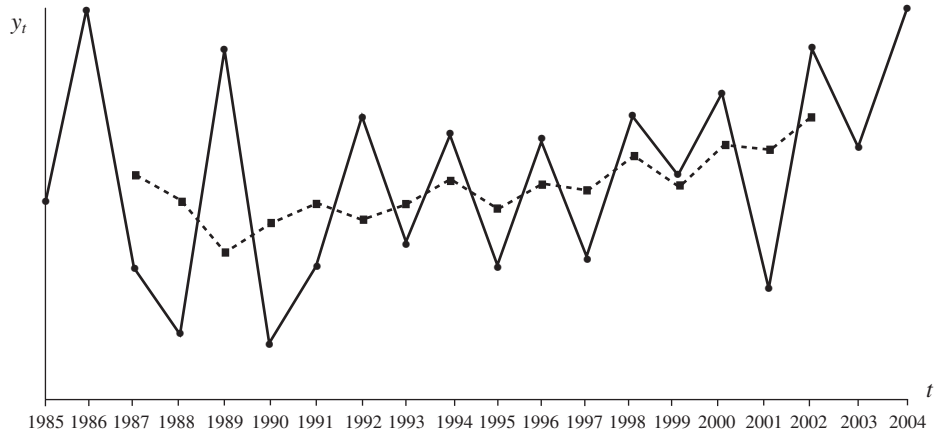
Año	Media móvil
1985	—
1986	—
1987	1 240
1988	1 110
1989	870
1990	1 010
1991	1 094
1992	1 014
1993	1 084
1994	1 204
1995	1 074
1996	1 190
1997	1 155
1998	1 315
1999	1 175
2000	1 365
2001	1 335
2002	1 490
2003	—
2004	—

Así, por ejemplo, la media móvil de la observación de la variable en el año 1995, esto es, de y_{11}, se calcula como

$$\bar{y}_{95} = \frac{y_{93} + y_{94} + y_{95} + y_{96} + y_{97}}{5} = \frac{920 + 1\,400 + 800 + 1\,400 + 850}{5} = 1\,074.$$

Como se ve, el método mecánico de las medias móviles conduce a una pérdida de observaciones como consecuencia de la obtención de promedios.

En la gráfica aparecen dos representaciones: la línea poligonal que une los pares de puntos (t, y_t) es la serie de tiempo de la magnitud estudiada, mientras que la poligonal que une los pares de puntos $(t, \bar{y}_t)$, mucho más suave que la anterior y más corta, como consecuencia de la pérdida de observaciones, es la línea de tendencia.



5.3 Obténgase la ecuación de ajuste de una tendencia potencial:

$$y_t = a \cdot t^b.$$

SOLUCIÓN

Para estimar a y b, utilizando el procedimiento de los mínimos cuadrados ordinarios, el método más sencillo consiste en *linealizar* la ecuación anterior, según vimos en el capítulo 2 en relación a la regresión entre dos variables X e Y. En efecto, tomando logaritmos neperianos, se tiene la ecuación:

$$\ln y_t = \ln a + b \cdot \ln t.$$

A partir de ella, haciendo los cambios de variable:

$$y_t' = \ln y_t$$

y

$$t' = \ln t,$$

y denotando por $c = \ln a$ ($a > 0$), resulta la ecuación lineal:

$$y_t' = c + b \cdot t'.$$

Hemos pasado, por tanto, de una tendencia potencial a una tendencia lineal, con lo cual, aplicando mínimos cuadrados, se estiman los coeficientes de la ecuación anterior del modo habitual:

$$b = \frac{S_{t',y_t'}}{S_{t'}^2}$$

y

$$c = \overline{y}' - b \cdot \overline{t}',$$

quedando también estimados los coeficientes de la ecuación de tendencia potencial, ya que

$$a = \exp(c).$$

5.4 Hállese la ecuación de ajuste de una tendencia exponencial:

$$y_t = a \cdot b^t.$$

SOLUCIÓN

Procediendo igual que en el problema anterior, se linealiza la ecuación de tendencia:

$$\ln y_t = \ln a + t \cdot \ln b.$$

Haciendo el cambio de variable

$$y'_t = \ln y_t,$$

y denotando por

$$c = \ln a$$

y

$$d = \ln b,$$

con $a, b > 0$, se obtiene la estimación de los coeficientes de la ecuación de ajuste de una tendencia exponencial:

$$b = \exp(d) = \exp\left(\frac{S_{t,y'_t}}{S_t^2}\right)$$

y

$$a = \exp(c) = \exp(\overline{y}' - d \cdot \overline{t}).$$

5.5 El canal de televisión privado Antena Norte ha realizado un estudio sobre la evolución de la audiencia en un periodo de 20 años con los siguientes datos, en millones, correspondiente al número de telespectadores:

N.º telespectadores
6
7
8
9
10
14
18
16
17
24
22
26
30
34
38
42
46
50
54
61

a) Obténgase, mediante el método de las medias móviles con periodo igual a 3, la serie suavizada.

b) Ajústese una recta de tendencia a la serie suavizada.

SOLUCIÓN

a) Puesto que los datos son anuales no existe componente estacional en esta serie, con lo cual, el cálculo de las medias móviles servirá para suavizarla, eliminado otras oscilaciones como pueden ser las debidas a las componentes cíclica y accidental.

La obtención de las medias móviles de orden 3 aplicando la expresión

$$\bar{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3},$$

conduce a los resultados que figuran en la segunda columna de la tabla siguiente:

Años	Media móvil
1	—
2	7
3	8
4	9
5	11
6	14
7	16
8	17
9	19
10	21
11	24
12	26
13	30
14	34
15	38
16	42
17	46
18	50
19	55
20	—

b) Hallamos, a continuación, la recta de tendencia de las medias móviles,

$$\bar{y}_t = a + b \cdot t,$$

con la aplicación del criterio de los mínimos cuadrados, cuyos cálculos de apoyo aparecen recogidos en la tabla siguiente:

t	$\bar{y}_t$	t^2	$\bar{y}_t^2$	$t \cdot \bar{y}_t$
1	7	1	49	7
2	8	4	64	16
3	9	9	81	27
4	11	16	121	44
5	14	25	196	70
6	16	36	256	96
7	17	49	289	119
8	19	64	361	152
9	21	81	441	189

t	$\overline{y}_t$	t^2	$\overline{y}_t^2$	$t \cdot \overline{y}_t$
10	24	100	576	240
11	26	121	676	286
12	30	144	900	360
13	34	169	1 156	442
14	38	196	1 444	532
15	42	225	1 764	630
16	46	256	2 116	736
17	50	289	2 500	850
18	55	324	3 025	990
171	**467**	**2 109**	**16 015**	**5 786**

Se obtienen, así,

$$S_{t,\overline{y}_t} = \frac{1}{N}\sum t \cdot \overline{y}_t - \overline{t} \cdot \overline{\overline{y}} = \frac{5\,786}{18} - \frac{171}{18} \cdot \frac{467}{18} = 321{,}44 - 9{,}5 \cdot 25{,}94 = 75{,}01$$

y

$$S_t^2 = \frac{1}{N}\sum t^2 - \overline{t}^2 = \frac{2\,109}{18} - 9{,}5^2 = 26{,}91,$$

donde N es el número de años e $\overline{\overline{y}}$ es la media de las observaciones de la serie de medias móviles.

En definitiva, los coeficientes de la recta de tendencia son

$$b = \frac{S_{t,\overline{y}_t}}{S_t^2} = \frac{75{,}01}{26{,}91} = 2{,}787$$

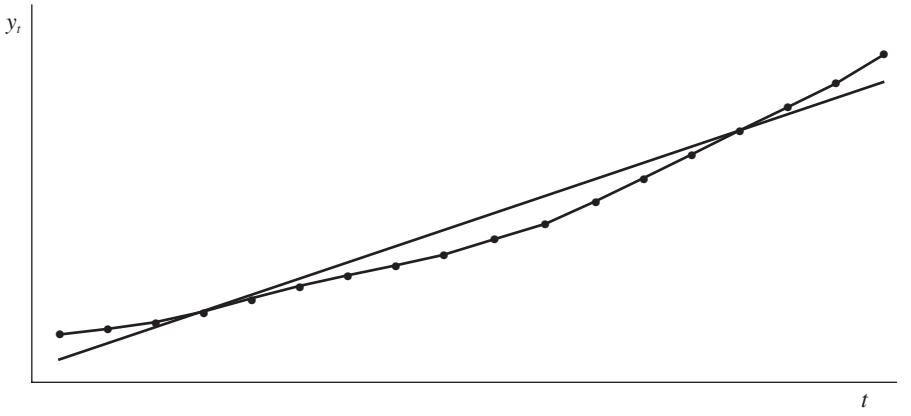
y

$$a = \overline{\overline{y}} - b \cdot \overline{t} = 25{,}94 - 2{,}787 \cdot 9{,}5 = 0{,}5365,$$

con lo cual, la recta de tendencia es

$$\overline{y}_t = 0{,}5365 + 2{,}787 \cdot t.$$

En la siguiente gráfica se representan la serie de medias móviles y la recta de tendencia.



5.6 Demuéstrese que, utilizando un esquema multiplicativo,

$$\sum_{j=1}^{k} e_{\cdot j} = k.$$

SOLUCIÓN

Por definición de componente estacional, para cualquier subperiodo j, se tiene que

$$e_{\cdot j} = \frac{\overline{y}'_{\cdot j}}{\overline{y}'},$$

donde, $\overline{y}'_{\cdot j}$ es la media del subperiodo e $\overline{y}'$ la media global, esto es, la media de las medias anteriores. Estas medias son las que se derivan de los métodos de obtención de la componente estacional a partir de un procedimiento de ajuste de tendencia o de medias móviles.

Por tanto,

$$\sum_{j=1}^{k} e_{\cdot j} = \sum_{j=1}^{k} \frac{\overline{y}'_{\cdot j}}{\overline{y}'} = \frac{1}{\overline{y}'} \sum_{j=1}^{k} \overline{y}'_{\cdot j}.$$

Ahora bien,

$$\overline{y}' = \frac{1}{k} \sum_{j=1}^{k} \overline{y}'_{\cdot j},$$

con lo cual, despejando,

$$\sum_{j=1}^{k} \overline{y}'_{\cdot j} = k \cdot \overline{y}',$$

y sustituyendo, resulta que

$$\sum_{j=1}^{k} e_{\cdot j} = \frac{k \cdot \overline{y}'}{\overline{y}'} = k.$$

5.7 Demuéstrese que, utilizando un esquema aditivo,

$$\sum_{j=1}^{k} e_{\cdot j} = 0.$$

SOLUCIÓN

Por definición de componente estacional se tiene que, para cualquier subperiodo j,

$$e_{\cdot j} = \overline{y}'_{\cdot j} - \overline{y}',$$

donde $\overline{y}'_{\cdot j}$ e $\overline{y}'$ son, respectivamente, la media del subperiodo y la media global correspondientes al método de obtención de la tendencia utilizado. Por tanto, sustituyendo y operando con sumatorios, se obtiene que

$$\sum_{j=1}^{k} e_{\cdot j} = \sum_{j=1}^{k} (\overline{y}'_{\cdot j} - \overline{y}') = \sum_{j=1}^{k} \overline{y}'_{\cdot j} - \sum_{j=1}^{k} \overline{y}' = \sum_{j=1}^{k} \overline{y}'_{\cdot j} - k \cdot \overline{y}' = 0,$$

ya que

$$\sum_{j=1}^{k} \overline{y}'_{\cdot j} = k \cdot \overline{y}',$$

según se demostró en el problema anterior.

5.8 Con objeto de organizar la zona de aparcamientos en un gran centro comercial, situado a las afueras de una ciudad, la dirección del centro ha estudiado la evolución del número de vehículos, en miles, que han estacionado en el periodo 2000-2004, obteniéndose los siguientes datos.

Trimestres	2000	2001	2002	2003	2004
1	16	15	17	20	30
2	20	24	25	34	40
3	25	25	27	32	60
4	35	40	51	58	70

Elimínese la influencia de la componente estacional, suponiendo un esquema multiplicativo.

SOLUCIÓN

Antes de *desestacionalizar* la serie, es decir, de eliminar la componente estacional, comprobaremos que las componentes de la misma se relacionan bajo un esquema multiplicativo; para ello se recogen en la siguiente tabla las medias y las desviaciones típicas de las observaciones de cada año.

Años	Medias	Desviaciones típicas
2000	24	7,11
2001	26	8,97
2002	30	12,69
2003	36	13,78
2004	50	15,81

Por ejemplo, la media del año 2003, que toma el valor 36, se ha obtenido calculando la media aritmética de las observaciones 20, 34, 32 y 58, siendo 13,78 su desviación típica.

Se comprueba que el coeficiente de regresión de la recta de regresión de las desviaciones típicas sobre las medias es $b = 0{,}31$, valor indicativo de que la desviación típica crece moderadamente al crecer la media, aumentando, por tanto, la amplitud de las oscilaciones de la serie a lo largo del tiempo, lo cual justifica la hipótesis de un esquema multiplicativo.

A partir de aquí procederemos a eliminar la componente estacional de la serie, mediante el *método de las relaciones de las medias de cada subperiodo*, en este caso trimestres, *con respecto a la tendencia*. A grandes rasgos, este procedimiento consiste en comparar los valores medios de cada subperiodo con la media de todas las observaciones, considerando que, si no existiera estacionalidad, ambos valores coincidirían. Con el fin de eliminar la influencia que el paso del tiempo tiene en la estacionalidad, en primer lugar, se corrigen estas cantidades a partir de un ajuste de tendencia realizado previamente.

La descripción del método requiere la formalización de los conceptos. Así, supondremos que y_{ij} es la observación genérica de la serie de tiempo, donde i, $i = 1, ..., N$, es el periodo al que está referida la observación y j, $j = 1, ..., k$, el subperiodo dentro periodo inicial. Generalmente los periodos son años, refiriéndose el primer subíndice al año concreto al que corresponde la observación; los subperiodos suelen ser meses, trimestres, cuatrimestres o semestres.

La primera etapa del procedimiento de desestacionalización consiste en obtener la ecuación de tendencia, para lo cual calculamos los valores medios de cada periodo, media aritmética de las observaciones referidas al mismo, según la expresión genérica:

$$\bar{y}_{i\cdot} = \frac{1}{k} \sum_{j=1}^{k} y_{ij},$$

que, en esta ocasión, se traduce en

$$\bar{y}_{00\cdot} = \frac{16 + 20 + 25 + 35}{4} = 24$$

$$\bar{y}_{01\cdot} = \frac{15 + 24 + 25 + 40}{4} = 26$$

$$\bar{y}_{02\cdot} = \frac{17 + 25 + 27 + 51}{4} = 30$$

$$\bar{y}_{03\cdot} = \frac{20 + 34 + 32 + 58}{4} = 36$$

$$\bar{y}_{04\cdot} = \frac{30 + 40 + 60 + 70}{4} = 50.$$

Obsérvese que estos valores ya fueron hallados al principio del problema a la hora de discutir el tipo de esquema de relación entre las componentes de la serie de tiempo.

A continuación, se estiman los parámetros de la recta de regresión mínimo cuadrática de los valores medios anuales por trimestre con respecto al tiempo,

$$\bar{y}'_{i\cdot} = a + b \cdot i',$$

donde $i' = i - 2002$.

Sirva la siguiente tabla de apoyo en la realización de los cálculos necesarios:

i	$i' = i - 2002$	$\bar{y}_{i\cdot}$	i'^2	$\bar{y}^2_{i\cdot}$	$\bar{y}_{i\cdot} \cdot i'$
2000	−2	24	4	576	−48
2001	−1	26	1	676	−26
2002	−0	30	0	900	0
2003	1	36	1	1 296	36
2004	2	50	4	2 500	100
	0	**166**	**10**	**5 948**	**62**

Como $\sum i' = 0$, a partir del sistema de ecuaciones normales, se obtiene, según vimos en el problema **5.1**, el sistema simplificado:

$$\sum \bar{y}_{i\cdot} = N \cdot a$$

$$\sum i' \cdot \bar{y}_{i\cdot} = b \sum i'^2,$$

cuya resolución conduce a los valores estimados:

$$b = \frac{\sum i' \cdot \bar{y}_{i\cdot}}{\sum i'^2} = \frac{62}{10} = 6{,}2,$$

pendiente y

$$a = \frac{1}{N} \sum \bar{y}_{i\cdot} = \frac{166}{5} = 33{,}2,$$

término independiente de la recta de tendencia, respectivamente, siendo, por tanto,

$$\bar{y}_{i\cdot} = 33{,}2 + 6{,}2 \cdot (i - 2002),$$

la ecuación de dicha recta.

La siguiente etapa del método consiste en calcular las medias de cada subperiodo, $\bar{y}_{\cdot j}$, que, en la situación que nos ocupa, serán las medias trimestrales:

$$\bar{y}_{\cdot 1} = \frac{1}{5}(16 + 15 + 17 + 20 + 30) = 19{,}6$$

$$\bar{y}_{\cdot 2} = \frac{1}{5}(20 + 24 + 25 + 34 + 40) = 28{,}6$$

$$\bar{y}_{\cdot 3} = \frac{1}{5}(25 + 25 + 27 + 32 + 60) = 33{,}8$$

$$\bar{y}_{\cdot 4} = \frac{1}{5}(35 + 40 + 51 + 58 + 70) = 50{,}8.$$

A continuación se corrigen las medias de cada subperiodo con la *eliminación de la variación debida al paso del tiempo*. Esta etapa del método consiste en restar a cada media la proporción que, sobre el efecto total del periodo, representa el hecho de encontrarse en un subperiodo concreto; el resultado de esta corrección son las llamadas *medias corregidas de cada subperiodo*:

$$\bar{y}'_{\cdot j} = \bar{y}_{\cdot j} - b \cdot \frac{j - 1}{k}.$$

Obsérvese que b es la pendiente de la ecuación de tendencia, es decir, el incremento que sufre el valor medio del periodo cuando la variable tiempo aumenta una unidad, es decir, un periodo, con lo cual, al subperiodo j le corresponde una proporción del incremento total, b, igual a $(j - 1)/k$; así, por ejemplo, al primer subperiodo le corresponde una proporción de incremento igual a cero, pues todavía no ha transcurrido ningún subperiodo, siendo, por ello, $\bar{y}'_{\cdot 1} = \bar{y}_{\cdot 1}$. Merece también la pena reseñar que este proceso de corrección presupone que el incremento total se reparte uniformemente a lo largo del periodo.

La medias trimestrales corregidas son

$$\overline{y}'_{\cdot 1} = \overline{y}_{\cdot 1} - \frac{6{,}2}{4}(1-1) = 19{,}6$$

$$\overline{y}'_{\cdot 2} = \overline{y}_{\cdot 2} - \frac{6{,}2}{4}(2-1) = 27{,}05$$

$$\overline{y}'_{\cdot 3} = \overline{y}_{\cdot 3} - \frac{6{,}2}{4}(3-1) = 30{,}7$$

$$\overline{y}'_{\cdot 4} = \overline{y}_{\cdot 4} - \frac{6{,}2}{4}(4-1) = 46{,}15.$$

Seguidamente habremos de comparar cada una de estas medias corregidas con la media de todas ellas, *media global corregida,*

$$\overline{y}' = \frac{1}{k}\sum_{j=1}^{k} \overline{y}'_{\cdot j},$$

ya que, si no existiera estacionalidad, es decir, si no influyera el hecho de que la observación pertenezca a uno u otro periodo, todas las medias corregidas serían iguales entre sí y, en consecuencia, iguales a la media global.

Así, en la última etapa del método se calcula la proporción que sobre la media global representa cada una de las medias corregidas de cada subperiodo, esto es,

$$e_{\cdot j} = \frac{\overline{y}'_{\cdot j}}{\overline{y}'},$$

siendo esta proporción de aumento o disminución debida, precisamente, a la existencia de estacionalidad.

A partir de estos cocientes se obtienen el índice de variación estacional para cada subperiodo, cuya expresión genérica es

$$I_j = e_{\cdot j} \cdot 100,$$

porcentaje de aumento o disminución sobre el valor medio global corregido que tiene una observación por pertenecer al subperiodo j.

El promedio de las medias trimestrales corregidas, esto es, la media global corregida, es, en este caso,

$$\overline{y}' = \frac{\overline{y}'_{\cdot 1} + \overline{y}'_{\cdot 2} + \overline{y}'_{\cdot 3} + \overline{y}'_{\cdot 4}}{4} = 30{,}875,$$

con lo cual, las componentes estacionales de los subperiodos son

$$e_{\cdot 1} = \frac{\overline{y}'_{\cdot 1}}{\overline{y}'} = \frac{19{,}6}{30{,}875} = 0{,}635$$

$$e_{\cdot 2} = \frac{\overline{y}'_{\cdot 2}}{\overline{y}'} = \frac{27{,}05}{30{,}875} = 0{,}876$$

$$e_{\cdot 3} = \frac{\overline{y}'_{\cdot 3}}{\overline{y}'} = \frac{30{,}7}{30{,}875} = 0{,}994$$

$$e_{\cdot 4} = \frac{\overline{y}'_{\cdot 4}}{\overline{y}'} = \frac{46{,}15}{30{,}875} = 1{,}495,$$

y los índices de variación estacional:

$$I_1 = e_{\cdot 1} \cdot 100 = 0{,}635 \cdot 100 = 63{,}5$$

$$I_2 = e_{\cdot 2} \cdot 100 = 0{,}876 \cdot 100 = 87{,}6$$

$$I_3 = e_{\cdot 3} \cdot 100 = 0{,}994 \cdot 100 = 99{,}4$$

$$I_4 = e_{\cdot 4} \cdot 100 = 1{,}495 \cdot 100 = 149{,}5.$$

La interpretación de estos índices es clara. Por ejemplo, el hecho de que una observación pertenezca al tercer trimestre hace que tenga un valor de un $100 - 99{,}4 = 0{,}6$ por ciento inferior al valor que tendría en el caso de que no existiera estacionalidad; igualmente, si la observación corresponde al cuarto trimestre, su valor es un $149{,}5 - 100 = 49{,}5$ por ciento superior al que tendría si no influyera el hecho de pertenecer a un trimestre u otro.

Además, según se demostró en **5.6**, la suma de las componentes estacionales es, en este caso, 4, número de subperiodos de cada periodo.

Dividiendo cada observación de la serie de tiempo por la componente estacional de cada uno de los trimestres, resultan las observaciones de la serie desestacionalizada. De este modo, la observación del trimestre *j* del año *i* se obtiene mediante la expresión:

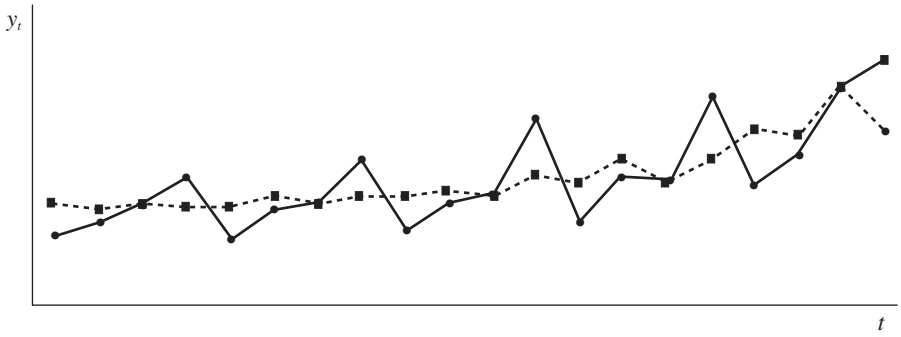
$$\frac{y_{ij}}{e_{\cdot j}},$$

que, aplicada a los datos del problema, proporciona los resultados que figuran en la siguiente tabla.

Trimestres	2000	2001	2002	2003	2004
1	25,20	23,62	26,77	31,50	47,24
2	22,83	27,40	28,54	38,81	45,66
3	25,15	25,15	27,16	32,19	60,36
4	23,41	26,76	34,11	38,80	46,82

Así, por ejemplo, la observación del segundo trimestre del año 2003, una vez eliminada la influencia de la componente estacional, es 34/0,876 = 38,81.

La representación de ambas series —la original y la desestacionalizada— muestra que esta última es más suave que la serie inicial como consecuencia de la eliminación de las oscilaciones debidas a la componente estacional.



5.9 La recta de tendencia de las ventas medias anuales por semestre, en miles de euros, de una empresa se ha obtenido a partir de los datos del periodo 2002-2004:

$$\bar{y}_{i\cdot} = 53{,}667 + 2{,}5(i - 2003).$$

Se sabe, además, que la varianza de dichos valores medios es 4,186 y que la varianza de la variable transformada, $i' = i - 2003$, es 0,667.

a) Obténgase una predicción de la venta media por semestre para el año 2005.

b) Analícese la fiabilidad del resultado.

SOLUCIÓN

a) Sustituyendo en la recta de ajuste de la tendencia la variable i por el valor 2005, se obtiene

$$\bar{y}^*_{05\cdot} = 53{,}667 + 2{,}5(2005 - 2003) = 58{,}67 \text{ miles de euros},$$

predicción de las ventas media por semestre para el año 2005.

b) La fiabilidad del resultado se estudia mediante el coeficiente de determinación lineal:

$$r^2 = \frac{S^2_{i',\bar{y}_{i\cdot}}}{S^2_{i'} \cdot S^2_{\bar{y}_{i\cdot}}}.$$

No se dispone de la covarianza, aunque sí implícitamente, ya que el coeficiente de la recta de tendencia es

$$b = \frac{S_{i',\bar{y}_{i\cdot}}}{S^2_{i'}},$$

con lo cual, despejando, se obtiene que

$$S_{i',\bar{y}_{i\cdot}} = b \cdot S^2_{i'},$$

esto es,

$$S_{i',\bar{y}_{i\cdot}} = 2{,}5 \cdot 0{,}667 = 1{,}667,$$

siendo, en definitiva, el coeficiente de determinación lineal:

$$r^2 = \frac{1{,}667^2}{0{,}667 \cdot 4{,}186} = 0{,}995,$$

valor próximo a 1, de lo que se concluye que la predicción es correcta.

5.10 El número de personas, en miles, que han realizado sus compras en un pequeño comercio de una ciudad en el periodo 2001-2004 ha sido:

Trimestres	2001	2002	2003	2004
1	4,3	4,7	5,2	5,5
2	2,8	3,1	3,6	4
3	1,5	1,9	2,4	2,9
4	4,4	4,9	5,6	6

a) Indíquese cuál es el tipo de esquema que relaciona las componentes de esta serie de tiempo.

b) Hállese la tendencia de la serie gráficamente.

c) Obténgase la componente estacional de la serie, utilizando el método de la relación a la tendencia.

d) Elimínese la componente estacional de la serie de tiempo.

SOLUCIÓN

a) En la siguiente tabla se recogen las medias y desviaciones típicas de los valores medios anualespor trimestre:

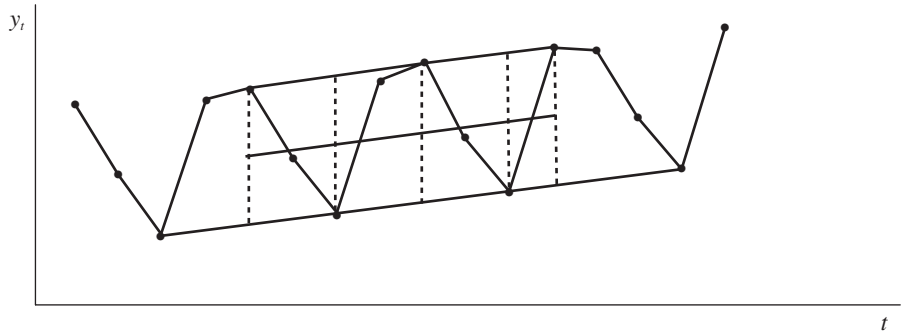
Años	Medias	Desviaciones típicas
2001	3,25	1,193
2002	3,65	1,228
2003	4,20	1,281
2004	4,60	1,227

Como puede comprobar el lector, la regresión lineal entre estas dos variables —medias y desviaciones típicas— conduce a un coeficiente de regresión $b = 0{,}035$, con lo cual, el ajuste lineal corresponde a una recta *prácticamente* paralela al eje horizontal. Este hecho prueba que la amplitud de las oscilaciones se mantiene aproximadamente constante a lo largo del tiempo, por lo que el esquema que debe considerarse en este caso es el aditivo.

b) Antes de estimar la recta de tendencia mediante el criterio de los mínimos cuadrados, vamos a aproximar esta componente *gráficamente*.

Para ello, se trazan dos líneas poligonales: una de ellas uniendo los puntos *máximos* de la serie y la otra los puntos *mínimos*, tal y como se recoge en la figura siguiente. Se dibujan, después, los segmentos de distancia entre las dos líneas, esto es, perpendiculares al eje de abscisas partiendo de cada máximo y de cada mínimo. La aproximación gráfica a la tendencia es la línea poligonal que une los puntos medios de dichos segmentos.

Téngase en cuenta que no se consideran las observaciones primera y última porque, al no disponerse de datos anteriores ni posteriores, no sabemos si constituyen máximos de la serie.



c) Para estimar los parámetros de la recta de tendencia, y dado que el número de años es una cantidad par, realizaremos el cambio de variable:

$$i' = 2(i - o),$$

siendo o la media aritmética de los dos años centrales:

$$o = \frac{2002 + 2003}{2} = 2\,002{,}5.$$

Utilizando este cambio de variable sigue cumpliéndose que $\sum i' = 0$, con la consiguiente simplificación de los cálculos, pero hay que tener en cuenta que ahora la ecuación de tendencia responde a la expresión:

$$\bar{y}_{i\cdot} = a + 2 \cdot b(i - 2\,002{,}5),$$

por lo cual, la pendiente de la recta o incremento de los valores medios anuales por trimestre debidos al transcurso de un año, cantidad por la que, proporcionalmente al trimestre de que se trate, habremos de corregir las medias trimestrales es, en esta situación, igual a $2 \cdot b$.

i	i'	$\bar{y}_{i\cdot}$	i'^2	$\bar{y}_{i\cdot} \cdot i'$
2001	−3	3,25	9	−9,75
2002	−1	3,65	1	−3,65
2003	1	4,20	1	4,20
2004	3	4,60	9	13,80
	0	**15,70**	**20**	**4,60**

La tabla anterior sirve de apoyo para la estimación del parámetro b, partiendo del sistema de ecuaciones normales:

$$\sum \bar{y}_{i\cdot} = N \cdot a$$

$$\sum i' \cdot \bar{y}_{i\cdot} = b \sum i'^2,$$

correspondiente a la ecuación

$$\bar{y}_{i\cdot} = a + b \cdot i',$$

sistema del que resulta el valor

$$b = \frac{4{,}6}{20} = 0{,}23,$$

siendo, por tanto, la pendiente de la recta de tendencia igual a

$$2 \cdot 0{,}23 = 0{,}46.$$

Mediante este valor se corrigen las medias trimestrales, $\overline{y}_{\cdot 1} = 4{,}925$, $\overline{y}_{\cdot 2} = 3{,}375$, $\overline{y}_{\cdot 3} = 2{,}175$ y $\overline{y}_{\cdot 4} = 5{,}225$, obteniéndose las siguientes medias trimestrales corregidas:

$$\overline{y}'_{\cdot 1} = 4{,}925$$

$$\overline{y}'_{\cdot 2} = 3{,}375 - \frac{0{,}46}{4} = 3{,}26$$

$$\overline{y}'_{\cdot 3} = 2{,}175 - \frac{0{,}46}{2} = 1{,}945$$

$$\overline{y}'_{\cdot 4} = 5{,}225 - \frac{3 \cdot 0{,}46}{4} = 4{,}88,$$

cuya media global corregida es

$$\overline{y}' = \frac{4{,}925 + 3{,}26 + 1{,}945 + 4{,}88}{4} = 3{,}7525.$$

Por tratarse de un esquema aditivo, la componente estacional se calcula por diferencia:

$$e_{\cdot 1} = \overline{y}'_{\cdot 1} - \overline{y}' = 4{,}925 - 3{,}7525 = 1{,}1725$$

$$e_{\cdot 2} = \overline{y}'_{\cdot 2} - \overline{y}' = 3{,}26 - 3{,}7525 = -0{,}4925$$

$$e_{\cdot 3} = \overline{y}'_{\cdot 3} - \overline{y}' = 1{,}945 - 3{,}7525 = -1{,}8075$$

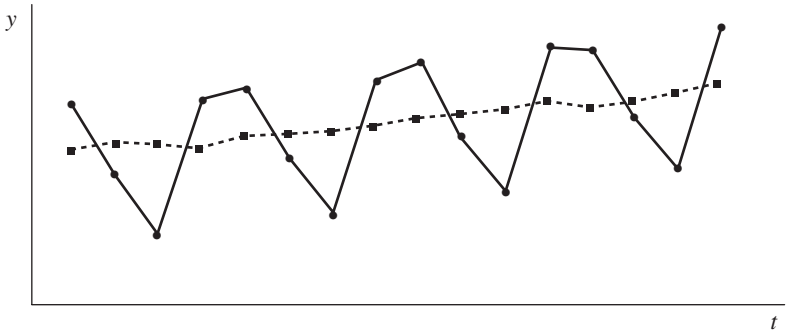
$$e_{\cdot 4} = \overline{y}'_{\cdot 4} - \overline{y}' = 4{,}88 - 3{,}7525 = 1{,}1275.$$

d) Para eliminar la estacionalidad en la serie de tiempo, se resta de cada observación la correspondiente componente estacional. El resultado de este proceso se recoge en la siguiente tabla:

Trimestre	2001	2002	2003	2004
1	4,3 – 1,1725 = 3,1275	4,7 – 1,1725 = 3,5275	5,2 – 1,1725 = 4,0275	5,5 – 1,1725 = 4,3275
2	2,8 + 0,4925 = 3,2925	3,1 + 0,4925 = 3,5925	3,6 + 0,4925 = 4,0925	4,0 + 0,4925 = 4,4925
3	1,5 + 1,8075 = 3,3075	1,9 + 1,8075 = 3,7075	2,4 + 1,8075 = 4,2075	2,9 + 1,8075 = 4,7075
4	4,4 – 1,1275 = 3,2725	4,9 – 1,1275 = 3,7725	5,6 – 1,1275 = 4,4725	6,0 – 1,1275 = 4,8725

Con el procedimiento empleado hemos eliminado la componente estacional, pues la serie obtenida carece de oscilaciones trimestrales. Además, el aspecto tan suavizado de la representación

de la serie desestacionalizada, muy próximo a la recta de tendencia, hace pensar que las oscilaciones que presenta la serie original son debidas casi en su totalidad a la componente estacional, siendo prácticamente inapreciable la influencia de las componentes cíclica y accidental, al menos para el periodo —sin duda corto— considerado.



5.11 El número de personas, en miles, que ha visitado un museo ubicado en una región norteña durante el periodo 2002-2004 ha sido:

Cuatrimestre	2002	2003	2004
1	30	40	35
2	50	60	50
3	35	52	20

a) La fundación que regenta el museo se está planteando la posibilidad de organizar una exposición en el año 2005 dedicada a los pintores del siglo XX cuyas obras han estado inspiradas en la región. ¿Qué cuatrimestre del año parece el más adecuado para llevarla a cabo?

b) Calcúlese una predicción para el número de visitantes en el año 2005 y para el cuatrimestre en el que parece más razonable celebrar la exposición.

SOLUCIÓN

a) Un factor que puede influir considerablemente en la decisión sobre el cuatrimestre más idóneo para organizar la exposición es conocer cuál es el cuatrimestre con mayor afluencia de visitantes. Con el fin de disponer de dicha información, realizamos un estudio de la estacionalidad de la serie, mediante el método de las relaciones de las media cuatrimestrales con respecto a la tendencia.

Los resultados de las distintas etapas de este método: cálculo de las medias cuatrimestrales y de las medias cuatrimestrales corregidas, así como la obtención de la componente estacional y de los índices de variación estacional, considerando que el esquema es multiplicativo, se recogen en la tabla siguiente.

Cuatrimestre	Media	Media corregida	Componente estacional
1	35	35	0,84
2	53,33	53,885	1,28
3	35,67	36,78	0,88

Téngase en cuenta que las medias cuatrimestrales corregidas son el resultado de eliminar la tendencia en las medias cuatrimestrales con la pendiente de la recta de regresión de las valores medios anuales por cuatrimestre con respecto al tiempo, que, como puede comprobar el lector, es $b = -1{,}665$.

De multiplicar por cien la componente estacional de cada uno de los cuatrimetres, resultan los índices de variación estacional:

$$I_1 = 0{,}84 \cdot 100 = 84$$
$$I_2 = 1{,}29 \cdot 100 = 128$$
$$I_3 = 0{,}88 \cdot 100 = 88.$$

A la vista de los datos parece adecuado celebrar la exposición en el segundo cuatrimestre, pues su índice de variación estacional es el mayor de todos, indicando que dicho cuatrimestre será el de mayor afluencia de visitantes.

b) La estimación de la ecuación de tendencia que, según puede comprobarse, es

$$\overline{y}_{i\cdot} = 41{,}33 - 1{,}665(i - 2003),$$

permite dar una predicción para el número medio de visitantes correspondientes al año 2005. En efecto, sustituyendo el año 2005 en dicha recta, se tiene que

$$\overline{y}^{*}_{05\cdot} = 41{,}33 - 1{,}665(2005 - 2003) = 38 \text{ mil visitantes},$$

número medio de visitantes *por cuatrimestre* previsto para el año 2005. La predicción del *total* de visitantes se halla multiplicando por 3, número de cuatrimestres, dicho valor medio:

$$y^{*}_{05\cdot} = 3 \cdot 38 = 114 \text{ mil visitantes}.$$

Por lo que se refiere al segundo cuatrimestre, subperiodo con mayor afluencia de visitantes, para la estimación de la cifra prevista se ha de considerar el hecho de que existe estacionali-

dad, según se ha comprobado con los datos del periodo 2002-2004. Ello obliga a multiplicar la predicción del número medio de visitantes por cuatrimestre para el año 2005 por la correspondiente componente estacional:

$$y^*_{05,2} = \bar{y}^*_{05\cdot} \cdot e_{\cdot 2} = 38 \cdot 1{,}29 = 49{,}02 \text{ miles de visitantes.}$$

Téngase en cuenta que esta predicción se realiza bajo la hipótesis de que la estacionalidad obtenida con los datos del periodo 2002-2004 se mantiene en el año 2005.

5.12 La tabla adjunta recoge los beneficios trimestrales, en miles de euros, para los años 2002, 2003 y 2004 de una red de comercios dedicados a la venta de artículos de playa:

Años	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2002	10	40	30	20
2003	10	52	30	40
2004	15	45	40	20

a) Obténgase la serie de tendencia, aplicando el método de las medias móviles.

b) Calcúlense los índices de variación estacional por el método de la razón a la media móvil, suponiendo un esquema multiplicativo.

c) Desestacionalícese la serie original.

SOLUCIÓN

a) Si se toman 4 observaciones para formar las medias móviles, quedarán promediadas las variaciones de cada estación y, por tanto, eliminada la componente estacional. Obtenemos, así, la siguiente serie de medias móviles:

$$\bar{y}_{2.5} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \frac{10 + 40 + 30 + 20}{4} = 25$$

$$\bar{y}_{3.5} = \frac{y_2 + y_3 + y_4 + y_5}{4} = \frac{40 + 30 + 20 + 10}{4} = 25$$

$$\bar{y}_{4.5} = \frac{y_3 + y_4 + y_5 + y_6}{4} = \frac{30 + 20 + 10 + 52}{4} = 28$$

$$\bar{y}_{5.5} = \frac{y_4 + y_5 + y_6 + y_7}{4} = \frac{20 + 10 + 52 + 30}{4} = 28$$

$$\bar{y}_{6.5} = \frac{y_5 + y_6 + y_7 + y_8}{4} = \frac{10 + 52 + 30 + 40}{4} = 33$$

$$\bar{y}_{7.5} = \frac{y_6 + y_7 + y_8 + y_9}{4} = \frac{52 + 30 + 40 + 15}{4} = 34{,}25$$

$$\bar{y}_{8.5} = \frac{y_7 + y_8 + y_9 + y_{10}}{4} = \frac{30 + 40 + 15 + 45}{4} = 32{,}5$$

$$\bar{y}_{9.5} = \frac{y_8 + y_9 + y_{10} + y_{11}}{4} = \frac{40 + 15 + 45 + 40}{4} = 35$$

$$\bar{y}_{10.5} = \frac{y_9 + y_{10} + y_{11} + y_{12}}{4} = \frac{15 + 45 + 40 + 20}{4} = 30.$$

Al haberse promediado un número par de observaciones, las medias móviles calculadas aparecen *descentradas*. Promediándolas nuevamente se obtienen las medias móviles *centradas*:

$$\bar{\bar{y}}_3 = \frac{\bar{y}_{2,5} + \bar{y}_{3,5}}{2} = \frac{25 + 25}{2} = 25$$

$$\bar{\bar{y}}_4 = \frac{\bar{y}_{3,5} + \bar{y}_{4,5}}{2} = \frac{25 + 28}{2} = 26{,}5$$

$$\bar{\bar{y}}_5 = \frac{\bar{y}_{4,5} + \bar{y}_{5,5}}{2} = \frac{28 + 28}{2} = 28$$

$$\bar{\bar{y}}_6 = \frac{\bar{y}_{5,5} + \bar{y}_{6,5}}{2} = \frac{28 + 33}{2} = 30{,}5$$

$$\bar{\bar{y}}_7 = \frac{\bar{y}_{6,5} + \bar{y}_{7,5}}{2} = \frac{33 + 34{,}25}{2} = 33{,}625$$

$$\bar{\bar{y}}_8 = \frac{\bar{y}_{7,5} + \bar{y}_{8,5}}{2} = \frac{34{,}25 + 32{,}5}{2} = 33{,}375$$

$$\bar{\bar{y}}_9 = \frac{\bar{y}_{8,5} + \bar{y}_{9,5}}{2} = \frac{32{,}5 + 35}{2} = 33{,}75$$

$$\bar{\bar{y}}_{10} = \frac{\bar{y}_{9,5} + \bar{y}_{10,5}}{2} = \frac{35 + 30}{2} = 32{,}5.$$

Hemos obtenido, de esta manera, una serie de observaciones de la denominada componente extraestacional o componente a largo plazo.

b) Dividiendo cada dato de la serie original, y_t, por la correspondiente media móvil, $\bar{\bar{y}}_t$, (teniendo en cuenta las observaciones perdidas por el procedimiento), resulta una nueva

serie, cuyas observaciones denotaremos por y'_t, que recoge conjuntamente las componentes estacional y accidental. En la siguiente tabla aparece esta nueva serie:

Años	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2002	—	—	30/25 = 1,2	20/26,5 = 0,75
2003	10/28 = 0,36	52/30,5 = 1,71	30/33,625 = 0,89	40/33,375 = 1,20
2004	15/33,75 = 0,44	45/32,5 = 1,38	—	—

Por último, para eliminar la componente accidental de la serie obtenida, se calculan, en primer lugar, las medias aritméticas de las observaciones de cada uno de los cuatro trimestres; con este modo de actuar estamos suponiendo implícitamente que las componentes accidentales de los trimestres se compensan unas con otras, siendo su media aritmética igual a cero:

$$\overline{y}'_{\cdot 1} = \frac{0{,}36 + 0{,}44}{2} = 0{,}40$$

$$\overline{y}'_{\cdot 2} = \frac{1{,}71 + 1{,}38}{2} = 1{,}55$$

$$\overline{y}'_{\cdot 3} = \frac{1{,}2 + 0{,}89}{2} = 1{,}05$$

$$\overline{y}'_{\cdot 4} = \frac{0{,}75 + 1{,}20}{2} = 0{,}98.$$

En segundo lugar, se promedian las medias anteriores,

$$\overline{y}' = \frac{0{,}40 + 1{,}55 + 1{,}05 + 0{,}98}{4} = 0{,}995.$$

Por último, comparando las medias trimestrales, $\overline{y}'_{\cdot j}$, con su promedio, $\overline{y}'$, y multiplicando por 100 el resultado, se obtienen los índices de variación estacional:

$$I_1 = \frac{\overline{y}'_{\cdot 1}}{\overline{y}'} \cdot 100 = \frac{0{,}40}{0{,}995} \cdot 100 = 40{,}20$$

$$I_2 = \frac{\overline{y}'_{\cdot 2}}{\overline{y}'} \cdot 100 = \frac{1{,}55}{0{,}995} \cdot 100 = 155{,}78$$

$$I_3 = \frac{\overline{y}'_{\cdot 3}}{\overline{y}'} \cdot 100 = \frac{1{,}05}{0{,}995} \cdot 100 = 105{,}53$$

$$I_4 = \frac{\overline{y}'_{\cdot 4}}{\overline{y}'} \cdot 100 = \frac{0{,}98}{0{,}995} \cdot 100 = 98{,}49.$$

c) Para desestacionalizar la serie original hay que dividir cada una de sus observaciones por su respectivo índice de variación estacional expresado en tanto por uno, es decir, por la componente estacional,

$$\frac{y_{ij}}{I_j/100}.$$

En la tabla siguiente se recogen los resultados de este proceso:

Años	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2002	24,88	25,68	28,43	20,31
2003	24,88	33,38	28,43	40,61
2004	37,31	28,89	37,90	20,31

Obsérvese que, por ejemplo, el dato del segundo trimestre del año 2002 se ha obtenido como

$$\frac{y_{02,2}}{I_2/100} = \frac{40}{155{,}78/100} = 25{,}68 \text{ miles de euros.}$$

Comentemos, por último, que, según puede comprobar el lector siguiendo el procedimiento de problemas anteriores, la suposición de un esquema multiplicativo para las observaciones de esta serie está fundamentada, porque del análisis de regresión entre los valores medios y las desviaciones típicas de las observaciones de cada año se obtiene un valor del coeficiente de regresión igual a 0,50, indicativo de una relación ligeramente creciente o positiva entre ambas variables.

5.13 La relación de la media anual de parados de un país, en miles de personas, con respecto a la tendencia es

$$\bar{y}_{i\cdot} = 2\,299{,}5 - 45{,}2(i - 2003).$$

a) ¿Qué indica el signo negativo de la pendiente de esta recta?

b) De los datos observados por cuatrimestres para cada uno de los años, se tienen las siguientes medias cuatrimestrales: $\bar{y}_{\cdot 1} = 2\,362{,}33$, $\bar{y}_{\cdot 2} = 2\,226{,}33$, $\bar{y}_{\cdot 3} = 2\,310$. Obténgase la predicción del número de parados para el segundo cuatrimestre de 2005, suponiendo un esquema multiplicativo.

SOLUCIÓN

a) El signo negativo del coeficiente b, pendiente de la recta de tendencia de la serie de tiempo del número de parados, indica que el número medio anual de parados por cuatrimestre *disminuye* como consecuencia del paso de un año.

b) Sustituyendo el año 2005 en la recta de regresión, se calcula la predicción del valor medio por cuatrimestre de ese año. Así,

$$\overline{y}^{*}_{05\cdot} = 2\,299{,}5 - 45{,}2(2005 - 2003) = 2\,209{,}1 \text{ miles de parados}$$

es el número medio previsto por cuatrimestre para el año 2005, siempre y cuando se considere que la serie temporal analizada mantendrá su estructura, por lo menos, hasta dicho año.

Para obtener la cifra de parados prevista para el segundo cuatrimestre hay que considerar la *posible* existencia de estacionalidad.

Siguiendo el método de las relaciones de las medias cuatrimestrales con respecto a la tendencia, se elimina la tendencia de dichas medias:

$$\overline{y}'_{\cdot 1} = 2\,362{,}33$$

$$\overline{y}'_{\cdot 2} = 2\,226{,}33 + \frac{45{,}2}{3} = 2\,241{,}40$$

$$\overline{y}'_{\cdot 3} = 2\,310 + \frac{45{,}2}{3} \cdot 2 = 2\,340{,}13,$$

siendo el promedio de estas medias cuatrimestrales corregidas igual a

$$\overline{y}' = \frac{\overline{y}'_{\cdot 1} + \overline{y}'_{\cdot 2} + \overline{y}'_{\cdot 3}}{3} = \frac{2\,362 + 2\,211{,}26 + 2\,279{,}87}{3} = 2\,314{,}62.$$

La estacionalidad del segundo cuatrimestre es, por tanto,

$$e_{\cdot 2} = \frac{\overline{y}'_{\cdot 2}}{\overline{y}'} = \frac{2\,241{,}40}{2\,314{,}62} = 0{,}968.$$

Multiplicando esta componente estacional por el valor medio de parados por cuatrimestre previsto para el año 2005, se obtiene una predicción del número de parados del segundo cuatrimestre:

$$y^{*}_{05,2} = 2\,209{,}1 \cdot 0{,}968 = 2\,138{,}40 \text{ miles de parados,}$$

cifra que está por debajo de 2 209,1 miles de parados, cantidad prevista para el segundo cuatrimestre si no hubiera existido estacionalidad.

5.14 Se espera que las ventas totales de una empresa alcancen un montante de 8 millones de euros para el año próximo, considerándose, además, que el sistema de índices de variación estacional es:

Trimestres	1	2	3	4
Índices	130	90	105	75

Si suponemos un esquema multiplicativo, ¿cuáles serán las cifras de ventas previstas para cada trimestre?

SOLUCIÓN

La media de ventas prevista por trimestre para el próximo año es

$$\overline{y}^*_{i\cdot} = \frac{8}{4} = 2 \text{ millones de euros,}$$

cantidad que coincidiría con las ventas de cada uno de los trimestres en el caso de que no hubiera habido estacionalidad.

Utilizando el sistema de índices de variación estacional, se obtienen las cifras de ventas previstas para cada trimestre; así, si $e_{\cdot j}$ es la componente estacional del trimestre genérico, para obtener el correspondiente valor trimestral de las ventas, y^*_{ij}, habrá que calcular:

$$y^*_{ij} = \overline{y}^*_{i\cdot} \cdot e_{\cdot j},$$

lo cual supone, como es sabido, una modificación del valor de $\overline{y}^*_{i\cdot}$, por exceso o por defecto, según el carácter de la estacionalidad.

En este caso, las cifras de ventas, en millones de euros, previstas para cada trimestre son

$$y^*_{i1} = 2 \cdot 1{,}30 = 2{,}6$$

$$y^*_{i2} = 2 \cdot 0{,}90 = 1{,}8$$

$$y^*_{i3} = 2 \cdot 1{,}05 = 2{,}1$$

$$y^*_{i4} = 2 \cdot 0{,}75 = 1{,}5.$$

5.15 En el periodo 2002-2004, la media anual de ingresos por cuatrimestre por la venta de entradas en un cine de una pequeña ciudad, en miles de euros, responde a la siguiente ecuación:

$$\overline{y}_{i\cdot} = 6 + 1{,}2(i - 2003).$$

Se sabe, también, que para el mismo periodo los ingresos medios por año de cada cuatrimestre han sido: 7,16, 2,5 y 8,33, respectivamente. Calcúlese la predicción de ventas para cada uno de los cuatrimestres del año 2005, suponiendo un esquema aditivo.

SOLUCIÓN

Sustituyendo el año 2005 en la recta de regresión de las medias anuales con respecto al tiempo, resulta la predicción de la media anual de ingresos por cuatrimestre para dicho año:

$$\overline{y}^{*}_{05\cdot} = 6 + 1{,}2(2005 - 2003) = 8{,}4 \text{ miles de euros,}$$

siendo tres veces esta cantidad la cifra de ventas prevista para el año 2005, esto es, $3 \cdot 8{,}4 = 25{,}2$ miles de euros. Para repartir este montante en cada uno de los cuatrimestres del año, será necesario conocer la componente estacional, cuya expresión genérica, al tratarse de un esquema aditivo, es

$$e_{\cdot j} = \overline{y}'_{\cdot j} - \overline{y}',$$

donde $\overline{y}'_{\cdot j}$ e $\overline{y}'$ son, respectivamente, la media cuatrimestral corregida genérica y la media global corregida.

Ahora bien, eliminando la tendencia de las medias cuatrimestrales, resultan los siguientes valores corregidos:

$$\overline{y}'_{\cdot 1} = 7{,}16$$

$$\overline{y}'_{\cdot 2} = 2{,}5 - \frac{1{,}2}{3} = 2{,}1$$

$$\overline{y}'_{\cdot 3} = 8{,}33 - \frac{1{,}2}{3} \cdot 2 = 7{,}53,$$

siendo la media global corregida:

$$\overline{y}' = \frac{\overline{y}'_{\cdot 1} + \overline{y}'_{\cdot 2} + \overline{y}'_{\cdot 3}}{3} = \frac{7{,}16 + 2{,}1 + 7{,}53}{3} = 5{,}5966.$$

Por consiguiente, las componentes estacionales resultan ser:

$$e_{\cdot 1} = \overline{y}'_{\cdot 1} - \overline{y}' = 7{,}16 - 5{,}5966 = 1{,}563$$

$$e_{\cdot 2} = \overline{y}'_{\cdot 2} - \overline{y}' = 2{,}1 - 5{,}5966 = -3{,}496$$

$$e_{\cdot 3} = \overline{y}'_{\cdot 3} - \overline{y}' = 7{,}53 - 5{,}5966 = 1{,}933,$$

que, como puede comprobarse, suman 0.

De la aplicación de la expresión genérica:

$$y^{*}_{05,\,j} = y^{*}_{05\cdot} + e_{\cdot j},$$

donde cada observación será igual a la media anual más la componente estacional del respectivo subperiodo, se obtienen las cifras de ventas, en miles de euros, previstas para cada cuatrimestre:

$$y^*_{05,1} = 1{,}563 + 8{,}4 = 9{,}963$$

$$y^*_{05,2} = -3{,}496 + 8{,}4 = 4{,}904$$

$$y^*_{05,3} = 1{,}933 + 8{,}4 = 10{,}333.$$

Hay que apreciar que, tal y como comentábamos al principio del problema, hemos repartido, efectivamente, el total de ventas entre los tres cuatrimestres, ya que la suma de las cantidades obtenidas por cuatrimestre resulta ser igual al total, 25,2:

$$y^*_{05,1} + y^*_{05,2} + y^*_{05,3} = y^*_{05\cdot}.$$

5.16 A partir de datos trimestrales del periodo 2002-2004, se han previsto las cifras de ingresos, en miles de euros, por las ventas de entradas en un zoo para cada trimestre del año 2006: 112,8, 126, 130 y 120. Analícese la estacionalidad de cada trimestre, suponiendo un esquema multiplicativo.

SOLUCIÓN

Para obtener la componente estacional, $e_{\cdot j}$, para cada trimestre del año 2006, hemos de conocer primero —en realidad, predecir— la media anual de ingresos por trimestre de dicho año.

La información proporcionada por el enunciado permite realizar esta predicción, promediando las cifras previstas para cada uno de los trimestres del año 2006 que, evidentemente, han sido halladas utilizando una ecuación de estimación de la tendencia:

$$\frac{y^*_{06,1} + y^*_{06,2} + y^*_{06,3} + y^*_{06,4}}{4} = \overline{y}^*_{06\cdot},$$

esto es,

$$\frac{112{,}8 + 126 + 130 + 120}{4} = 122{,}2 \text{ miles de euros.}$$

Si tenemos en cuenta que la observación de cada trimestre del año 2006 ha sido obtenida considerando la existencia de estacionalidad, bajo la hipótesis de un esquema multiplicativo, es decir,

$$y^*_{06,j} = \overline{y}^*_{06\cdot} \cdot e_{\cdot j},$$

se obtiene, despejando, que

$$e_{\cdot j} = \frac{y^*_{06\cdot}}{\overline{y}^*_{06\cdot}}.$$

En consecuencia, las componentes estacionales de cada cuatrimestre son

$$e_{\cdot 1} = \frac{y^*_{06,1}}{\overline{y}^*_{06\cdot}} = \frac{112{,}8}{122{,}2} = 0{,}9231$$

$$e_{\cdot 2} = \frac{y^*_{06,2}}{\overline{y}^*_{06\cdot}} = \frac{126}{122{,}2} = 1{,}0311$$

$$e_{\cdot 3} = \frac{y^*_{06,3}}{\overline{y}^*_{06\cdot}} = \frac{130}{122{,}2} = 1{,}0638$$

$$e_{\cdot 4} = \frac{y^*_{06,4}}{\overline{y}^*_{06\cdot}} = \frac{120}{122{,}2} = 0{,}9820.$$

5.17 Utilizando un sistema de índices de variación estacional, se ha previsto que el número de visitantes a una exposición, en miles de personas, para cada uno de los trimestres del año 2006 será:

Años	1	2	3	4
N.º de visitantes	8,1	10,8	11,7	5,4

a) ¿En qué sentido influye la existencia de estacionalidad sobre el número de visitantes del segundo trimestre? (Supóngase un esquema aditivo).

b) Hállese el número de visitantes para cada uno de los trimestres, si no hubiera estacionalidad.

SOLUCIÓN

a) Puesto que se conoce la previsión del número de visitantes para cada trimestre del año 2006, puede predecirse el número medio de visitantes por trimestre para dicho año, puesto que, necesariamente,

$$\overline{y}^*_{06\cdot} = \frac{1}{k}\sum_{j=1}^{k} y^*_{06,j}.$$

Por tanto, para los datos del problema,

$$\overline{y}^{*}_{06\cdot} = \frac{8{,}1 + 10{,}8 + 11{,}7 + 5{,}4}{4} = 9 \text{ mil visitantes.}$$

La predicción del número de visitantes del segundo trimestre del año 2006 se ha realizado considerando la existencia de estacionalidad, es decir,

$$y^{*}_{06,2} = \overline{y}^{*}_{06\cdot} + e_{\cdot 2},$$

bajo hipótesis de esquema aditivo.

En consecuencia, despejando,

$$e_{\cdot 2} = y^{*}_{06,2} - \overline{y}^{*}_{06\cdot},$$

es decir,

$$e_{\cdot 2} = 10{,}8 - 9 = 1{,}8 \text{ miles de visitantes.}$$

Si no existiera estacionalidad, la componente estacional sería igual a cero, con lo cual, la cifra media de visitantes por trimestre para el año 2006, 9 mil, habría coincidido con la cifra del segundo trimestre de dicho año. Cualquier desviación por encima o por debajo de cero es indicativo de la influencia que tiene la estacionalidad del trimestre correspondiente, en este caso, el segundo. Dado que $e_{\cdot 2}$ es igual a 1,8, el número de visitantes al museo previsto para el segundo trimestre supera, como sabemos, en 1,8 miles de visitantes al valor medio por trimestre previsto.

b) Como consecuencia de todo lo comentado anteriormente, si no existiera componente estacional, las cifras de visitantes tendrían que ser iguales para todos los trimestres del año 2006 e iguales a la cantidad media de visitantes por trimestre; en este caso 9 mil visitantes.

5.18 Un grupo de empresas de publicidad ha realizado un estudio para el periodo 2000-2004, estimando la ecuación de la tendencia a partir de la media anual del número de artículos adquiridos a través de teletienda, en miles de unidades. De esta estimación se ha hallado que el incremento trimestral del número medio por trimestre de artículos adquiridos mediante este procedimiento de compra es del 15 por ciento, obteniéndose, además, que el esquema es aditivo y que la componente estacional, también en miles de unidades, es:

Trimestres	1	2	3	4
Componente estacional	6,12	−8,4	−15,3	17,58

a) Se espera que las ventas totales para el año 2005 alcancen un montante de 122,4 miles de artículos. ¿Cuáles son las cifras de ventas previstas para cada trimestre de dicho año?

b) Calcúlese la predicción de ventas para el año 2006.

SOLUCIÓN

a) Puesto que para el año 2005 se prevén unas ventas de 122,4 miles de unidades, dividiendo dicha cantidad entre cuatro, resulta la predicción del número medio de unidades por trimestre para dicho año, esto es,

$$\overline{y}^{*}_{05\cdot} = \frac{122{,}4}{4} = 30{,}6 \text{ miles de unidades,}$$

cantidad que coincidiría con la cifra de ventas de cada trimestre, siempre y cuando no existiera estacionalidad. El hecho de contar con una componente estacional nos obliga a *corregir* este valor trimestral por la correspondiente componente estacional. Así, y puesto que el esquema es aditivo, la cifra de ventas del trimestre genérico es

$$y^{*}_{05,j} = \overline{y}^{*}_{05\cdot} + e_{\cdot j}.$$

En definitiva, las cifras de ventas trimestrales, en miles de unidades, son

$$y^{*}_{05,1} = 30{,}6 + 6{,}12 = 36{,}72$$

$$y^{*}_{05,2} = 30{,}6 + (-8{,}4) = 22{,}2$$

$$y^{*}_{05,3} = 30{,}6 + (-15{,}3) = 15{,}3$$

$$y^{*}_{05,4} = 30{,}6 + 17{,}58 = 48{,}18.$$

b) Como consecuencia del transcurso de *un año* el número medio anual de ventas por trimestre aumenta en cuatro veces el incremento debido al paso de un trimestre. Puesto que dicho incremento ha sido del 15 por ciento, es decir, 0,15, se tiene que $0{,}15 \cdot 4 = 0{,}6$ es el aumento anual del valor medio por trimestre, esto es, la pendiente de la recta de tendencia.

En definitiva,

$$\overline{y}^{*}_{06\cdot} = 30{,}6 + 0{,}6 = 31{,}2 \text{ miles de artículos.}$$

Finalmente, las ventas totales previstas para el año 2006 serán de

$$y^{*}_{06\cdot} = 4 \cdot \overline{y}^{*}_{06\cdot} = 124{,}8 \text{ miles de artículos.}$$

5.19 Calcúlense los índices que reflejen la variación estacional de las ventas de un empresa en un cierto año, sabiendo que durante el primer trimestre el nivel de ventas fue un 12 por ciento superior al segundo y que en el segundo y tercer trimestre no hubo estacionalidad.

SOLUCIÓN

Llamando I_1, I_2, I_3 e I_4 a los índices de variación estacional, se tiene, por la información del enunciado, que

$$I_1 = I_2 + 0{,}12 \cdot I_2$$

$$I_2 = 100$$

$$I_3 = 100.$$

De las dos primeras igualdades se deduce, de modo inmediato, que

$$I_1 = 112.$$

Por último, para obtener el valor del índice correspondiente al cuarto trimestre, hay que considerar que la suma de los índices es igual a 400, con lo cual,

$$I_4 = 400 - (I_1 + I_2 + I_3) = 400 - 112 - 100 - 100 = 88.$$

5.20 Descríbase el método de las relaciones de las medias de cada subperiodo respecto a la tendencia cuando el modelo sea estacionario.

SOLUCIÓN

El hecho de que el modelo sea estacionario implica que, en la estimación con el criterio de los mínimos cuadrados de la recta de tendencia a partir de las medias de los periodos, se ha obtenido la ecuación:

$$\bar{y}_{i\cdot} = a,$$

siendo, por tanto, igual a cero el coeficiente de regresión, b, y, en consecuencia, nulo el incremento de los valores medios de los periodos debido al paso de un periodo.

En tal caso, a la hora de calcular la componente estacional, ni las medias de cada subperiodo,

$$\bar{y}_{\cdot j} = \frac{1}{N} \sum_{i=1}^{N} y_{ij},$$

ni la media global,

$$\bar{y} = \frac{1}{k} \sum_{j=1}^{k} \bar{y}_{\cdot j},$$

han de ser corregidas, con lo cual, la componente estacional genérica es

$$e_{\cdot j} = \frac{\bar{y}_{\cdot j}}{\bar{y}},$$

bajo un esquema multiplicativo, y

$$e_{\cdot j} = \bar{y}_{\cdot j} - \bar{y},$$

si el esquema es aditivo.

5.21 Las componentes de la serie de ventas, en miles de euros, de un producto durante el periodo 1990-2004 están relacionadas según un esquema aditivo. Además, dicha serie es estacionaria pero presenta estacionalidad; en concreto, la componente estacional del primer semestre es de 10 mil euros.

Calcúlese la cifra media de ventas de cada semestre, sabiendo que la media de ventas de todo el periodo es de 74 mil euros.

SOLUCIÓN

El hecho de que la serie sea estacionaria significa que no tiene tendencia, con lo cual, a la hora de calcular la componente estacional no hay que corregir las medias semestrales por la influencia del paso del tiempo. Ello implica que la media global, es decir, la media aritmética de las medias semestrales, coincide con la media de las observaciones que, en este caso, es $\bar{y} = 74$.

Por tanto, se tiene, por un lado, que

$$\bar{y} = \frac{\bar{y}_{\cdot 1} + \bar{y}_{\cdot 2}}{2},$$

y, por otro, al tratarse de un esquema aditivo,

$$e_{\cdot 1} = \bar{y}_{\cdot 1} - \bar{y}.$$

En consecuencia, sustituyendo las cantidades conocidas en las relaciones anteriores resulta que

$$74 = \frac{\bar{y}_{\cdot 1} + \bar{y}_{\cdot 2}}{2}$$

y que

$$10 = \bar{y}_{\cdot 1} - 74,$$

con lo cual,

$$\bar{y}_{\cdot 1} = 84$$

y

$$\bar{y}_{\cdot 2} = 64,$$

valores medios por año, en miles de euros, de las ventas del primer y segundo semestre, respectivamente.

Obsérvese, además, que como ha de cumplirse que $e_{\cdot 1} + e_{\cdot 2} = 0$, entonces, la componente estacional del segundo semestre es igual a -10, es decir, el hecho de que una cifra de ventas sea del segundo semestre significa que será 10 mil euros inferior al valor que tendría en el caso de que no existiera estacionalidad.

5.22 El número medio de turistas por trimestre que visitaron cierta estación de esquí en 2001 fue de 18 mil, registrándose idéntica afluencia media en los años 2002 y 2003.

a) ¿Cuál es la ecuación de la recta de tendencia obtenida a partir de los valores medios anuales por trimestre?

b) De las observaciones recogidas para cada uno de los trimestres del periodo considerado se sabe, además, que las medias trimestrales son: 40 mil, 10 mil, 2 mil y 20 mil, respectivamente. Obténgase un sistema de índices de variación estacional, considerando un esquema multiplicativo. ¿Cómo influye la estacionalidad en el número de turistas correspondiente al segundo trimestre?

SOLUCIÓN

a) La ecuación de la recta de tendencia obtenida mediante el criterio de los mínimos cuadrados para el periodo 2001-2003 es

$$\bar{y}_{i\cdot} = a + b \cdot i,$$

donde b, pendiente de la recta, refleja la variación (creciente o decreciente) producida en los valores medios de la variable número de turistas, en miles, debida exclusivamente al paso del tiempo. En este caso, los valores medios anuales por trimestre son iguales para los tres años, esto es,

$$\bar{y}_{01\cdot} = \bar{y}_{02\cdot} = \bar{y}_{03\cdot} = 18,$$

lo cual significa que el paso del tiempo no hace que las observaciones varíen; por ello, b es igual a cero, resultando, entonces, la ecuación de tendencia:

$$\bar{y}_{i\cdot} = 18.$$

El modelo de tendencia obtenido —recta paralela al eje de abscisas a la altura de la ordenada en el origen— es un modelo estacionario o de media constante.

b) Las medias trimestrales son

$$\bar{y}_{\cdot 1} = 40$$

$$\bar{y}_{\cdot 2} = 10$$

$$\bar{y}_{\cdot 3} = 2$$

$$\bar{y}_{\cdot 4} = 20,$$

y, consecuentemente, para hallar la componente estacional de cada trimestre, ha de compararse, en esta ocasión, cada media trimestral, $\bar{y}_{\cdot j}$, con la media, $\bar{y} = 18$, según vimos en el problema **5.20**:

$$e_{\cdot j} = \frac{\bar{y}_{\cdot j}}{\bar{y}}.$$

En definitiva,

$$e_{\cdot 1} = \frac{40}{18} = 2{,}222$$

$$e_{\cdot 2} = \frac{10}{18} = 0{,}556$$

$$e_{\cdot 3} = \frac{2}{18} = 0{,}111$$

$$e_{\cdot 4} = \frac{20}{18} = 1{,}111.$$

Multiplicando por 100 los valores anteriores, se tiene el sistema de índices de variación estacional: 22,22, 55,6, 11,1 y 111,1 por ciento.

Concluimos, por tanto, que durante el segundo trimestre visitaron la estación de esquí un 44,4 por ciento menos de turistas de los que la habrían visitado si no hubiera existido estacionalidad.

5.23 Utilizando datos trimestrales del periodo 1996-2004, un grupo de expertos ha estimado la relación con respecto a la tendencia de la media anual de hectáreas arrasadas en los bosques de una cierta comarca. Dicha estimación ha permitido:

- Conocer que el incremento del número medio de hectáreas arrasadas debido al transcurso de un trimestre es 0,5.
- Prever que el total de hectáreas arrasadas durante el año 2007 será de 108.

Obténgase la correspondiente predicción para el año 2008.

SOLUCIÓN

Con los datos trimestrales del periodo 1996-2004, se ha estimado la ecuación de la recta de tendencia a partir de los valores medios anuales por trimestre, $\overline{y}_{i\cdot}$, siendo la pendiente de dicha recta igual al incremento que se produce en los valores medios anuales por el transcurso de un año y correspondiendo, por tanto, la cuarta parte de la pendiente al incremento del número medio anual por trimestre de hectáreas arrasadas debido al transcurso de un trimestre que, según la información que proporciona el enunciado es 0,5.

En consecuencia, la pendiente de la recta de tendencia resulta ser

$$0{,}5 \cdot 4 = 2.$$

Se sabe, además, que la previsión del número total de hectáreas arrasadas en el año 2007, $y^*_{07\cdot}$, es igual a 108, con lo cual, la predicción del valor medio por trimestre para dicho año es

$$\overline{y}^*_{07\cdot} = \frac{108}{4} = 27 \text{ hectáreas.}$$

Incrementando este valor medio en la cantidad correspondiente al transcurso de un año, es decir, en 2 hectáreas, resulta la predicción del número medio por trimestre de hectáreas arrasadas en el año 2008:

$$\overline{y}^*_{08\cdot} = 27 + 2 = 29 \text{ hectáreas,}$$

y, en definitiva, el total de hectáreas arrasadas previsto para dicho año es

$$y^*_{08\cdot} = 4 \cdot 29 = 116 \text{ hectáreas.}$$

5.24 La serie de ventas de productos textiles en una localidad costera, en miles de euros, en el periodo 2001-2004 ha sido:

Estación	2001	2002	2003	2004
Primavera	4	4	5	6
Verano	14	16	19	20
Otoño	4	5	5	6
Invierno	2	3	3	4

a) Bajo un esquema multiplicativo, y sabiendo que en el ajuste lineal mínimo-cuadrático de la media de ventas anuales por estación respecto a la tendencia el valor de la pendiente es uno, ¿qué ventas pueden esperarse para el verano de año 2005, atendiendo a este modelo?

b) Constrúyase una serie de números índices para las ventas anuales con base 2001.

c) Sabiendo que la serie de índices de precios de este tipo de mercancías para el periodo considerado ha sido: 100, 98, 103 y 104, calcúlese la tasa media de variación experimentada por las ventas anuales en términos reales con base en el año 2001.

SOLUCIÓN

a) Para calcular el índice de variación estacional del verano, aplicando el método de las relaciones de la media de cada estación respecto a la tendencia, ha de hallarse la media de ventas de cada estación:

$$\bar{y}_{\cdot p} = \frac{4 + 4 + 5 + 6}{4} = 4{,}75$$

$$\bar{y}_{\cdot v} = \frac{14 + 16 + 19 + 20}{4} = 17{,}25$$

$$\bar{y}_{\cdot o} = \frac{4 + 5 + 5 + 6}{4} = 5$$

$$\bar{y}_{\cdot i} = \frac{2 + 3 + 3 + 4}{4} = 3.$$

Puesto que el enunciado proporciona el valor de la pendiente en la regresión mínimo-cuadrática de la media de ventas anuales con respecto a la tendencia, pueden corregirse las anteriores medias:

$$\bar{y}'_{\cdot p} = 4{,}75$$

$$\bar{y}'_{\cdot v} = 17{,}25 - \frac{1}{4} \cdot 1 = 17$$

$$\bar{y}'_{\cdot o} = 5 - \frac{1}{4} \cdot 2 = 4{,}5$$

$$\bar{y}'_{\cdot i} = 3 - \frac{1}{4} \cdot 3 = 2{,}25,$$

siendo la media global corregida:

$$\overline{y}' = \frac{\overline{y}'_{\cdot p} + \overline{y}'_{\cdot v} + \overline{y}'_{\cdot o} + \overline{y}'_{\cdot i}}{4} = \frac{4{,}75 + 17 + 4{,}5 + 2{,}25}{4} = 7{,}125.$$

Por tanto, el índice de variación estacional del verano es

$$I_v = \frac{\overline{y}'_{\cdot v}}{\overline{y}'} \cdot 100 = \frac{17}{7{,}125} \cdot 100 = 238{,}59.$$

En consecuencia, en verano se vende un 138,59 por ciento más de lo que se vendería si no existiera estacionalidad.

La pendiente de la recta de regresión de la media anual de ventas respecto a la tendencia es igual a uno, lo cual implica que la cifra media de ventas aumenta en una unidad, es decir, en mil euros, al pasar del año 2004 al año 2005. Ahora bien, la media de ventas en el año 2004 fue, en miles de euros,

$$\overline{y}_{04\cdot} = \frac{6 + 20 + 6 + 4}{4} = 9,$$

esto es, 9 mil euros, por lo que en el año 2005 pasa a ser de 10 mil euros.

Invitamos al lector a que compruebe el valor de la pendiente de la ecuación de tendencia mediante estimación mínimo-cuadrática con los datos del enunciado, teniendo en cuenta que el cambio de variable adecuado en esta ocasión es

$$\overline{y}_{i\cdot} = a + 2 \cdot b(i - 2\,002{,}5).$$

En definitiva, para el verano de 2005 las ventas previstas se calculan como

$$y^*_{05,v} = \overline{y}^*_{05\cdot} \cdot e_{\cdot v},$$

por lo que éstas serán

$$y^*_{05,v} = 10 \cdot 2{,}3859 = 23{,}859 \text{ miles de euros.}$$

b) Se puede obtener de modo sencillo una serie de números índices simples para las cifras de ventas anuales con base en el año 2001, sin más que aplicar la expresión conocida del capítulo 4:

$$\frac{y_{i\cdot}}{y_{01\cdot}},$$

donde $y_{01\cdot}$ son las ventas correspondientes al año 2001 e $y_{i\cdot}$ las ventas de cada uno de los años del periodo 2001-2004 que se calculan sumando las observaciones de todas las estaciones para cada uno de ellos.

En la siguiente tabla se recogen los valores de las ventas, en miles de euros, así como los índices simples de los años considerados.

Años	2001	2002	2003	2004
Ventas anuales	24	28	32	36
Índices	1	1,167	1,333	1,5

c) La tasa pedida responde a la expresión:

$$\text{tm} = \sqrt[4-1]{\frac{y'_{04\cdot}}{y'_{01\cdot}}} - 1,$$

donde $y'_{01\cdot}$ son las ventas tanto en términos nominales como reales del año 2001, pues éste es el año base, e $y'_{04\cdot}$ las ventas del año 2004, en términos reales con base en el año 2001.

Considerando que, según vimos en el capítulo 4,

$$y'_{04\cdot} = \frac{y_{04\cdot}}{D_{01}^{04}},$$

donde D_{01}^{04} es el deflactor del año 2004 con base en el año 2001, resulta que

$$y'_{04\cdot} = \frac{36}{1{,}04} = 34{,}61.$$

En definitiva, sustituyendo las ventas en términos reales, la tasa media de variación de las ventas anuales en términos reales con base en el año 2001 es

$$\text{tm} = \sqrt[3]{\frac{34{,}61}{24}} - 1 = 0{,}1298.$$

5.25 Demuéstrese que a lo largo de una tendencia lineal la tasa de crecimiento es decreciente.

SOLUCIÓN

Considerando una tendencia lineal,

$$y_t = a + b \cdot t,$$

la tasa de variación entre $t - 1$ y t,

$$\dot{y}_t = \frac{y_t}{y_{t-1}} - 1,$$

responde a la expresión:

$$\dot{y}_t = \frac{a + b \cdot t}{a + b\,(t-1)} - 1 = \frac{a + b \cdot t - a - b\,(t-1)}{a + b\,(t-1)},$$

sin más que sustituir los valores de tendencia para dichos periodos.

Si $b > 0$, es decir, si la variable aumenta con el paso del tiempo, hecho que suponemos al hablar de tasa de *crecimiento*, entonces,

$$\dot{y}_t = \frac{b}{a + b\,(t-1)}$$

disminuye a medida que aumenta el valor de t. Se concluye, así, que la tasa de crecimiento a lo largo de una recta de tendencia *disminuye* a medida que transcurre el tiempo.

5.26 Demuéstrese que a lo largo de una tendencia exponencial la tasa de variación se mantiene constante.

SOLUCIÓN

Dada una tendencia exponencial,

$$y_t = a \cdot b^t,$$

se obtiene la tasa de variación entre los periodos $t - 1$ y t, sustituyendo los valores de tendencia para dichos periodos y simplificando:

$$\dot{y}_t = \frac{y_t}{y_{t-1}} - 1 = \frac{a \cdot b^t}{a \cdot b^{t-1}} - 1 = b - 1,$$

cantidad independiente de t.

Como consecuencia, puede comprobar el lector que la tasa media acumulativa, tm, es también igual a $b - 1$.

5.27 Se ha estimado la ecuación de tendencia de la serie de precios de un producto, en euros, durante los últimos 15 años: $y_t = 3{,}5 + 4{,}3 \cdot t$, para $t = 1, ..., 15$, siendo el coeficiente de determinación de la regresión mínimo-cuadrática igual a 0,98. Se sabe, además, que, a partir del próximo año, se renovará la maquinaría de fabricación del producto con la consiguiente reducción de los costes de producción.

a) De la ecuación anterior se ha obtenido una previsión del precio del producto para el próximo año de 72,3 euros. ¿Qué fiabilidad tiene esta predicción?

b) Estímese el incremento relativo interanual de los precios con la información disponible.

c) Obténgase una aproximación a la tasa media del precio correspondiente a los últimos 15 años.

SOLUCIÓN

a) Sustituyendo en la ecuación el valor $t = 16$ se obtiene, efectivamente, una previsión del precio del producto para el año próximo, y^*_{16}, igual a 72,3, que, si hacemos caso al valor del coeficiente de determinación lineal, 0,98, cercano a 1, diríamos que el ajuste de la ecuación de tendencia ha sido bueno y, consecuentemente, fiable la predicción.

Sin embargo, tenemos información adicional sobre una reducción de costes que conducirá a un descenso del precio del producto, lo cual producirá, a su vez, un cambio en la estructura de la tendencia de la serie que hace que surjan serias dudas sobre nuestra confianza en la predicción efectuada.

b) Al no disponer de los datos originales de la serie de tiempo, una buena aproximación al cálculo de las tasas de variación de los precios es considerar la tendencia de la serie y aplicar el resultado **5.25**.

Así, la tasa de variación entre los periodos $t - 1$ y t de los valores de la recta de tendencia es

$$\dot{y}_t = \frac{b}{a + b\,(t - 1)},$$

con lo cual, aplicando la expresión genérica anterior para $t = 2, ..., 15$, con $a = 3{,}5$ y $b = 4{,}3$, se tienen las tasas de variación de los valores de tendencia, y, por tanto, una estimación de las tasas de crecimiento de los precios, para el periodo considerado que figuran en la siguiente tabla.

Años	Tasas de variación
2	0,551
3	0,355
4	0,262
5	0,208
6	0,172
7	0,147
8	0,128
9	0,113

Años	Tasas de variación
10	0,102
11	0,092
12	0,085
13	0,078
14	0,072
15	0,068

c) Una aproximación a la tasa media de los precios, del periodo resulta de considerar las tasas de variación obtenidas en el apartado anterior y aplicar la definición de tasa media:

$$\text{tm} = \sqrt[15-1]{(1 + \dot{y}_2) \ldots (1 + \dot{y}_T)} - 1 = 0{,}167.$$

5.28 Aplicando el criterio de los mínimos cuadrados, se han estimado, a partir de datos anuales, los valores de tendencia que figuran en la tabla adjunta correspondientes al número de asistentes, en miles, a un congreso científico de celebración anual, suponiendo un modelo exponencial:

Años	Tendencia
1	0,840
2	1,008
3	1,209
4	1,451
5	1,741
6	2,090
7	2,508
8	3,009
9	3,611
10	4,334

Estímese la tasa media acumulativa del número de asistentes en el periodo considerado.

SOLUCIÓN

Como no se dispone de los valores de la magnitud para dicho periodo, una posible aproximación al cálculo de la tasa media de la misma es la tasa media de los valores de tendencia:

$$\text{tm} = \sqrt[10-1]{\frac{y^*_{10}}{y^*_1}} - 1,$$

donde y_{10}^* e y_1^* son las observaciones final e inicial, respectivamente, de la serie de tendencia y, consecuentemente, valores estimados.

Sustituyendo por los datos que proporciona el problema resulta la estimación de la tasa media acumulativa del número de asistentes al congreso para el periodo considerado:

$$\text{tm} = \sqrt[9]{\frac{4{,}334}{0{,}840}} - 1 = 0{,}2.$$

Por otro lado, si aplicamos el resultado **5.26**, sabemos que, a lo largo de una tendencia exponencial, la tasa media acumulativa es

$$\text{tm} = b - 1,$$

donde b es el parámetro de la ecuación de tendencia:

$$y_t = a \cdot b^t.$$

Puesto que tenemos los datos de la serie de tendencia, es posible hallar los parámetros del modelo porque, por ejemplo, los puntos (5;1,741) y (6;2,090) pertenecen a esta curva exponencial. Así, sustituyendo los pares de puntos anteriores, se tiene el siguiente sistema de ecuaciones con dos incógnitas:

$$1{,}741 = a \cdot b^5$$

$$2{,}090 = a \cdot b^6,$$

cuya resolución, que se lleva a cabo tomando logaritmos, conduce al valor $b = 1{,}2$, como puede comprobar fácilmente el lector.

En definitiva, este camino nos conduce también al valor:

$$\text{tm} = b - 1 = 1{,}2 - 1 = 0{,}2.$$

5.29 Se ha realizado un estudio sobre el comportamiento que en los últimos años han seguido los ingresos que una determinada ONG ha recibido por las donaciones de particulares.

La siguiente tabla recoge los ingresos medios anuales por trimestre, en miles de euros, para el periodo 2000-2004, así como la correspondiente serie de índices de precios.

Años	Medias anuales	Índices
2000	15	110
2001	25	115
2002	27	122
2003	30	128
2004	32	130

Además, se ha estimado la relación con respecto a la tendencia de la media anual de ingresos por trimestre, y ha resultado la siguiente ecuación:

$$\bar{y}_{i\cdot} = 25{,}8 + 3{,}9(i - 2002).$$

a) Hállese la serie de ingresos totales en términos reales con base en 2002.

b) Obténgase la tasa media de variación de los valores de tendencia para el periodo considerado.

SOLUCIÓN

a) Antes de calcular la serie de ingresos totales en términos reales, hay que obtener la serie de ingresos totales; puesto que el enunciado proporciona el valor medio por trimestre para cada año, $\bar{y}_{i\cdot}$, puede hallarse el ingreso total para cada uno de los años del periodo considerado:

$$y_{i\cdot} = 4 \cdot \bar{y}_{i\cdot}.$$

Así, la serie de ingresos totales anuales, en miles de euros, a precios corrientes, es decir, en términos nominales, es

$$y_{00\cdot} = 4 \cdot 15 = 60$$

$$y_{01\cdot} = 4 \cdot 25 = 100$$

$$y_{02\cdot} = 4 \cdot 27 = 108$$

$$y_{03\cdot} = 4 \cdot 30 = 120$$

$$y_{04\cdot} = 4 \cdot 32 = 128.$$

Para expresar la serie anterior en términos reales con base en 2002, es necesario contar con un deflactor. Puesto que se dispone de los valores del índice de precios para el periodo 2000-2004, puede utilizarse como deflactor el índice que mide la variación del índice de precios entre cada año considerado y el año 2002.

Por tanto, dividiendo el índice de precios de cada año entre el enlace, $I_0^{02} = 122$, se tiene que

$$D_{02}^{00} = \frac{110}{122} = 0{,}90$$

$$D_{02}^{01} = \frac{115}{122} = 0{,}94$$

$$D_{02}^{02} = \frac{122}{122} = 1$$

$$D_{02}^{03} = \frac{128}{122} = 1{,}05$$

$$D_{02}^{04} = \frac{130}{122} = 1{,}065.$$

En definitiva, aplicando la conocida expresión,

$$\text{precios constantes año } i \text{ (base 02)} = \frac{\text{precios corrientes año } i}{D_{02}^{i}},$$

se obtiene la siguiente serie de ingresos totales, en miles de euros, a precios constantes del año 2002:

$$y'_{00\cdot} = \frac{60}{0{,}9} = 66{,}6$$

$$y'_{01\cdot} = \frac{100}{0{,}94} = 106{,}38$$

$$y'_{02\cdot} = \frac{108}{1} = 108$$

$$y'_{03\cdot} = \frac{120}{1{,}05} = 114{,}28$$

$$y'_{04\cdot} = \frac{128}{1{,}065} = 120{,}18.$$

b) La tasa media de variación de los valores de tendencia para el periodo considerado es

$$\text{tm} = \sqrt[5-1]{\frac{\overline{y}^{*}_{04\cdot}}{\overline{y}^{*}_{00\cdot}}} - 1,$$

donde $\overline{y}^{*}_{04\cdot}$ e $\overline{y}^{*}_{00\cdot}$ son, respectivamente, los valores de tendencia *estimados* en los años 2000 y 2004 mediante la recta de regresión mínimo cuadrática:

$$\overline{y}_{i\cdot} = 25{,}8 + 3{,}9(i - 2002).$$

Sustituyendo en la ecuación anterior, se tiene la estimación de la tendencia, para el año 2000,

$$\overline{y}^{*}_{00\cdot} = 25{,}8 + 3{,}9(2000\text{-}2002) = 18 \text{ mil euros},$$

siendo la estimación para 2004 igual a

$$\overline{y}^{*}_{04\cdot} = 25{,}8 + 3{,}9(2004\text{-}2002) = 33{,}6 \text{ miles de euros}.$$

En definitiva, la tasa media de variación de los valores de tendencia para el periodo considerado es

$$\text{tm} = \sqrt[4]{\frac{\overline{y}^*_{04\cdot}}{\overline{y}^*_{00\cdot}}} - 1 = \sqrt[4]{\frac{33{,}6}{18}} - 1 = 0{,}1688.$$

Capítulo 6

Introducción al cálculo de probabilidades

Principales conceptos y resultados

Un experimento es **aleatorio** cuando al repetirse en las mismas condiciones no da lugar al mismo resultado.

Se denomina **espacio muestral**, Ω, al conjunto de los resultados posibles de un experimento aleatorio. Cada resultado, ω, es un **punto muestral**.

Un **suceso**, A, es un subconjunto del espacio muestral formado, por tanto, por puntos muestrales. Un suceso **elemental** consta de un único punto muestral, mientras que un suceso **compuesto** está formado por más de un punto muestral.

Dado que los sucesos son, en realidad, conjuntos, las operaciones (complementariedad, unión, intersección, diferencia y diferencia simétrica) y relaciones (inclusión, igualdad e incompatibilidad) entre conjuntos son igualmente válidas para sucesos. Así, las diferentes operaciones entre sucesos conducen a las siguientes definiciones:

- Suceso **complementario** de un suceso A:

$$\bar{A} = \{\omega \in \Omega \,/\, \omega \notin A\}.$$

- Suceso **unión** de los sucesos A y B:

$$A \cup B = \{\omega \in \Omega \,/\, \omega \in A, \text{ o bien, } \omega \in B\}.$$

- Suceso **intersección** de los sucesos A y B:

$$A \cap B = \{\omega \in \Omega \,/\, \omega \in A \text{ y } \omega \in B\}.$$

- Suceso **diferencia** de los sucesos A y B:

$$A - B = \{\omega \in \Omega \,/\, \omega \in A \text{ y } \omega \notin B\}.$$

- Suceso **diferencia simétrica** de los sucesos A y B:

$$A \,\Delta\, B = \{\omega \in \Omega \,/\, (\omega \in A \text{ o bien } \omega \in B) \text{ y } \omega \notin A \cap B\}.$$

Análogamente, se tienen las siguientes relaciones entre sucesos:

- El suceso A está **contenido** en el suceso B, si cualquier punto muestral que pertenece al suceso A también pertenece al suceso B:

$$A \subset B \text{ si } \omega \in A \Rightarrow \omega \in B.$$

- Los sucesos A y B son **iguales**, si cualquier punto muestral de A está en B y viceversa:

$$A = B \text{ si } \omega \in A \Leftrightarrow \omega \in B.$$

- Los sucesos A y B son **incompatibles**, **disjuntos** o **mutuamente excluyentes**, si no tienen puntos muestrales en común:

$$A \cap B = \phi.$$

La unión y la intersección de suceso cumplen las propiedades *asociativa* y *conmutativa* y, entre ellas, se verifica la propiedad *distributiva*.

Con las dos leyes de Morgan se relacionan la unión, la intersección y la complementariedad:

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Se dice que *ha ocurrido un suceso A*, si al realizar el experimento aleatorio se obtiene cualquier punto muestral perteneciente a A[1].

El conjunto de todos los sucesos[2] se llama **partes de** Ω, $\wp(\Omega)$, y el par $(\Omega, \wp(\Omega))$ se llama **espacio probabilizable**[3].

Una **probabilidad** es una aplicación, p, de $\wp(\Omega)$ en la recta real, $\Re$, tal que a cada suceso, A, le hace corresponder su *medida teórica de ocurrencia*[4], $p(A)$.

---

[1] En este sentido, Ω es el suceso *seguro* y ϕ el suceso *imposible*.

[2] En muchas ocasiones no interesan todos los sucesos del espacio muestral sino únicamente una parte de ellos, verificando una serie de propiedades de interés, que se denomina σ-álgebra. Aunque la introducción de este concepto no es necesaria en el contexto que nos ocupa, recomendamos al lector interesado la consulta de textos donde el estudio del cálculo de probabilidades se hace de un modo más pormenorizado.

[3] En realidad, el espacio probabilizable aquí definido es un caso particular de la situación general en la que, en lugar de $\wp(\Omega)$, la mayor de las σ-álgebras, se considera una σ-álgebra cualquiera.

[4] La definición de probabilidad con la que trabajaremos en este capítulo es, en realidad, un caso particular de la definición de probabilidad sobre los sucesos de una σ-álgebra cualquiera.

Una probabilidad cumple los tres axiomas siguientes[5]:

- Para cualquier suceso A de $\wp(\Omega)$, $p(A) \geq 0$.
- $p(\Omega) = 1$.
- Dada cualquier colección infinita numerable de sucesos $\{A_i\}_{i=1}^{\infty}$, disjuntos dos a dos, se cumple: $p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$.

De la definición axiomática de probabilidad se derivan las siguientes propiedades:

P.1 $p(\phi) = 0$.

P.2 Dada cualquier colección finita de sucesos $\{A_i\}_{i=1}^{\infty}$, disjuntos dos a dos, se cumple: $p\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} p(A_i)$.

P.3 Para cualquier suceso A, $p(\bar{A}) = 1 - p(A)$.

P.4 Dados dos sucesos A y B tales que $A \subset B$, $p(B - A) = p(B) - p(A)$.

P.5 Dados dos sucesos A y B tales que $A \subset B$, entonces $p(A) \leq p(B)$.

P.6 Para cualquier suceso A, $p(A) \leq 1$.

P.7 Dados dos sucesos A y B, entonces, $p(B - A) = p(B) - p(A \cap B)$.

P.8 Dados dos sucesos[6] A y B, entonces, $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

La terna $(\Omega, \wp(\Omega), p)$ recibe el nombre de **espacio probabilístico**[7].

Dado un espacio probabilístico $(\Omega, \wp(\Omega), p)$, y dado un suceso B de probabilidad distinta de cero, se llama **probabilidad condicionada** por B a una aplicación de $\wp(\Omega)$ en la recta real, $\Re$, tal que a cada suceso A le hace corresponder el número real:

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

Entendida la probabilidad del suceso A como la medida de su ocurrencia, la probabilidad[8] condicionada es la medida de la ocurrencia de A, sabiendo que ha ocurrido el suceso B, esto es, la medida de la ocurrencia del suceso A *dentro* del suceso B.

---

[5] La definición axiomática de probabilidad se debe al matemático ruso Kolmogorov.

[6] Esta propiedad puede generalizarse para un número n de sucesos.

[7] Cualquier asignación de probabilidad a los sucesos elementales en un espacio muestral finito o infinito numerable de modo que la probabilidad de cada suceso elemental sea no negativa y la suma de todas ellas sea la unidad, define una probabilidad. En particular, la definición de una probabilidad sobre un espacio muestral finito y equiprobable conduce a la conocida regla de Laplace, para cuya aplicación se requiere la obtención del *número de casos posibles* y del *número de casos favorables* al suceso del cual se desea calcular su probabilidad. Para ello se utiliza el análisis combinatorio, como se podrá comprobar en algunos de los problemas de este capítulo.

[8] La probabilidad condicionada cumple los tres axiomas de Kolmogorov, siendo, por tanto, una probabilidad.

La **regla de la multiplicación** es una consecuencia de la definición de probabilidad condicionada. Así, dados n sucesos de un espacio probabilístico, tales que $p(A_1 \cap \ldots \cap A_{n-1}) \neq 0$, entonces,

$$p(A_1 \cap \ldots \cap A_n) = p(A_1) \cdot p(A_2/A_1) \ldots p(A_n/A_1 \cap \ldots \cap A_{n-1}).$$

Otras dos importantes consecuencias de la definición de probabilidad condicionada son el **teorema de la probabilidad total** y el **teorema de Bayes**. Así, dado un espacio probabilístico y una partición[9] del espacio muestral formada por una colección infinita numerable de sucesos $\{A_i\}_{i=1}^{\infty}$ se cumplen los dos resultados siguientes:

1. La probabilidad de un suceso cualquiera B es

$$p(B) = \sum_{i=1}^{\infty} p(A_i) \cdot p(B/A_i).$$

2. Si B es un suceso de probabilidad no nula, entonces, la probabilidad de cualquier suceso de la partición, A_j, condicionada por B puede escribirse como

$$p(A_j/B) = \frac{p(A_j) \cdot p(B/A_j)}{\sum_{i=1}^{\infty} p(A_i) \cdot p(B/A_i)}.$$

Se dice que dos sucesos A y B de un espacio probabilístico son **independientes**, si

$$p(A \cap B) = p(A) \cdot p(B).$$

De la definición de independencia de sucesos se deducen varias consecuencias: un suceso de probabilidad nula es independiente de cualquier otro; un suceso de probabilidad igual a la unidad es independiente de cualquier otro; y, si un suceso es independiente de otro, también lo es de su complementario.

Dados dos sucesos A y B de un espacio probabilístico, A y B son independientes si uno de ellos es de probabilidad nula, o $p(A/B) = p(A)$.

Se dice que tres sucesos de un espacio probabilístico, A, B y C, son **mutuamente independientes** si

$$p(A \cap B) = p(A) \cdot p(B)$$

$$p(A \cap C) = p(A) \cdot p(C)$$

$$p(B \cap C) = p(B) \cdot p(C)$$

$$p(A \cap B \cap C) = p(A) \cdot p(B) \cdot p(C).$$

En ocasiones, un experimento aleatorio se puede descomponer en varias etapas[10], con lo cual, el cálculo de las probabilidades de los sucesos requiere el conocimiento de las relaciones de *independencia* —el resultado de cada una de ellas no influye en los resultados del resto— o de *dependencia* —existe influencia entre los resultados— de dichas etapas.

---

[9] Una partición del espacio muestral Ω está formada por una colección infinita numerable de sucesos $\{A_i\}_{i=1}^{\infty}$, disjuntos dos a dos y tales que su unión es igual Ω.

[10] Se trata de los llamados *experimentos compuestos*, cuya compleja formalización omitiremos en esta obra, aunque sí calcularemos probabilidades de sucesos pertenecientes a este tipo de experimentos.

APLICACIÓN DE CONCEPTOS Y DEMOSTRACIÓN DE RESULTADOS

6.1 Dado un espacio probabilístico $(\Omega, \wp(\Omega), p)$, demuéstrense las propiedades que se derivan de la definición de probabilidad.

SOLUCIÓN

Antes de comenzar con la demostración de estas propiedades hemos de mencionar que, para poder llevarla a cabo, solamente pueden utilizarse los tres axiomas de la definición de probabilidad y cada una de las propiedades que, sucesivamente, se vayan comprobando; es necesario hacer este comentario porque el hecho de que el lector "conozca" algunas propiedades derivadas del concepto de probabilidad hace que caiga en el frecuente error de aplicarlas para demostrar otras, aunque aquellas no estén todavía probadas. En este sentido, recomendamos al lector que, aunque hay otras posibilidades igualmente válidas, intente resolver este problema *siguiendo el orden* en el que son presentadas en el resumen teórico que aparece al principio de este capítulo.

P.1 $\mathrm{p}(\phi) = 0$.

Para demostrar la primera propiedad tomaremos una colección infinita numerable de sucesos, $\{A_i\}_{i=1}^{\infty}$, tal que $A_i = \phi$, para todo suceso de la colección, es decir, una colección infinita numerable en la que todos los sucesos son el suceso imposible. Se cumple, entonces, que la unión de todos ellos es el suceso imposible,

$$\bigcup_{i=1}^{\infty} A_i = \phi,$$

con lo cual,

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = p(\phi).$$

Aplicando el tercer axioma de la probabilidad, se verifica:

$$p(\phi) = p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i) = \sum_{i=1}^{\infty} p(\phi),$$

concluyéndose así que, necesariamente,

$$p(\phi) = 0.$$

P.2 Dada cualquier colección finita de sucesos del espacio probabilístico, $\{A_i\}_{i=1}^{n}$, disjuntos dos a dos, se verifica que $p\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} p(A_i)$.

Para completar la colección *finita* de sucesos disjuntos y convertirla en una colección *infinita* de sucesos, $\{A_i\}_{i=1}^{\infty}$, también disjuntos, añadiremos sucesos iguales al suceso imposible, es decir, $A_{n+1} = \phi, A_{n+2} = \phi, \ldots$, cumpliéndose, entonces, que

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{n} A_i,$$

y, por tanto,

$$p\left(\bigcup_{i=1}^{n} A_i\right) = p\left(\bigcup_{i=1}^{\infty} A_i\right).$$

Teniendo en cuenta el tercer axioma de la probabilidad, se tiene que

$$p\left(\bigcup_{i=1}^{n} A_i\right) = p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i),$$

con lo cual, dividiendo el sumatorio en dos sumatorios y teniendo en cuenta, por la propiedad **P.1** ya demostrada, que el segundo sumatorio es una suma de sumandos todos iguales a cero, resulta:

$$\sum_{i=1}^{\infty} p(A_i) = \sum_{i=1}^{n} p(A_i) + \sum_{i=n+1}^{\infty} p(A_i) = \sum_{i=1}^{n} p(A_i) + \sum_{i=n+1}^{\infty} p(\phi) = \sum_{i=1}^{n} p(A_i),$$

quedando probada esta propiedad.

P.3 Para cualquier suceso A, $p(\bar{A}) = 1 - p(A)$.

Si realizamos una partición del espacio muestral, considerando el suceso A como uno de los suceso implicados,

$$A \cup \bar{A} = \Omega,$$

entonces,

$$p(A \cup \bar{A}) = p(\Omega) = 1.$$

Aplicando la propiedad **P.2** para el caso de dos sucesos, se cumple que

$$1 = p(A \cup \bar{A}) = p(A) + p(\bar{A}),$$

esto es,

$$p(\bar{A}) = 1 - p(A).$$

P.4 Dados dos sucesos A y B tales que $A \subset B$, entonces, $p(B - A) = p(B) - p(A)$.

Si el suceso A está contenido en B, el suceso B puede expresarse como unión de los sucesos disjuntos A y $B - A$:

$$B = A \cup (B - A).$$

En efecto, el suceso B puede escribirse como B intersección el suceso seguro, esto es,

$$B = B \cap \Omega.$$

Por otro lado, el suceso A y su complementario, $\bar{A}$, constituyen una partición del espacio muestral, con lo cual, el suceso B admite expresarse como

$$B = B \cap (A \cup \bar{A}).$$

Por último, aplicando la propiedad distributiva de la intersección con respecto a la unión, se tiene que

$$B = (B \cap A) \cup (B \cap \bar{A}).$$

o, lo que es igual,

$$B = (A \cap B) \cup (B - A),$$

siendo esta unión disjunta, pues $A \cap B \subset A$ y $B - A \subset \bar{A}$.

Ahora bien, como $A \subset B$, entonces, $B \cap A = A$, con lo cual,

$$B = A \cup (B - A).$$

Esta igualdad permite aplicar la propiedad **P.2**, por lo que

$$p(B) = p(A) + p(B - A),$$

y, despejando,

$$p(B - A) = p(B) - p(A).$$

P.5 Dados dos sucesos A y B tales que $A \subset B$, entonces, $p(A) \leq p(B)$.

Esta propiedad es inmediata, teniendo en cuenta que, por un lado, acabamos de probar que $p(B) = p(A) + p(B - A)$ y que, por otro lado, el primer axioma de la probabilidad garantiza que la probabilidad de cualquier suceso es positiva, con lo cual, $p(B - A) \geq 0$ y, necesariamente, la probabilidad de A ha de ser, a lo sumo, igual que la probabilidad de B.

P.6 $p(A) \leq 1$.

Como $A \subset \Omega$, de la propiedad anterior se deriva de modo inmediato que

$$p(A) \leq p(\Omega).$$

Pero como

$$p(\Omega) = 1,$$

entonces,

$$p(A) \leq 1.$$

P.7 Dados dos sucesos A y B, se cumple: $p(B - A) = p(B) - p(A \cap B)$.

Para demostrar esta propiedad, tendremos en cuenta que, según vimos en **P.4**, el suceso B puede escribirse como unión de dos sucesos disjuntos,

$$B = (B \cap A) \cup (B \cap \bar{A}).$$

Entonces, considerando de nuevo la propiedad **P.2**,

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

con lo cual, despejando,

$$p(B - A) = p(B) - p(A \cap B),$$

quedando así demostrada esta propiedad.

Obsérvese que estamos ante una generalización de la propiedad **P.4**, ya que, si $A \subset B$, entonces, $p(A \cap B) = p(A)$.

P.8 Dados dos sucesos A y B, se verifica: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

El suceso $A \cup B$ puede expresarse como unión de tres sucesos disjuntos,

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A),$$

ya que, según vimos en **P.4**, el suceso A puede escribirse como

$$A = (A \cap B) \cup (A \cap \bar{B}),$$

y el suceso B como

$$B = (B \cap A) \cup (B \cap \bar{A}),$$

por lo que,

$$A \cup B = (A \cap B) \cup (A \cap \bar{B}) \cup (A \cap B) \cup (\bar{A} \cap B)$$

o, lo que es igual,

$$A \cup B = (A - B) \cup (A \cap B) \cup (B - A).$$

Aplicando la propiedad **P.2**, ahora para el caso de tres sucesos, se tiene que

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

Ahora bien, por la propiedad **P.7** se obtiene, por un lado,

$$p(A - B) = p(A) - p(A \cap B),$$

y, por otro lado,

$$p(B - A) = p(B) - p(A \cap B),$$

con lo que, sustituyendo,

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

y, en consecuencia,

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

Observe el lector que, en la situación particular de que A y B fueran sucesos disjuntos, su intersección tendría probabilidad cero, con lo cual, esta propiedad es más general que **P.2**.

6.2 Analícese la probabilidad de un espacio probabilístico $(\Omega, \wp(\Omega), p)$, donde Ω es un espacio muestral finito y equiprobable.

SOLUCIÓN

El hecho de que el espacio muestral finito, $\Omega = \{\omega_1, \ldots, \omega_n\}$, sea equiprobable significa que todos los sucesos elementales tienen idéntica probabilidad, esto es,

$$p(\{\omega_1\}) = \ldots = p(\{\omega_n\}).$$

y, dado que, por aplicación de **P.2**:

$$p(\Omega) = p(\{\omega_1\} \cup \ldots \cup \{\omega_n\}) = p(\{\omega_1\}) + \ldots + p(\{\omega_n\}) = 1.$$

entonces, para $i = 1, \ldots, n$, necesariamente,

$$p(\{\omega_i\}) = \frac{1}{n}.$$

Además, como cualquier suceso A puede expresarse como unión finita de sucesos elementales,

$$A = \bigcup_{\omega_i \in A} \{\omega_i\},$$

entonces, volviendo a aplicar la propiedad **P.2**, resulta que la probabilidad del suceso A es igual a

$$p(A) = p\left(\bigcup_{\omega_i \in A} \{\omega_i\}\right) = \sum_{\omega_i \in A} p(\{\omega_i\}) = \frac{k}{n},$$

donde k es el número de puntos muestrales que pertenecen al suceso A.

Téngase en cuenta que el denominador de la probabilidad obtenida, esto es, n, resulta ser igual al número de resultados del experimento aleatorio, o equivalentemente, al *número de casos posibles*; en cuanto al numerador, k, es el número de resultados que son favorables a la ocurrencia del suceso A, es decir, *el número de casos favorables al suceso A*. En consecuencia, la probabilidad del suceso A se obtiene aplicando la conocida regla de Laplace: cociente entre el número de casos favorables y el número de casos posibles.

6.3 Sean A y B dos sucesos tales que $p(A) = 0{,}4$, $p(B) = 0{,}3$ y $p(A \cup B) = 0{,}6$. Calcúlese $p(A \cap B)$, $p(\overline{A})$, $p(\overline{B})$, $p(A - B)$, $p(B - A)$, $p(\overline{A} \cup \overline{B})$ y $p(\overline{A} \cap \overline{B})$.

SOLUCIÓN

La propiedad **P.8**,

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

permite, despejando, calcular

$$p(A \cap B) = p(A) + p(B) - p(A \cup B) = 0{,}4 + 0{,}3 - 0{,}6 = 0{,}1.$$

Las probabilidades de $\bar{A}$ y $\bar{B}$ se obtienen por aplicación de **P.3**:

$$p(\bar{A}) = 1 - p(A) = 1 - 0{,}4 = 0{,}6$$

y

$$p(\bar{B}) = 1 - p(B) = 1 - 0{,}3 = 0{,}7.$$

Mediante **P.7** se obtiene, de modo inmediato, que

$$p(A - B) = p(A) - p(A \cap B) = 0{,}4 - 0{,}1 = 0{,}3$$

y

$$p(B - A) = p(B) - p(A \cap B) = 0{,}3 - 0{,}1 = 0{,}2.$$

Por último, con las leyes de Morgan se calculan:

$$p(\bar{A} \cup \bar{B}) = p(\overline{A \cap B}) = 1 - p(A \cap B) = 1 - 0{,}1 = 0{,}9$$

y

$$p(\bar{A} \cap \bar{B}) = p(\overline{A \cup B}) = 1 - p(A \cup B) = 1 - 0{,}6 = 0{,}4.$$

6.4 Se consideran dos sucesos, A y B, de los cuales se conocen $p(A) = 0{,}7$, $p(A - B) = 0{,}3$ y $p(B - A) = 0{,}2$. Hállense las probabilidades de los sucesos: $A \cup B$, B, $\bar{A}$, $\bar{B}$, $A \cap B$, $\bar{A} \cap \bar{B}$, $\bar{A} \cup \bar{B}$.

SOLUCIÓN

El suceso $C = A \cup B$ puede escribirse, según vimos en el problema **6.1**, como

$$(C \cap A) \cup (C \cap \bar{A}).$$

Ahora bien, por un lado, el primer suceso de la unión anterior es

$$C \cap A = (A \cup B) \cap A = A,$$

ya que $A \subset A \cup B$.

Por otro lado, el segundo suceso es

$$C \cap \bar{A} = (A \cup B) \cap \bar{A} = (A \cap \bar{A}) \cup (B \cap \bar{A}) = \phi \cup (B \cap \bar{A}) = B \cap \bar{A}.$$

En definitiva,

$$A \cup B = A \cup (B - A),$$

unión disjunta, pues $B - A \subset \bar{A}$, con lo cual, puede aplicarse **P.2**:

$$p(A \cup B) = p(A) + p(B - A) = 0{,}7 + 0{,}2 = 0{,}9.$$

Por otro lado, teniendo en cuenta **P.7**,

$$p(A - B) = p(A) - p(A \cap B),$$

resulta, despejando, que

$$p(A \cap B) = p(A) - p(A - B) = 0{,}7 - 0{,}3 = 0{,}4^{1}.$$

De **P.8** se obtiene:

$$p(B) = p(A \cup B) - p(A) + p(A \cap B) = 0{,}9 - 0{,}7 + 0{,}4 = 0{,}6,$$

resultado al que también puede llegarse teniendo en cuenta que $p(B - A) = p(B) - p(A \cap B)$.

Las probabilidades de los sucesos $\bar{A}$ y $\bar{B}$ son inmediatas, considerando **P.3**:

$$p(\bar{A}) = 1 - p(A) = 1 - 0{,}7 = 0{,}3$$

y

$$p(\bar{B}) = 1 - p(B) = 1 - 0{,}6 = 0{,}4.$$

Finalmente, utilizando las leyes de Morgan:

$$p(\bar{A} \cup \bar{B}) = p(\overline{A \cap B}) = 1 - p(A \cap B) = 1 - 0{,}4 = 0{,}6$$

y

$$p(\bar{A} \cap \bar{B}) = p(\overline{A \cup B}) = 1 - p(A \cup B) = 1 - 0{,}9 = 0{,}1.$$

Téngase en cuenta que, aplicando **P.8** a los sucesos $\bar{A}$ y $\bar{B}$, se cumple que $p(\bar{A} \cup \bar{B}) = p(\bar{A}) + p(\bar{B}) - p(\bar{A} \cap \bar{B})$.

6.5 Se ha realizado un estudio sobre los hábitos en el desayuno de una población. Entre otros, se han obtenido los siguientes resultados:

- El 53 por ciento bebe una taza de leche.
- El 33 por ciento desayuna con cereales.
- El 65 por ciento alguna de las dos cosas.

[1] Proponemos al lector el cálculo de $p(A \cup B)$ hallando previamente $p(A \cap B)$ y utilizando parte de la demostración de la propiedad **P.8** que figura en **6.1**.

Calcúlese el porcentaje de individuos de la población que:

a) No desayuna ni con leche ni con cereales.

b) Desayuna con una taza de leche sin cereales.

SOLUCIÓN

Se considera el experimento aleatorio consistente en elegir un individuo de la población y ver qué desayuna (taza de leche y cereales). El espacio muestral asociado a este experimento es

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\},$$

donde

ω_1 = el individuo sólo bebe una taza de leche.

ω_2 = el individuo sólo desayuna con cereales.

ω_3 = el individuo desayuna con leche y cereales.

ω_4 = el individuo ni toma leche ni toma cereales.

El enunciado del problema proporciona las probabilidades de los sucesos, L y C, *beber una taza de leche* y *desayunar con cereales*, respectivamente, y de $L \cup C$, esto es, *desayunar alguna de las dos cosas*:

$$p(L) = p(\{\omega_1, \omega_3\}) = 0{,}53,$$

$$p(C) = p(\{\omega_2, \omega_3\}) = 0{,}33$$

y

$$p(L \cup C) = p(\{\omega_1, \omega_2, \omega_3\}) = 0{,}65.$$

a) El suceso descrito en este apartado se corresponde con $\overline{L} \cap \overline{C}$ o, lo que es igual, con el suceso complementario de la unión, $\overline{L \cup C}$. Por tanto,

$$p(\overline{L \cup C}) = 1 - p(L \cup C) = 1 - 0{,}65 = 0{,}35,$$

esto es, el 35 por ciento de los individuos no desayuna ni leche ni cereales.

Observe el lector que el suceso $\overline{L \cup C}$ es, en realidad, el suceso elemental $\{\omega_4\}$.

b) La probabilidad del suceso $L \cap \bar{C} = L - C$ se calcula como

$$p(L - C) = p(L) - p(L \cap C),$$

para lo cual es necesario conocer $p(L \cap C)$.

Ahora bien, de la relación

$$p(L \cup C) = p(L) + p(C) - p(L \cap C),$$

se obtiene, despejando, que

$$p(L \cap C) = p(L) + p(C) - p(L \cup C) = 0{,}53 + 0{,}33 - 0{,}65 = 0{,}21,$$

probabilidad del suceso elemental $\{\omega_3\}$.

En definitiva, la probabilidad pedida en este apartado, probabilidad de $\{\omega_1\}$, resulta:

$$p(L - C) = 0{,}53 - 0{,}21 = 0{,}32,$$

es decir, el 32 por ciento de los individuos de la muestra desayuna con una taza de leche sin cereales.

6.6 Se considera el espacio muestral $\Omega = \{0, 1, \ldots\}$ y una probabilidad, p, de $\wp(\Omega)$ en $\Re$, tal que

$$p(i) = 0{,}2^i \cdot 0{,}8^k.$$

Hállese la probabilidad de obtener el resultado 2.

SOLUCIÓN

Para el cálculo de la constante k ha de considerarse que, como la unión de todos los sucesos elementales es igual a Ω, entonces, por aplicación del tercer axioma de la probabilidad, la suma de las probabilidades asignadas a cada uno de los sucesos elementales tiene que ser igual a la unidad:

$$1 = \sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{\infty} 0{,}2^i \cdot 0{,}8^k = 0{,}8^k \sum_{i=0}^{\infty} 0{,}2^i = 0{,}8^k \cdot \frac{1}{1 - 0{,}2} = 0{,}8^k \cdot \frac{1}{0{,}8} = 0{,}8^{k-1},$$

resultado al que hemos llegado, teniendo en cuenta que $\sum_{i=0}^{\infty} 0{,}2^i$ es la suma de los infinitos términos de una progresión geométrica de razón menor que la unidad[2].

[2] La suma de los infinitos términos de una progresión geométrica de razón, r, tal que, $|r| < 1$, es $\frac{a_1}{1 - r}$, donde a_1 es el primer término de la progresión.

Por tanto,

$$0{,}8^{k-1} = 1,$$

con lo cual, necesariamente, ha de cumplirse que

$$k - 1 = 0,$$

y, en consecuencia, la constante k es igual a la unidad.

Una vez calculado el valor de k, la probabilidad pedida se halla de modo inmediato:

$$p(2) = 0{,}2^2 \cdot 0{,}8 = 0{,}032.$$

6.7 Una pandilla formada por 6 amigos, de los cuales dos son gemelos, ha asistido a un partido de baloncesto y se han sentado todos en la misma fila de modo aleatorio. ¿Cuál es la probabilidad de que los dos gemelos se sienten juntos?

SOLUCIÓN

El número de posibles reordenaciones de los 6 amigos en la fila es igual a las permutaciones de 6 elementos[3], es decir,

$$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 720,$$

siendo todas ellas igualmente probables, hecho que permite aplicar la regla de Laplace a la hora de hallar la probabilidad pedida.

Para calcular el número de casos favorables ha de considerarse a los dos gemelos como un único elemento. De este modo, serán situaciones favorables el número de reordenaciones de 5 elementos multiplicando el resultado por dos, ya que cada una de estas ordenaciones admite, a su vez, que los dos gemelos permuten entre sí:

$$2 \cdot 5! = 2 \cdot 5 \cdot 4 \cdot 3 \cdot 2 = 240.$$

En definitiva,

$$p(\text{gemelos se sienten juntos}) = \frac{240}{720} = 0{,}33.$$

[3] Recuérdese que, dado un conjunto de m elementos $\{a_1, \ldots, a_m\}$, el número de *permutaciones* de estos m elementos es el número de todas sus posibles ordenaciones, esto es, $P_m = m!$.

6.8 El conserje de una facultad está de baja por enfermedad. Su sustituto ha olvidado a qué departamento pertenece cada uno de los cuatro casilleros de correspondencia asignados a su secretaría, con lo cual decide distribuir el correo de cada departamento al azar.

a) ¿Qué probabilidad hay de que realice la asignación correctamente?

b) Hállese la probabilidad de que la secretaría del departamento de economía tenga bien repartido su correo.

SOLUCIÓN

a) El número de todas las asignaciones que el sustituto puede realizar del correo de cada departamento es igual al número de posibles reordenaciones del correo en cada uno de los 4 casilleros, esto es, permutaciones de 4 elementos:

$$4! = 24.$$

Puesto que todas estas reordenaciones son igualmente probables, es posible aplicar la regla de Laplace para hallar la probabilidad requerida.

En cuanto al número de casos favorables, observe el lector que únicamente hay una situación en la cual todo el correo está perfectamente asignado. En definitiva:

$$p(\text{asignación correcta del correo}) = \frac{1}{24}.$$

b) Para determinar los casos favorables en esta ocasión, hay que tener en cuenta que, una vez asignado de manera correcta el correo de la secretaría del departamento de economía, los otros tres departamentos pueden tener su correo reordenado en cualquiera de los tres casilleros restantes. Por tanto, el número de casos favorables es el de permutaciones de 3 elementos:

$$3! = 6.$$

En consecuencia,

$$p(\text{economía tiene asignación correcta}) = \frac{6}{24} = 0{,}25.$$

6.9 Sea $(\Omega, \wp(\Omega), p)$ un espacio probabilístico y sea B un suceso cualquiera de probabilidad no nula. Demuéstrese que la aplicación de $\wp(\Omega)$ en la recta real, tal que a cada suceso le hace corresponder su probabilidad condicionada por B, es una probabilidad sobre el espacio probabilizable $(\Omega, \wp(\Omega))$.

SOLUCIÓN

Para demostrar este resultado hay que comprobar que dicha aplicación cumple los tres axiomas de Kolmogorov.

En primer lugar, resulta inmediato que, para cualquier suceso A, se cumple que

$$p(A/B) = \frac{p(A \cap B)}{p(B)} \geq 0,$$

ya que, tanto numerador como denominador, son cantidades no negativas.

En segundo lugar, aplicando las propiedades de la probabilidad,

$$p(\Omega/B) = \frac{p(\Omega \cap B)}{p(B)} = \frac{p(B)}{p(B)} = 1.$$

Y, por último, dada una colección infinita numerable de sucesos, disjuntos dos a dos, $\{A_i\}_{i=1}^{\infty}$, se verifica, por definición de probabilidad condicionada, que

$$p\left(\bigcup_{i=1}^{\infty} A_i/B\right) = \frac{p\left[\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right]}{p(B)} = \frac{p\left[\bigcup_{i=1}^{\infty} (A_i \cap B)\right]}{p(B)},$$

siendo la última igualdad resultado de aplicar la propiedad distributiva de la intersección con respecto a la unión.

Utilizando el tercer axioma de la probabilidad, cosa que puede hacerse puesto que los sucesos de la colección $\{A_i \cap B\}_{i=1}^{\infty}$ son disjuntos dos a dos, resulta que

$$p\left(\bigcup_{i=1}^{\infty} A_i/B\right) = \frac{\sum_{i=1}^{\infty} p(A_i \cap B)}{p(B)} = \sum_{i=1}^{\infty} \frac{p(A_i \cap B)}{p(B)} = \sum_{i=1}^{\infty} p(A_i/B),$$

quedando así, demostrados los tres axiomas de Kolmogorov para la probabilidad condicionada.

Obsérvese que, si $B = \Omega$, entonces,

$$p(A/\Omega) = \frac{p(A \cap \Omega)}{p(\Omega)} = \frac{p(A)}{1} = p(A).$$

6.10 Un estudio estadístico tiene por objeto evaluar los resultados de una campaña publicitaria destinada al lanzamiento de un nuevo producto. Una de las conclusiones de este análisis es que el 60 por ciento de los individuos que ha visto el anuncio ha comprado posteriormente el producto.

Si un individuo adquiere el producto, ¿puede decirse que hay una probabilidad igual a 0,4 de que no haya visto el anuncio?

SOLUCIÓN

Llamando A y P a los sucesos *ver el anuncio* y *comprar el producto*, respectivamente, se conoce la probabilidad de P condicionada por A, esto es,

$$p(P/A) = 0{,}6.$$

Y, únicamente con esta información, no puede deducirse $p(\overline{A}/P)$, probabilidad del suceso $\overline{A}$ condicionada por la ocurrencia del suceso P.

El lector que haya contestado afirmativamente a esta pregunta ha confundido la cuestionada probabilidad con

$$p(\overline{P}/A) = 1 - p(P/A) = 1 - 0{,}6 = 0{,}4,$$

es decir, con la probabilidad de no adquirir el producto habiendo visto el anuncio.

6.11 Demuéstrese el teorema de la probabilidad total.

SOLUCIÓN

Sea un espacio probabilístico, $(\Omega, \wp(\Omega), p)$, y una partición del espacio muestral, $\{A_i\}_{i=1}^{\infty}$, formada por sucesos de probabilidad distinta de cero. Por tratarse de una partición de Ω, el suceso seguro puede escribirse como

$$\Omega = \bigcup_{i=1}^{\infty} A_i,$$

con lo cual, cualquier suceso B admite la siguiente expresión:

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{\infty} A_i \right) = \bigcup_{i=1}^{\infty} (B \cap A_i),$$

siendo los sucesos $\{B \cap A_i\}_{i=1}^{\infty}$ disjuntos dos a dos.

Aplicando el tercer axioma de la probabilidad, se tiene que

$$p(B) = \sum_{i=1}^{\infty} p(B \cap A_i).$$

Ahora bien, puesto que, para cada suceso A_i, se cumple que

$$p(B/A_i) = \frac{p(B \cap A_i)}{p(A_i)},$$

resulta, despejando, que

$$p(B \cap A_i) = p(B/A_i) \cdot p(A_i),$$

esto es,

$$p(B) = \sum_{i=1}^{\infty} p(B \cap A_i) = \sum_{i=1}^{\infty} p(B/A_i) \cdot p(A_i).$$

6.12 Demuéstrese el teorema de Bayes.

SOLUCIÓN

Sea un espacio probabilístico, $(\Omega, \wp(\Omega), p)$, y una partición del espacio muestral, $\{A_i\}_{i=1}^{\infty}$, formada por sucesos de probabilidad distinta de cero y sea B un suceso de probabilidad no nula, entonces, por definición de probabilidad condicionada,

$$p(A_j/B) = \frac{p(B \cap A_j)}{p(B)},$$

para cualquier suceso A_j de la partición.

Considerando que, por un lado, el numerador de la expresión anterior, por aplicación de la regla de la multiplicación, es

$$p(B \cap A_j) = p(B/A_j) \cdot p(A_j),$$

y que, por otro lado, mediante el teorema de la probabilidad total, el denominador es

$$p(B) = \sum_{i=1}^{\infty} p(B/A_i) \cdot p(A_i),$$

se tiene, sustituyendo, que

$$p(A_j/B) = \frac{p(B/A_j) \cdot p(A_j)}{\sum_{i=1}^{\infty} p(B/A_i) \cdot p(A_i)}.$$

6.13 Sean A y B dos sucesos tales que $p(A) = 0{,}5$, $p(B) = 0{,}5$ y $p(A/B) = 0{,}4$. Calcúlense las siguientes probabilidades: $p(\overline{A}/B)$, $p(B/A)$ y $p(\overline{B}/A)$.

SOLUCIÓN

Puesto que la probabilidad condicionada es una probabilidad, se cumple que

$$p(\overline{A}/B) = 1 - p(A/B) = 1 - 0{,}4 = 0{,}6.$$

Para hallar $p(B/A)$, aplicamos la definición de probabilidad condicionada,

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

y calculamos el numerador de la expresión anterior mediante la regla de la multiplicación:

$$p(A \cap B) = p(B) \cdot p(A/B) = 0{,}5 \cdot 0{,}4 = 0{,}2.$$

En definitiva,

$$p(B/A) = \frac{0{,}2}{0{,}5} = 0{,}4.$$

Por último,

$$p(\overline{B}/A) = 1 - p(B/A) = 1 - 0{,}4 = 0{,}6.$$

6.14 Pruébese que, si un suceso es de probabilidad igual a cero, entonces, es independiente de cualquier otro suceso.

SOLUCIÓN

Sea el suceso A tal que $p(A) = 0$ y sea B un suceso cualquiera. Entonces, por un lado,

$$p(A) \cdot p(B) = 0 \cdot p(B) = 0,$$

y, por otro lado, como el suceso $A \cap B$ está contenido en el suceso A, aplicando la propiedad **P.5** se obtiene que

$$p(A \cap B) \leq p(A) = 0,$$

siendo, en consecuencia, igual a cero la probabilidad del suceso $A \cap B$.

En definitiva,

$$p(A \cap B) = p(A) \cdot p(B)$$

y el suceso A es independiente de cualquier suceso B.

6.15 Demuéstrese que, si los sucesos A y B son independientes, entonces, también lo son A y $\overline{B}$.

SOLUCIÓN

Para probar la independencia de los sucesos A y $\overline{B}$ ha de comprobarse que

$$p(A \cap \overline{B}) = p(A) \cdot p(\overline{B}).$$

Ahora bien,

$$p(A \cap \overline{B}) = p(A - B) = p(A) - p(A \cap B),$$

sin más que aplicar la propiedad **P.7**.

Teniendo en cuenta que los sucesos A y B son independientes, se verifica que

$$p(A \cap B) = p(A) \cdot p(B),$$

con lo cual, sustituyendo en la igualdad anterior resulta que

$$p(A \cap \overline{B}) = p(A) - p(A \cap B) = p(A) - p(A) \cdot p(B).$$

Sacando factor común a $p(A)$:

$$p(A \cap \overline{B}) = p(A) \cdot [1 - p(B)] = p(A) \cdot p(\overline{B}),$$

siendo, por tanto, A y $\overline{B}$ sucesos independientes.

Téngase en cuenta que, si A y $\overline{B}$ son independientes, volviendo a aplicar el resultado que acabamos de demostrar, también lo serán $\overline{B}$ y el complementario de A, $\overline{A}$.

6.16 Sean A y B dos sucesos independientes tales que $p(A) = 0{,}4$, y $p(B) = 0{,}5$. Hállense las probabilidades de los sucesos: $\overline{A}$, $\overline{B}$, $A \cap B$, $A \cup B$, $\overline{A} \cap \overline{B}$, $\overline{A} \cup \overline{B}$, $A \cap \overline{B}$ y $\overline{A} \cap B$.

SOLUCIÓN

Por las propiedades de la probabilidad, se calculan, tanto

$$p(\overline{A}) = 1 - p(A) = 1 - 0{,}4 = 0{,}6,$$

como

$$p(\overline{B}) = 1 - p(B) = 1 - 0{,}5 = 0{,}5.$$

Para hallar la probabilidad del suceso intersección hay que tener en cuenta que los sucesos A y B son independientes y, por consiguiente,

$$p(A \cap B) = p(A) \cdot p(B) = 0{,}4 \cdot 0{,}5 = 0{,}2.$$

Por otro lado, la probabilidad del suceso unión, se obtiene de modo inmediato, pues

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) = 0{,}4 + 0{,}5 - 0{,}2 = 0{,}7.$$

En cuanto a las probabilidades de los sucesos $\overline{A} \cap \overline{B}$ y $\overline{A} \cup \overline{B}$, basta aplicar las leyes de Morgan. Así, por un lado,

$$p(\overline{A} \cap \overline{B}) = p(\overline{A \cup B}) = 1 - p(A \cup B) = 1 - 0{,}7 = 0{,}3,$$

y, por otro,

$$p(\overline{A} \cup \overline{B}) = p(\overline{A \cap B}) = 1 - p(A \cap B) = 1 - 0{,}2 = 0{,}8.$$

Por último, y puesto que, como ya se ha dicho, los sucesos A y B son independientes, también los son A y $\overline{B}$ y $\overline{A}$ y B, con lo que

$$p(A \cap \overline{B}) = p(A) \cdot p(\overline{B}) = 0{,}4 \cdot 0{,}5 = 0{,}2$$

y

$$p(\overline{A} \cap B) = p(\overline{A}) \cdot p(B) = 0{,}6 \cdot 0{,}5 = 0{,}3.$$

6.17 Se realiza el experimento aleatorio consistente en extraer dos cartas de una baraja española. Calcúlense las probabilidades de los sucesos:

a) Las dos cartas son de oros.

b) La primera carta es de oros y la segunda de copas.

c) Una carta es de oros y la otra es de copas.

Considérense, para ello, las dos situaciones siguientes:

- Las cartas se extraen con reemplazamiento.
- Las cartas se extraen sin reemplazamiento.

SOLUCIÓN

Sirva este problema para introducir el concepto de *experimento compuesto* de varias etapas o experimentos; en este ejemplo, la primera y segunda etapas consisten en extraer la primera y la segunda carta, respectivamente. Aunque no formalizaremos los conceptos, sí daremos las pautas para trabajar en este tipo de situaciones.

a) En este orden de cosas, si denotamos por O_i ($i = 1,2$) el suceso la *i-ésima carta extraída es de oros,* ha de calcularse la probabilidad de que *la primera carta sea de oros y la segunda carta sea de oros,* suceso que puede interpretarse como una intersección entre los sucesos O_1 y O_2 es decir,

$$p(O_1 O_2),$$

probabilidad de un suceso en cuya expresión, por pertenecer a un experimento compuesto, no escribimos el símbolo de la intersección.

Esta probabilidad, aplicando la regla de la multiplicación[4], es igual a

$$p(O_1 O_2) = p(O_1) \cdot p(O_2/O_1).$$

Ahora bien, si se considera la primera situación en la que las cartas se extraen con reemplazamiento, la probabilidad $p(O_2/O_1)$ coincide con $p(O_2)$, puesto que el resultado de la primera etapa del experimento no influye en lo que ocurra en la segunda; se dice que los experimentos que componen el experimento compuesto son *independientes*. Así, teniendo en cuenta que los posibles resultados de cada uno de los experimentos que constituyen el experimento compuesto dan lugar a un número finito de resultados igualmente probables, se puede aplicar la regla de Laplace, por lo que

$$p(O_1 O_2) = p(O_1) \cdot p(O_2) = \frac{10}{40} \cdot \frac{10}{40} = 0{,}0625.$$

Si, por el contrario, la carta no es devuelta al mazo, en la segunda etapa del experimento quedarán únicamente 9 cartas de oros de un total de 39 resultados equiprobables, con lo cual, utilizando de nuevo la regla de Laplace, la probabilidad pedida es

$$p(O_1 O_2) = p(O_1) \cdot p(O_2/O_1) = \frac{10}{40} \cdot \frac{9}{39} = 0{,}0577.$$

En este caso, estamos trabajando con experimentos *dependientes.*

Antes de dar respuesta a los siguientes apartados, vamos a detenernos en otras posibles vías de resolución de problemas con experimentos compuestos.

La situación de independencia de experimentos se puede resolver, también, planteando un espacio muestral finito y equiprobable que consta de todas las posibles *ordenaciones* de dos cartas, repetidas o no, tomadas de entre las 40 que forman el mazo, lo cual constituye un total de 40^2 casos, es decir, el número de *variaciones con repetición* de 40 elementos tomados de dos en dos[5]. El número de casos favorables, variaciones con repetición de 10 elementos tomados de 2 en 2,

---

[4] Tanto la regla de la multiplicación, como los teoremas de la probabilidad total y de Bayes, se aplican sobre todo en situaciones de trabajo con experimentos compuestos.

[5] Dado un conjunto de m elementos $\{a_1, ..., a_m\}$, el número de ordenaciones de n elementos repetidos o no, que se pueden obtener es el número de *variaciones con repetición* de m elementos, tomados de n en n, $VR_{m,n} = m^n$.

son $10 \cdot 10$, ya que, por cada carta de oros —de un total de 10—, hay otras 10 cartas de oros para constituir una ordenación favorable. El lector puede preguntarse el porqué de considerar importante el orden cuando lo que aquí realmente interesa son las cartas que se reciben; la respuesta está en que nuestro interés reside en conseguir un espacio muestral que, además de finito, sea equiprobable, a partir del cual poder obtener las probabilidades de cualquier suceso compuesto.

En cuanto al caso de experimentos dependientes, también podría calcularse la probabilidad anterior, aplicando la regla de Laplace, para lo cual habría que considerar como espacio muestral finito y equiprobable el formado por todas las posibles ordenaciones de dos elementos que se pueden formar a partir de las 40 cartas de la baraja, esto es, el número de *variaciones* de 40 elementos tomadas de dos en dos[6], $40 \cdot 39$, siendo los casos favorables el número de variaciones de las 10 cartas de oros tomadas de dos en dos, $10 \cdot 9$.

Otra alternativa de cálculo de la probabilidad planteada en el caso de dependencia de experimentos surge de suponer el espacio muestral finito y equiprobable formado por todos los *grupos*, sin importar el orden, de dos cartas que se pueden elegir de 40. Así, el número de casos posibles, es decir, el número de grupos, todos ellos igualmente probables, es igual a las *combinaciones*[7] de 40 elementos tomados de dos en dos, $\binom{40}{2}$. En cuanto al número de casos favorables, éste será igual al número de grupos de dos cartas que pueden elegirse de un total de 10 cartas de oros: $\binom{10}{2}$.

Este último camino es muy recomendable cuando el experimento compuesto consta de un gran número de etapas.

b) Sea C_i $(i = 1,2)$ el suceso *la i-ésima carta extraída es de copas*. La probabilidad de que la primera carta sea de oros y la segunda de copas se calcula, igual que en el apartado anterior, aplicando la regla de la multiplicación:

$$p(O_1 C_2) = p(O_1) \cdot p(C_2/O_1).$$

Los comentarios realizados en el apartado ***a)*** son válidos a la hora de hallar esta probabilidad, tanto cuando hay reemplazamiento de la carta elegida en la primera etapa, esto es, cuando los experimentos son independientes,

$$p(O_1 C_2) = p(O_1) \cdot p(C_2) = \frac{10}{40} \cdot \frac{10}{40} = 0{,}0625,$$

como cuando no hay devolución de la carta, es decir, cuando los experimentos son dependientes,

$$p(O_1 C_2) = p(O_1) \cdot p(C_2/O_1) = \frac{10}{40} \cdot \frac{10}{39} = 0{,}0641.$$

---

[6] Dado un conjunto de m elementos $\{a_1, ..., a_m\}$, el número de ordenaciones de n elementos es el número de *variaciones* de m elementos, tomados de n en n, $V_{m,n} = m(m-1) \ldots (m-n+1)$.

[7] Recuérdese que, dado un conjunto de m elementos, $\{a_1, ..., a_m\}$, el número de combinaciones de n elementos que se pueden obtener a partir de él es igual al número de posibles grupos de n elementos, sin importar el orden de los mismos, $C_{m,n} = \binom{m}{n}$.

c) La probabilidad pedida resulta de sumar dos probabilidades:

$$p(\text{una carta de oros y otra de copas}) = p(O_1\, C_2) + p(C_1\, O_2).$$

El primer sumando está calculado en el apartado ***b)***. En cuanto al segundo sumando, se obtiene siguiendo los mismos razonamientos de anteriores apartados.

Así, cuando no hay reemplazamiento y, por tanto, los experimentos son independientes,

$$p(C_1\, O_2) = p(C_1) \cdot p(O_2) = \frac{10}{40} \cdot \frac{10}{40},$$

y, cuando la carta elegida no es devuelta al mazo, y, en consecuencia, los experimentos son dependientes,

$$p(C_1\, O_2) = p(C_1) \cdot p(O_2/C_1) = \frac{10}{40} \cdot \frac{10}{39}.$$

En definitiva,

$$p(\text{una carta de oros y otra de copas}) = 2 \cdot \frac{10}{40} \cdot \frac{10}{40} = 0{,}125,$$

en la primera situación, y

$$p(\text{una carta de oros y otra de copas}) = 2 \cdot \frac{10}{40} \cdot \frac{10}{39} = 0{,}128,$$

en la segunda.

Cuando no hay devolución al mazo de la carta elegida, la probabilidad pedida puede obtenerse, también, aplicando la regla de Laplace. En esta ocasión, el número de casos favorables es $\binom{10}{1} \cdot \binom{10}{1}$, esto es, por cada una de las 10 cartas de oros que se pueden elegir, hay otras 10 cartas de copas de entre las cuales tomar una.

En consecuencia,

$$p(\text{una carta de oros y otra de copas}) = \frac{\binom{10}{1} \cdot \binom{10}{1}}{\binom{40}{2}}.$$

6.18 Una entidad financiera ha concedido a sus clientes exclusivamente tres tipos de créditos. El 80 por ciento son hipotecarios, de los cuales un 10 por ciento son a interés fijo. El 15 por ciento son créditos personales, de los que un 6 por ciento son a

interés fijo. El resto son «supercrédito coche», todos a interés fijo. Con objeto de regalar un viaje, se elige al azar un cliente de entre los que poseen un crédito en la entidad.

a) ¿Cuál es la probabilidad de que el cliente elegido posea un crédito a interés fijo?

b) El cliente elegido tiene un crédito a interés fijo. ¿Cuál es la probabilidad de que posea un crédito hipotecario?

SOLUCIÓN

a) Sea F el suceso *el crédito es a interés fijo* y sean H, P y C, los sucesos *el crédito concedido es hipotecario*, *personal* o *supercrédito coche*, respectivamente. El enunciado del problema proporciona las siguientes probabilidades:

$$p(H) = 0{,}8,$$

$$p(F/H) = 0{,}1,$$

$$p(P) = 0{,}15,$$

$$p(F/P) = 0{,}06,$$

$$p(C) = 0{,}05$$

y

$$p(F/C) = 1.$$

Aplicando el teorema de la probabilidad total, la probabilidad de que, elegido un cliente al azar, posea un crédito a interés fijo es

$$p(F) = p(F/H) \cdot p(H) + p(F/P) \cdot p(P) + p(F/C) \cdot p(C)$$

y, sustituyendo,

$$p(F) = 0{,}1 \cdot 0{,}8 + 0{,}06 \cdot 0{,}15 + 1 \cdot 0{,}05 = 0{,}139.$$

b) La probabilidad de que, teniendo un crédito a interés fijo, sea un crédito hipotecario es

$$p(H/F)$$

que, aplicando el teorema de Bayes, resulta ser

$$p(H/F) = \frac{p(H) \cdot p(F/H)}{p(F)} = \frac{0{,}8 \cdot 0{,}1}{0{,}139} = 0{,}576.$$

6.19 La empresa láctea El buen vaquero tiene 1 000 empleados, de los cuales el 10 por ciento son directivos, el 15 por ciento técnicos, el 20 por ciento administrativos y resto operarios. El porcentaje de trabajadores que poseen estudios superiores dentro de cada categoría es del 90, 80, 20 y 4, respectivamente.

El consejo de administración ha decidido otorgar becas de formación a la mitad de los empleados que no posean titulación superior. ¿Cuántas becas otorgará la empresa?

SOLUCIÓN

Dados los sucesos: D, *ser directivo*, T, *ser técnico*, A, *ser administrativo*, O, *ser operario* y E, *poseer estudios superiores*, el problema proporciona las probabilidades siguientes:

$$p(D) = 0{,}1,\ p(T) = 0{,}15,\ p(A) = 0{,}2,\ p(O) = 0{,}55,$$

$$p(E/D) = 0{,}9,\ p(E/T) = 0{,}8,\ p(E/A) = 0{,}2 \text{ y } p(E/O) = 0{,}04.$$

Para hallar el número de becas que el consejo de administración ha decidido otorgar hay que calcular, en primer lugar, el porcentaje que, sobre el conjunto de empleados, representan aquellos que no poseen titulación superior, esto es, $p(\overline{E})$. Ahora bien,

$$p(\overline{E}) = 1 - p(E),$$

y, aplicando el teorema de la probabilidad total, se tiene que

$$p(E) = p(E/D) \cdot p(D) + p(E/T) \cdot p(T) + p(E/A) \cdot p(A) + p(E/O) \cdot p(O),$$

es decir,

$$p(E) = 0{,}9 \cdot 0{,}1 + 0{,}8 \cdot 0{,}15 + 0{,}2 \cdot 0{,}2 + 0{,}04 \cdot 0{,}55 = 0{,}272,$$

y, por tanto,

$$p(\overline{E}) = 1 - p(E) = 1 - 0{,}272 = 0{,}728.$$

En consecuencia, el 72,8 por ciento de los empleados de la empresa láctea no posee estudios superiores, o lo que es lo mismo, 728 empleados. Ello supone que el consejo de administración otorgará $782/2 = 364$ becas de formación.

6.20 La cadena de televisión privada Canalmenos desea captar socios en una determinada ciudad. Para ello envía propaganda al 75 por ciento de los domicilios. Los datos que figuran en la siguiente tabla corresponden a las probabilidades de abonarse a esta cadena de las familias, según reciban o no la citada propaganda.

	Recibe	No recibe
Probabilidad	0,20	0,05

Una familia se ha abonado a esta nueva cadena. ¿Cuál es la probabilidad de que no haya recibido la propaganda?

SOLUCIÓN

Según los datos del problema, la probabilidad del suceso, *recibir propaganda*, R, es 0,75. Se sabe, además, que la probabilidad del suceso, *la familia se abona*, A, es 0,20, si ha recibido la información y 0,05, si no la ha recibido, es decir,

$$p(A/R) = 0{,}20$$

y

$$p(A/\bar{R}) = 0{,}05.$$

En definitiva, de la aplicación del teorema de Bayes resulta:

$$p(\bar{R}/A) = \frac{p(A/\bar{R}) \cdot p(\bar{R})}{p(A/R) \cdot p(R) + p(A/\bar{R}) \cdot p(\bar{R})} = \frac{0{,}05 \cdot 0{,}25}{0{,}20 \cdot 0{,}75 + 0{,}05 \cdot 0{,}25} = 0{,}077,$$

probabilidad de que una familia no haya recibido la propaganda habiéndose abonado.

6.21 Durante el pasado año el organismo público encargado de la formación de los trabajadores del sector del metal convocó un total de 462 planes de formación. Publicada la convocatoria se presentó un cierto número de solicitudes, de las cuales el 50 por ciento correspondía a planes individuales solicitados por las empresas, el 40 por ciento a planes agrupados solicitados por agrupaciones de empresas y el resto a otras entidades.

Una vez revisadas las solicitudes se concedió a las empresas el 40 por ciento de los planes solicitados, a las agrupaciones de empresas el 50 por ciento y al resto de las entidades el 20 por ciento.

a) ¿Cuántas solicitudes fueron presentadas?

b) Se elige una solicitud al azar. ¿Cuál es la probabilidad de que corresponda a un plan agrupado y haya sido denegada?

c) ¿Cuál es la probabilidad de que una solicitud que ha sido denegada corresponda a un plan agrupado?

SOLUCIÓN

Se consideran los sucesos, I, A y E, *la solicitud corresponde a un plan individual*, *la solicitud corresponde a un plan agrupado*, *la solicitud corresponde a otro tipo de entidad*, y sea C el suceso *la solicitud ha sido concedida.*

La información disponible es

$$p(I) = 0{,}5,$$

$$p(A) = 0{,}4,$$

$$p(E) = 0{,}1,$$

$$p(C/I) = 0{,}4,$$

$$p(C/A) = 0{,}5$$

y

$$p(C/E) = 0{,}2.$$

a) Aplicando el teorema de la probabilidad total al suceso *la solicitud ha sido concedida*, resulta que

$$p(C) = p(C/I) \cdot p(I) + p(C/A) \cdot p(A) + p(C/E) \cdot p(E),$$

esto es,

$$p(C) = 0{,}4 \cdot 0{,}5 + 0{,}5 \cdot 0{,}4 + 0{,}2 \cdot 0{,}1 = 0{,}42,$$

o, equivalentemente, al 42 por ciento de las solicitudes presentadas se les concedió un plan de formación.

Dado que la convocatoria consta de 462 planes, el total de solicitudes presentadas es igual a

$$n = \frac{462 \cdot 100}{42} = 1\,100.$$

b) La probabilidad de que una solicitud elegida al azar *corresponda a un plan agrupado y haya sido denegada* es

$$p(A\overline{C}),$$

donde $\overline{C}$ es el complementario del suceso *la solicitud ha sido aceptada*.

Aplicando la regla de la multiplicación,

$$p(A\overline{C}) = p(A) \cdot p(\overline{C}/A),$$

y teniendo en cuenta que

$$p(\overline{C}/A) = 1 - p(C/A) = 1 - 0{,}5 = 0{,}5,$$

resulta:

$$p(A\overline{C}) = 0{,}4 \cdot 0{,}5 = 0{,}2.$$

c) La probabilidad condicionada

$$p(A/\overline{C})$$

se halla aplicando el teorema de Bayes:

$$p(A/\overline{C}) = \frac{p(A) \cdot p(\overline{C}/A)}{P(\overline{C})}.$$

Aunque el denominador de la expresión anterior se obtiene directamente a partir de la probabilidad del suceso C calculada en el apartado ***a)***, ya que

$$p(\overline{C}) = 1 - p\,(C) = 1 - 0{,}42 = 0{,}58,$$

presentamos un procedimiento alternativo para que el lector se familiarice con las propiedades de la probabilidad.

Así, por el teorema de probabilidad total,

$$p(\overline{C}) = p(I) \cdot p\,(\overline{C}/I) + p(A) \cdot p(\overline{C}/A) + p(E) \cdot p(\overline{C}/E),$$

donde

$$p(\overline{C}/I) = 1 - p\,(C/I) = 1 - 0{,}4 = 0{,}6,$$

$$p(\overline{C}/E) = 1 - p\,(C/E) = 1 - 0{,}2 = 0{,}8$$

y

$$p(\overline{C}/A) = 0{,}5,$$

probabilidad calculada en el apartado anterior.

En definitiva,

$$p(\overline{C}) = 0{,}5 \cdot 0{,}6 + 0{,}4 \cdot 0{,}5 + 0{,}1 \cdot 0{,}8 = 0{,}58,$$

y, por tanto,

$$p(A/\overline{C}) = \frac{p(A) \cdot p(\overline{C}/A)}{p(\overline{C})} = \frac{0{,}4 \cdot 0{,}5}{0{,}58} = 0{,}345.$$

Téngase en cuenta que, en las dos probabilidades halladas en los apartados ***b)*** y ***c)***, intervienen los sucesos A y $\overline{C}$ y pero, mientras la primera es la probabilidad de la *intersección* de ambos sucesos, la segunda es una probabilidad *condicionada.*

6.22 Las probabilidades que figuran en la tabla siguiente corresponden al número de hamburguesas encargadas semanalmente a Fono-Burguer por las familias de una zona residencial:

N.º de hamburguesas	0	1	2	3	4
Probabilidad	0,3	0,15	0,18	0,25	0,12

Se sabe, además, que el 65 por ciento de las familias que no encargan hamburguesas no tienen hijos.

a) Si una familia ha encargado al menos 3 hamburguesas, ¿cuál es la probabilidad de que encargue exactamente 4 hamburguesas?

b) ¿Qué porcentaje de familias tienen hijos y no encargan hamburguesas?

SOLUCIÓN

a) Denotando por A al suceso *una familia encarga al menos 3 hamburguesas* y por B al suceso *una familia encarga exactamente 4 hamburguesas*, la probabilidad pedida puede expresarse como

$$p(B/A) = \frac{p(AB)}{p(A)}.$$

A la vista de la información que proporciona el enunciado, el suceso A puede considerarse como unión de dos sucesos: *una familia encarga exactamente 3 hamburguesas* y *una familia encarga exactamente 4 hamburguesas*; en consecuencia, el denominador de la fracción anterior es

$$p(A) = 0{,}25 + 0{,}12 = 0{,}37.$$

Por lo que respecta a la intersección entre A y B, obviamente coincide con el suceso B, cuya probabilidad es 0,12.

En definitiva,

$$p(B/A) = \frac{0{,}12}{0{,}37} = 0{,}324.$$

b) Llamando H al suceso *una familia tiene hijos* y N al suceso *una familia no encarga hamburguesas*, por el enunciado se conoce la probabilidad:

$$p(\overline{H}/N) = 0{,}65.$$

Teniendo en cuenta que

$$p(H/N) = 1 - p(\overline{H}/N) = 1 - 0{,}65 = 0{,}35,$$

y, aplicando la regla de la multiplicación, se obtiene la probabilidad pedida:

$$p(HN) = p(N) \cdot p(H/N) = 0{,}3 \cdot 0{,}35 = 0{,}105,$$

esto es, el 10,5 por ciento de las familias tienen hijos y no encargan hamburguesas.

6.23 El 90 por ciento de los electrodomésticos que se venden en la cadena de tiendas Electronuevo son de la marca Agnus. Se sabe que la probabilidad de que un cliente adquiera una lavadora y pertenezca a la marca Agnus es 0,35; de que sea un frigorífico y de esta marca es 0,25; y de que sea un televisor y de esta marca es 0,20.

a) Un cliente entra en el local y adquiere un electrodoméstico Agnus. ¿Cuál es la probabilidad de que sea una lavadora?

b) Si la probabilidad de que un cliente compre una lavadora de otra marca es 0,18, ¿cuál es la probabilidad de que un cliente elegido al azar adquiera una lavadora?

SOLUCIÓN

Sea A el suceso, *el electrodoméstico que se vende pertenece a la marca Agnus* y sean L, F y T, respectivamente, los sucesos *el electrodoméstico que se vende es una lavadora*, *es un frigorífico* y *es un televisor*.

Se sabe que

$$p(LA) = 0{,}35$$

$$p(FA) = 0{,}25$$

$$p(TA) = 0{,}20$$

y, además,

$$p(A) = 0{,}9.$$

Obsérvese que

$$p\,(LA) + p\,(FA) + p(TA) = 0{,}8,$$

siendo 0,9 la probabilidad de que *un electrodoméstico sea de la marca Agnus*; ello significa que lavadoras, frigoríficos y televisores no son los únicos electrodomésticos de esta marca que se venden en este establecimiento, pues, si así fuera, la suma de las tres probabilidades anteriores debería ser 0,9.

a) Para calcular la probabilidad del suceso L sabiendo que ha ocurrido el suceso A, se aplica la definición de probabilidad condicionada:

$$p\,(L/A) = \frac{p(LA)}{p(A)} = \frac{0{,}35}{0{,}9} = 0{,}39.$$

b) Como, según el enunciado, la probabilidad de que un cliente compre una lavadora de otra marca es

$$p(L\bar{A}) = 0{,}18,$$

se obtiene que la probabilidad de que un cliente adquiera una lavadora es

$$p(L) = p\,(LA) + p\,(L\bar{A}) = 0{,}35 + 0{,}18 = 0{,}53.$$

6.24 La empresa Castiguija, S. A., para reparar las aceras de una localidad recibe las baldosas en lotes de 10 unidades. De un estudio previo se sabe que las probabilidades que aparecen en la siguiente tabla corresponden al número de baldosas defectuosas del producto en un lote:

N.º de defectuosas	0	1	2
Probabilidad	0,39	0,56	0,05

Cada lote pasa por un proceso de control de calidad de modo que se eligen dos baldosas y, si ambas son buenas, se acepta, y, en caso contrario, se rechazan. ¿Cuál es la probabilidad de que un lote sea aceptado?

SOLUCIÓN

Sea B el suceso *el lote es aceptado* y sea L_i, con $i = 0, 1, 2$, el suceso *el lote tiene i baldosas defectuosas*. Para calcular la probabilidad de que el lote se acepte hay que tener en cuenta cuántas baldosas defectuosas hay en él.

Así, la probabilidad de que el lote sea aceptado si no tiene baldosas defectuosas es uno,

$$p\,(B/L_0) = 1,$$

la probabilidad de que sea aceptado, teniendo una baldosa defectuosa es

$$p\,(B/L_1) = \frac{9}{10} \cdot \frac{8}{9} = 0{,}8,$$

y, por último, la probabilidad de que sea aceptado, si tiene dos baldosas defectuosas es

$$p\,(B/L_2) = \frac{8}{10} \cdot \frac{7}{9} = 0{,}62.$$

El cálculo de las dos últimas probabilidades se ha realizado considerando que se trata de un experimento compuesto de dos extracciones sin reemplazamiento y que, por tanto, el resultado de la primera extracción influye en el resultado de la segunda.

Aplicando, entonces, el teorema de la probabilidad total, se tiene que

$$p\,(B) = p\,(L_0) \cdot p\,(B/L_0) + p\,(L_1) \cdot p(B/L_1) + p(L_2) \cdot p(B/L_2),$$

y, en definitiva,

$$p(B) = 0{,}39 \cdot 1 + 0{,}56 \cdot 0{,}8 + 0{,}05 \cdot 0{,}62 = 0{,}869.$$

6.25 El 50 por ciento de la población activa de un país se dedica al sector servicios, el 12 por ciento al de la construcción, el 3 por ciento al sector primario y el resto al industrial. La tasa de paro de este país es del 23 por ciento, siendo en el sector servicios del 18,6 por ciento, en el sector primario del 10 por ciento y en el sector industrial del 28 por ciento.

Si un individuo está en paro, ¿cuál es la probabilidad de que pertenezca al sector de la construcción?

SOLUCIÓN

Se consideran los sucesos, S, *pertenecer al sector servicios*, C, *pertenecer al sector de la construcción*, P, *pertenecer al sector primario*, I, *pertenecer al sector industrial*, y, E, *estar en paro*.

Se conocen las probabilidades:

$$p(S) = 0{,}5,\ p(C) = 0{,}12,\ p(P) = 0{,}03,\ p(I) = 0{,}35 \text{ y } p(E) = 0{,}23,$$

junto con las probabilidades condicionadas:

$$p(E/S) = 0{,}186,\ p(E/P) = 0{,}1 \text{ y } p(E/I) = 0{,}28.$$

Aplicando la definición de probabilidad condicionada, se tiene que la probabilidad pedida es

$$p(C/E) = \frac{p(CE)}{p(E)}.$$

Ahora bien, por la regla de la multiplicación, el numerador de la expresión anterior resulta ser

$$p(CE) = p(E/C) \cdot p(C),$$

con lo cual, sustituyendo,

$$p(C/E) = \frac{p(E/C) \cdot p(C)}{p(E)}.$$

Para hallar $p(E/C)$, única probabilidad desconocida en la fracción anterior, basta tener en cuenta, por el teorema de la probabilidad total, que

$$p(E) = p(E/S) \cdot p(S) + p(E/C) \cdot p(C) + p(E/P) \cdot p(P) + p(E/I) \cdot p(I),$$

por lo cual, despejando,

$$p(E/C) = \frac{p(E) - p(E/S) \cdot p(S) - p(E/P) \cdot p(P) - p(E/I) \cdot p(I)}{p(C)},$$

y, sustituyendo, se tiene que

$$p(E/C) = \frac{0{,}23 - 0{,}186 \cdot 0{,}5 - 0{,}03 \cdot 0{,}1 - 0{,}28 \cdot 0{,}35}{0{,}12} = 0{,}3.$$

En definitiva,

$$p(C/E) = \frac{0{,}3 \cdot 0{,}12}{0{,}23} = 0{,}1565$$

es la probabilidad pedida.

6.26 Sean A, B y C tres sucesos tales que A y B son independientes y, además, $A \cup B \subset C$, demuéstrese que, entonces,

$$p(\overline{C}) \leq p(\overline{A}) \cdot p(\overline{B}).$$

SOLUCIÓN

La relación de inclusión entre los sucesos $A \cup B$ y C se verifica de modo inverso entre los complementarios de estos sucesos, esto es,

$$\overline{C} \subset \overline{A \cup B},$$

y, por consiguiente,

$$p(\overline{C}) \leq p(\overline{A \cup B}) = p(\overline{A} \cap \overline{B}),$$

siendo esta última igualdad el resultado de aplicar una de las leyes de Morgan.

Por otro lado, según se probó en el problema **6.15**, si los sucesos A y B son independientes, también lo son sus complementarios, con lo que

$$p(\overline{C}) \leq p(\overline{A} \cap \overline{B}) = p(\overline{A}) \cdot p(\overline{B}),$$

como se quería demostrar.

6.27 Debido al consumo de piensos en mal estado, se sospecha que algunas granjas de pollos de la región de Belcagio tienen animales intoxicados.

El departamento de sanidad ha enviado un inspector con la misión de clausurar aquellas granjas que posean animales enfermos. Para ello, tiene orden de examinar un 5 por ciento de los pollos que haya en cada granja y clausurar aquellas en las que se encuentre al menos uno intoxicado.

a) ¿Cuál es la probabilidad de que sea clausurada una granja que tiene 100 pollos y 3 de ellos en mal estado?

b) ¿Cuál es la probabilidad de que en la inspección de dicha granja se encuentre un pollo intoxicado y éste sea el último que se examina?

SOLUCIÓN

a) El suceso *al menos un pollo está intoxicado* es complementario del suceso *ningún pollo está intoxicado*, con lo que

$$p(\text{al menos uno intoxicado}) = 1 - p(\text{ninguno intoxicado}).$$

Teniendo en cuenta que son 5 los pollos que se van a examinar y que 3 de los 100 pollos están en mal estado, y denotando por $\bar{I}_i$ al suceso el *i-ésimo pollo examinado está en buen estado*, resulta:

$$p(\text{ninguno intoxicado}) = p(\bar{I}_1 \bar{I}_2 \bar{I}_3 \bar{I}_4 \bar{I}_5),$$

con lo cual, aplicando la regla de la multiplicación, la probabilidad pedida es

$$p(\bar{I}_1) \cdot p(\bar{I}_2/\bar{I}_1) \cdot p(\bar{I}_3/\bar{I}_1\bar{I}_2) \cdot p(\bar{I}_4/\bar{I}_1\bar{I}_2\bar{I}_3) \cdot p(\bar{I}_5/\bar{I}_1\bar{I}_2\bar{I}_3\bar{I}_4) = \frac{97}{100} \cdot \frac{96}{99} \cdot \frac{95}{98} \cdot \frac{94}{97} \cdot \frac{93}{96} = 0{,}856.$$

Y, en definitiva, la probabilidad de que la granja sea clausurada es

$$p(\text{al menos uno intoxicado}) = 1 - 0{,}856 = 0{,}144.$$

Resolviendo este problema mediante la regla de Laplace, el número de casos posibles es $\binom{100}{5}$, esto es, las posibles elecciones de grupos de 5 pollos de entre los 100 pollos de la granja; de igual forma, el número de casos favorables es el número de elecciones de grupos de 5 pollos de entre los 97 pollos que están en buen estado, es decir, $\binom{97}{5}$.

b) Siguiendo la misma notación y metodología del apartado anterior, la probabilidad pedida es

$$p(\bar{I}_1\bar{I}_2\bar{I}_3\bar{I}_4I_5) = \frac{97}{100}\cdot\frac{96}{99}\cdot\frac{95}{98}\cdot\frac{94}{97}\cdot\frac{3}{96} = 0{,}0276.$$

6.28 Un bar de una ciudad entrega papeletas al azar a sus 100 primeros clientes, de las cuales 10 tienen como premio una segunda consumición. El primer grupo que entra en el local está formado por cuatro amigos. Si todos consumen, ¿cuál es la probabilidad de que exactamente uno de ellos gane una segunda copa?

SOLUCIÓN

Puede considerarse que se eligen 4 papeletas sin reemplazamiento (ninguna papeleta es devuelta) de un total de 100 papeletas de las cuales 10 tienen premio. Entonces, la probabilidad de que *exactamente uno de ellos gane una segunda copa* es equivalente a la probabilidad de que *tres no ganen la copa y uno sí.*

Denotando por G_i y $\overline{G}_i$ los sucesos *papeleta premiada* y *papeleta no premiada* en la *i*-ésima elección, y suponiendo que el primero de los amigos es quien gana la consumición, la probabilidad de este suceso es

$$p(G_1\overline{G}_2\overline{G}_3\overline{G}_4) = p(G_1)\cdot p(\overline{G}_2/G_1)\cdot p(\overline{G}_3/G_1\overline{G}_2)\cdot p(\overline{G}_4/G_1\overline{G}_2\overline{G}_3) = \frac{10}{100}\cdot\frac{90}{99}\cdot\frac{89}{98}\cdot\frac{88}{97}.$$

Pero esta no es la única forma de que resulte premiado uno solo de los cuatro amigos: hay tantas situaciones favorables como ordenaciones puedan hacerse en la elección de tres papeletas sin premio y una premiada, esto es, *permutaciones con repetición*[8] de 4 elementos, de los cuales 3 están repetidos:

$$PR_4^{1,3} = \frac{4!}{3!\cdot 1!} = 4.$$

[8] Recuérdese que, dado un conjunto de m elementos $\{a_1, \ldots, a_m\}$, de los cuales hay r iguales entre sí y distintos al resto, …, v iguales entre sí, y distintos a los demás, el número de *permutaciones con repetición* de estos m elementos es el número de todas sus posibles ordenaciones, esto es, $PR_m^{r,\ldots,v} = \dfrac{m!}{r!\cdot\ldots\cdot v!}$.

Adviértase que este número se obtiene de modo inmediato, *en este caso*, ya que el número de ordenaciones coincide con los cuatro amigos a los que les puede tocar la papeleta premiada.

Como, evidentemente, todos estos sucesos tienen la misma probabilidad, la probabilidad pedida es

$$p(\text{exactamente uno gane la segunda copa}) = 4 \cdot \frac{10}{100} \cdot \frac{90}{99} \cdot \frac{89}{98} \cdot \frac{88}{97} = 0{,}3.$$

También puede hallarse esta probabilidad mediante la regla de Laplace. Así, el número de casos posibles en la elección simultánea de las 4 papeletas de un total de 100 es $\binom{100}{4}$, siendo el número de casos favorables $\binom{10}{1} \cdot \binom{90}{3}$.

6.29 El restaurante Comersano recibe diariamente un pedido de cajas con dos docenas de huevos cada una. El encargado de cocina revisa la mercancía, y es aceptada una caja, si elegida una muestra aleatoria de 2 huevos, resulta a lo sumo uno roto. Suponiendo que la caja que va a ser examinada tiene 3 huevos rotos, ¿cuál es la probabilidad de que se acepte?

SOLUCIÓN

Para hallar la probabilidad pedida lo más fácil es considerar el suceso complementario:

$$p(\text{la caja es aceptada}) = 1 - p(\text{la caja no es aceptada}).$$

Ahora bien, para que una caja no sea aceptada, los dos huevos han de estar rotos, por lo que, denotando por R_i el suceso *el huevo i-ésimo está roto*, resulta que

$$p(\text{la caja no es aceptada}) = p(R_1R_2) = p(R_1) \cdot p(R_2/R_1).$$

Puesto que la caja que va a ser examinada contiene 3 huevos rotos y 21 no rotos, la probabilidad anterior es igual a

$$\frac{3}{24} \cdot \frac{2}{23} = 0{,}0109.$$

Esta probabilidad puede obtenerse también utilizando la regla de Laplace, siendo el número de casos posibles $\binom{24}{2}$ y el número de casos favorables $\binom{3}{2}$.

En definitiva, la probabilidad de que la caja sea aceptada es $1 - 0{,}0109 = 0{,}9891$.

6.30 Una empresa de transformados metálicos tiene 15 empleados, de los cuales 8 llevan más de diez años trabajando en la empresa, 4 llevan dos años y el resto tienen contrato de prueba.

Se selecciona al azar una muestra de 4 empleados para realizar un curso de especialización. ¿Cuál es la probabilidad de que se elijan los 3 empleados con contrato de prueba?

SOLUCIÓN

Denominando P_i al suceso *el i-ésimo empleado elegido tiene contrato de prueba*, un suceso con el cual se cumple la situación de que 3 de los empleados elegidos tienen contrato de prueba es, $P_1P_2P_3\overline{P}_4$, cuya probabilidad es

$$p(P_1P_2P_3\overline{P}_4) = p(P_1) \cdot p(P_2/P_1) \cdot p(P_3/P_1P_2) \cdot p(\overline{P}_4/P_1P_2P_3) = \frac{3}{15} \cdot \frac{2}{14} \cdot \frac{1}{13} \cdot \frac{12}{12}.$$

Según el orden de elección del empleado que *no* tiene contrato de prueba, hay 4 casos que dan lugar al suceso cuya probabilidad tenemos que calcular, todos ellos con la misma probabilidad, con lo cual,

$$p(\text{elegir 3 empleados con contrato de prueba}) = 4 \cdot \frac{3}{15} \cdot \frac{2}{14} \cdot \frac{1}{13} \cdot \frac{12}{12} = 0{,}0088.$$

Aplicando la regla de Laplace para la resolución de este problema, resulta que el número de casos posibles es igual a

$$\binom{15}{4} = \frac{15 \cdot 14 \cdot 13 \cdot 12}{4 \cdot 3 \cdot 2} = 1\,365,$$

esto es, todos los grupos de 4 empleados que pueden elegirse de un total de 15 empleados que tiene la empresa.

Para hallar el número de casos favorables hay que considerar que el grupo de empleados que se elija ha de incluir los 3 empleados con contrato de prueba que pueden combinarse con cualquiera de los 12 empleados restantes; en definitiva, son 12 los casos favorables.

Por tanto, la probabilidad pedida es, como ya sabíamos,

$$\frac{12}{1\,365} = 0{,}0088.$$

6.31 Una empresa de recolección de aceituna contrata a 50 trabajadores en la época de recogida. De ellos, 40 tienen familia numerosa, 6 tienen dos hijos y el resto no tiene hijos.

Antes de iniciar la recogida, el empresario concede al azar 10 ayudas familiares. ¿Cuál es la probabilidad de que todas ellas recaigan en los trabajadores con familia numerosa?

SOLUCIÓN

Dado que el número de becas es elevado, la resolución de este ejercicio resulta más sencilla aplicando la regla de Laplace. Así, el número de casos posibles es el número de elecciones de 10 trabajadores de un total de 50, esto es, $\binom{50}{10}$.

En cuanto al número de casos favorables, será igual al número de elecciones de 10 trabajadores que se puedan realizar dentro del grupo de 40 trabajadores con familia numerosa, es decir, $\binom{40}{10}$.

En consecuencia, la probabilidad pedida resulta ser igual a

$$p(\text{las ayudas sean para los trabajadores con familia numerosa}) = \frac{\binom{40}{10}}{\binom{50}{10}} = \frac{\dfrac{40!}{10! \cdot 30!}}{\dfrac{50!}{10! \cdot 40!}} = 0{,}0825.$$

6.32 El servicio de mantenimiento del Ayuntamiento de Villahermosa cuenta con tres secciones: limpieza, reparaciones menores y jardinería. A cada sección se han adscrito 10 trabajadores.

Con objeto de llevar a cabo un control de asistencia al trabajo, el jefe cita a 3 trabajadores de una sección elegida al azar: si al menos uno de ellos falta de su puesto de trabajo, descuenta el plus de productividad a todos los trabajadores de la sección. ¿Cuál es la probabilidad de que un día en el que han asistido a su puesto 4, 8 y 3 trabajadores de las secciones de limpieza, reparaciones y jardinería, respectivamente, se descuente de la elegida dicho plus?

SOLUCIÓN

Sea P el suceso *al menos uno de los trabajadores elegidos falta de su puesto de trabajo,* cuya probabilidad se calcula más cómodamente, teniendo en cuenta que $p(P) = 1 - p(\overline{P})$, donde $\overline{P}$ es el suceso *ninguno de los trabajadores elegidos falta de su puesto de trabajo.*

Aplicando el teorema de probabilidad total, se tiene que

$$p(\overline{P}) = p(\overline{P}/L) \cdot p(L) + p(\overline{P}/R) \cdot p(R) + p(\overline{P}/J) \cdot P(J),$$

donde L, R y J son, respectivamente, los sucesos *pertenecer a la sección de limpieza*, *a la de reparaciones* y *a la de jardines.*

Si suponemos que todas las secciones tienen la misma probabilidad de ser elegidas, entonces,

$$p(L) = p(R) = p(J) = \frac{1}{3}.$$

En cuanto a las probabilidades condicionadas, mediante la regla de Laplace,

$$p(\overline{P}/L) = \frac{\binom{4}{3}}{\binom{10}{3}} = 0{,}033,$$

$$p(\overline{P}/R) = \frac{\binom{8}{3}}{\binom{10}{3}} = 0{,}467$$

y

$$p(\overline{P}/J) = \frac{\binom{3}{3}}{\binom{10}{3}} = 0{,}0083.$$

El cálculo de las anteriores probabilidades se ha realizado, teniendo en cuenta que las tres secciones tienen el mismo número de trabajadores y, por tanto, el número de casos posibles es $\binom{10}{3}$ en las tres secciones; para hallar el número de casos favorables se han elegido 3 trabajadores de entre los que han asistido al trabajo en cada sección.

Por consiguiente,

$$p(\overline{P}) = 0{,}033 \cdot \frac{1}{3} + 0{,}467 \cdot \frac{1}{3} + 0{,}0083 \cdot \frac{1}{3} = 0{,}1694,$$

y, en definitiva,

$$p(P) = 1 - 0{,}1694 = 0{,}8306.$$

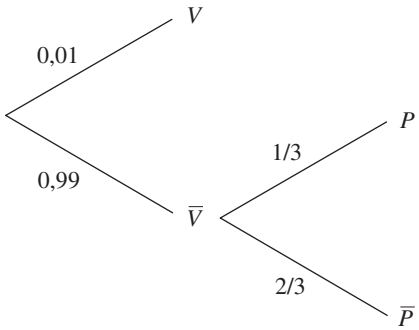
6.33 El 1 por ciento de las tabletas de chocolate La Ricura tienen como premio directo un viaje a Eurodisney. El resto tienen una etiqueta en su envoltorio con 3 casillas ocultas de las cuales una, y sólo una, tiene premio de consolación; para optar a dicho premio, el comprador deberá «rascar» una sola casilla.

Pepito Pérez ha obtenido premio con su tableta. ¿Cuál es la probabilidad de que haya sido su ansiado viaje?

SOLUCIÓN

Dados los sucesos, *G*, *ganar algún tipo de premio*, *V*, *obtener el viaje*, y, *P*, *ganar el premio de consolación*, el siguiente *diagrama de árbol,* representación gráfica de apoyo que permite calcular de forma cómoda probabilidades de sucesos pertenecientes a experimentos compuestos de varias etapas, ilustra las diferentes situaciones que pueden presentarse.

En un diagrama de árbol el paso de una etapa a otra se representa con *ramas* que parten del mismo origen y reflejan los distintos estados de llegada.



Según se comprueba en el que aquí se presenta, las ramas de paso de cada etapa a la siguiente tienen probabilidades cuyas suma es igual a uno. Así, en la primera etapa,

$$p(V) = 0{,}01$$

y

$$p(\overline{V}) = 1 - p(V) = 0{,}99.$$

Además, si no se consigue el viaje, la probabilidad de ganar todavía un premio es la probabilidad de acertar con la casilla premiada del envoltorio,

$$p(P/\overline{V}) = \frac{1}{3}$$

y, por tanto,

$$p(\overline{P}/\overline{V}) = \frac{2}{3}.$$

Para calcular, la probabilidad pedida,

$$p(V/G) = \frac{p(VG)}{p(G)},$$

hay que tener en cuenta que el suceso *VG*, *obtener viaje y ganar algún tipo de premio*, coincide con el suceso *V*, ya que si se obtiene el viaje, se está, de hecho, ganando un premio; en consecuencia,

$$p(VG) = p(V).$$

En cuanto al denominador de la fracción anterior, los compradores con premio son los que directamente ganan el viaje, o bien, los que no ganan el viaje pero rascan la casilla con premio de consolación. Utilizando el diagrama de árbol se tiene que

$$p(G) = p(V) + p(\overline{V}P) = p(V) + p(\overline{V}) \cdot p(P/\overline{V}) = 0{,}01 + 0{,}99 \cdot \frac{1}{3} = 0{,}34.$$

Se concluye, entonces, que la probabilidad de obtener el viaje, habiendo obtenido premio es

$$p(V/G) = \frac{0{,}01}{0{,}34} = 0{,}029.$$

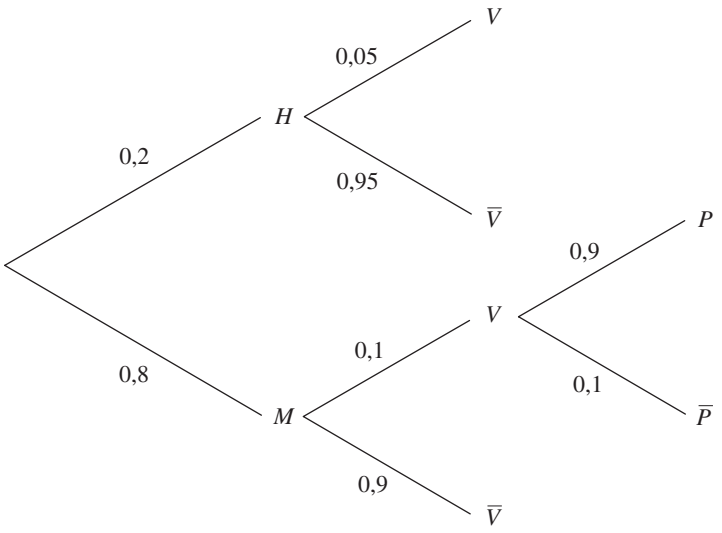
6.34 En una compañía aseguradora el 20 por ciento de los trabajadores son varones y, de ellos, el 5 por ciento son vendedores de seguros, siendo este porcentaje de un 10 por ciento en las mujeres. Además, el 90 por ciento de las vendedoras de seguros está realizando un curso de perfeccionamiento de ventas.

Hállese la probabilidad de que un trabajador elegido al azar esté realizando un curso de perfeccionamiento de ventas.

SOLUCIÓN

El experimento que se describe en este problema puede descomponerse en tres etapas: en una primera etapa se considera el sexo del trabajador, en una segunda etapa se clasifica a los trabajadores de cada sexo según sean o no vendedores de seguros y, por último, dentro de las mujeres vendedoras de seguros se distingue entre las que realizan o no un curso de perfeccionamiento de ventas.

En el siguiente diagrama de árbol se describen las tres etapas del experimento, donde H y M son los sucesos *ser hombre* y *ser mujer*; V y $\overline{V}$, los sucesos *ser vendedor de seguros* y *no ser vendedor de seguros*; y, por último, P y $\overline{P}$ los sucesos *recibir un curso de perfeccionamiento* y *no recibirlo*.



La probabilidad de que, elegido un trabajador al azar, *realice un curso de perfeccionamiento* es $p(P)$; ahora bien, solamente realizan el curso de perfeccionamiento las mujeres vendedoras de seguros, con lo que la probabilidad de este suceso es, de hecho, la probabilidad *de ser mujer, vendedora de seguros* y *realizar el curso de perfeccionamiento*:

$$p(P) = p(MVP).$$

Utilizando el diagrama de árbol se obtiene la probabilidad pedida:

$$p(MVP) = p(M) \cdot p(V/M) \cdot p(P/MV) = 0{,}8 \cdot 0{,}1 \cdot 0{,}9 = 0{,}072.$$

6.35 Con el fin de promocionar el turismo de una zona costera, se elabora un informe del que se desprenden, entre otros, los datos que se detallan a continuación.

El 65 por ciento de la población mayor de edad de la próspera comarca de Marcerón vive en la localidad de Marcera de Arriba y el resto en Marcera de Abajo. El 30 por ciento de los habitantes con mayoría de edad en Marcera de Arriba son jóvenes (entre 18 y 30 años), el 25 por ciento adultos (entre 30 y 65 años) y el resto ancianos (más de 65 años); estos porcentajes son del 28, 32 y 40, respectivamente, en Marcera de Abajo.

En la tabla siguiente figura el porcentaje de personas que prefieren la playa a la hora de disfrutar de sus vacaciones veraniegas y su distribución, por grupos de edad en cada una de las dos localidades:

Edad	Marcera de Arriba	Marcera de Abajo
18-30	40	35
30-65	36	32
>65	14	18

a) ¿Qué porcentaje de personas con mayoría de edad vive en Marcera de Arriba, son jóvenes y no tienen la playa como preferencia a la hora de elegir sus vacaciones?

b) Hállese el porcentaje de personas de toda la comarca de Marcerón que prefieren la playa a la hora de disfrutar de sus vacaciones.

SOLUCIÓN

Dados los sucesos, R, B, J, D, A y P, *vivir en Marcera de Arriba*, *vivir en Marcera de Abajo*, *ser joven*, *ser adulto*, *ser anciano* y *preferir la playa*, respectivamente, el enunciado proporciona las siguientes probabilidades:

$$p(R) = 0{,}65 \text{ y } p(B) = 0{,}35,$$

junto con las probabilidades condicionadas correspondientes a cada intervalo de edad en cada una de las dos localidades:

$$p(J/R) = 0{,}3,\ p(D/R) = 0{,}25 \text{ y } p(A/R) = 0{,}45,$$

para Marcera de Arriba, y

$$p(J/B) = 0{,}28,\ p(D/B) = 0{,}32 \text{ y } p(A/B) = 0{,}4,$$

para Marcera de Abajo.

Por último, para cada una de las localidades y dentro de cada grupo de edad, se dispone de la probabilidad de preferir la playa a la hora de disfrutar de las vacaciones:

$$p(P/JR) = 0{,}4,\ p(P/DR) = 0{,}36 \text{ y } p(P/AR) = 0{,}14;$$

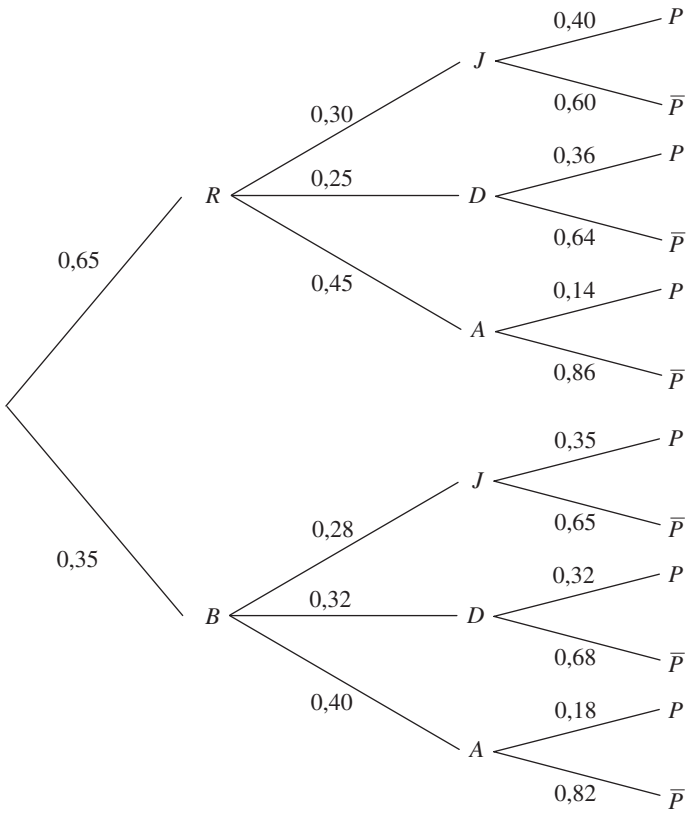
$$p(P/JB) = 0{,}35,\ p(P/DB) = 0{,}32 \text{ y } p(P/AB) = 0{,}18.$$

a) La probabilidad pedida se halla por aplicación de la regla de la multiplicación. En efecto,

$$p(RJ\overline{P}) = p(R) \cdot p(J/R) \cdot p(\overline{P}/JR) = 0{,}65 \cdot 0{,}30 \cdot 0{,}6 = 0{,}117.$$

Por tanto, el 11,7 por ciento de la población mayor de edad de Marcerón vive en Marcera de Arriba, es joven y no tiene como preferencia la playa.

b) El siguiente diagrama de árbol ayudará a calcular esta probabilidad:



En efecto, teniendo en cuenta las seis ramas del árbol anterior que corresponden a preferir la playa a la hora de disfrutar de las vacaciones, resulta:

$$p(P) = p(R) \cdot p(J/R) \cdot p(P/JR) + p(R) \cdot p(D/R) \cdot p(P/DR) + p(R) \cdot p(A/R) \cdot p(P/AR) +$$

$$+ p(B) \cdot p(J/B) \cdot p(P/JB) + p(B) \cdot p(D/B) \cdot p(P/DB) + p(B) \cdot p(A/B) \cdot p(P/AB).$$

Sustituyendo, se tiene que

$$p(P) = 0{,}65 \cdot 0{,}30 \cdot 0{,}40 + 0{,}65 \cdot 0{,}25 \cdot 0{,}36 + 0{,}65 \cdot 0{,}45 \cdot 0{,}14 +$$

$$+ 0{,}35 \cdot 0{,}28 \cdot 0{,}35 + 0{,}35 \cdot 0{,}32 \cdot 0{,}32 + 0{,}35 \cdot 0{,}40 \cdot 0{,}18 = 0{,}273,$$

es decir, el 27,3 por ciento de los habitantes mayores de edad de Marcerón prefieren la playa a la hora de disfrutar de sus vacaciones.

Puede el lector resolver el apartado ***a)*** con ayuda del diagrama de árbol.

6.36 En el II Encuentro sobre Economía de la Educación, se ha dividido el trabajo en dos áreas: una dedicada a la gestión y evaluación de la educación, A, y otra, B, a la financiación de la educación. En el área A está prevista la participación de 30 asistentes,

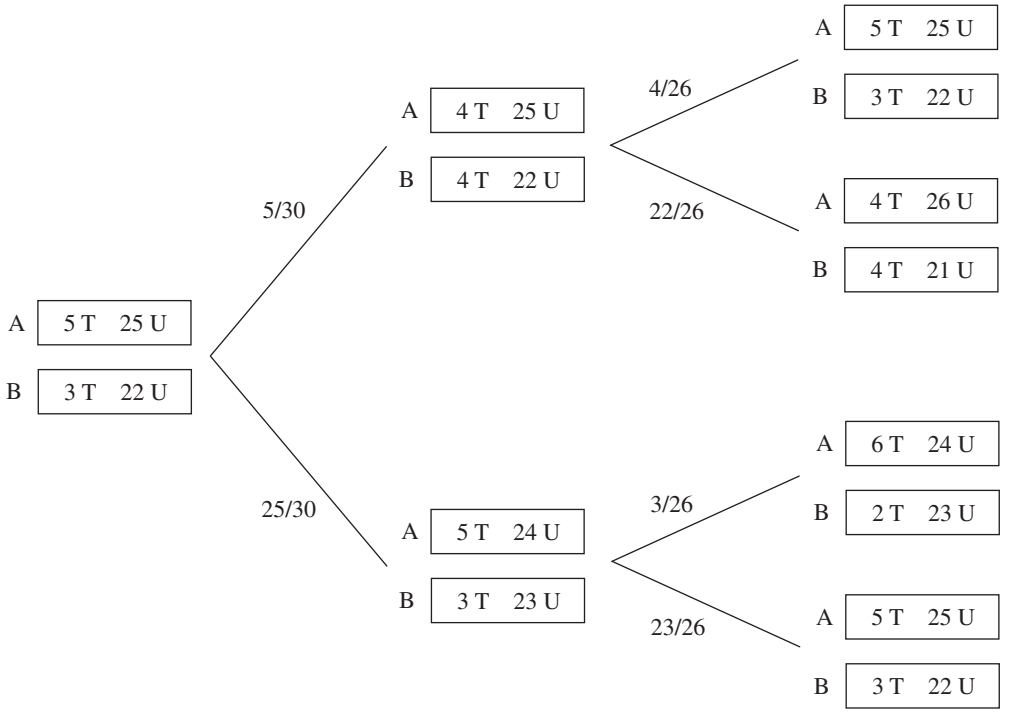
de los cuales 5 son técnicos de gestión y el resto profesores universitarios. En el área B, hay 25 participantes, de los cuales 3 son técnicos y el resto profesores.

La secretaría del encuentro ha cometido un error en la confección de la relación de asistentes; así, uno de los congresistas que debería estar incluido en la lista del área A, ha sido incorporado a la segunda, correspondiente al área B. Advertido el fallo, se le comunica a la azafata encargada de distribuir en las salas —una por área— a los congresistas, la cual toma al azar un congresista de la segunda sala y lo ubica en la primera.

Iniciado el encuentro, se elige aleatoriamente un congresista de la primera sala para que modere la sesión. ¿Cuál es la probabilidad de que sea un técnico de gestión?

SOLUCIÓN

Las diferentes etapas de este experimento aleatorio quedan descritas en el siguiente diagrama de árbol. Como puede observar el lector, en cada rama se considera la posibilidad de que el congresista que *se mueve* —bien en las listas confeccionadas en una primera etapa, bien físicamente en una segunda etapa—, de un área de trabajo a otra sea o no un técnico de gestión.



En una primera etapa, se incluye un congresista del área A en la lista del área B. La probabilidad de que *el congresista incluido sea un técnico* es igual 5/30, puesto que son 5 los técnicos del área A de un total de 30.

En la segunda etapa, si fue un técnico el que se «traspapeló» de la lista del área A a la del área B, la probabilidad de que *el congresista cambiado de área de trabajo por la azafata vuelva a ser un técnico* es ahora igual a 4/26, ya que el área B tiene, después del error en los listados, 4 técnicos de un total de 26 congresistas. Por el contrario, si fue un profesor el que vio cambiado su nombre de lista, la probabilidad de que *la azafata cambie de área a un técnico* es 3/26, ya que seguirán siendo 3 los técnicos de la lista B, de un total de 26 congresistas.

Para calcular la probabilidad del suceso *T*, *la persona elegida de la sala A es técnico*, han de considerarse las diferentes situaciones, esto es, las cuatro ramas del diagrama de árbol:

$$p(T) = \frac{5}{30} \cdot \frac{4}{26} \cdot \frac{5}{30} + \frac{5}{30} \cdot \frac{22}{26} \cdot \frac{4}{30} + \frac{25}{30} \cdot \frac{3}{26} \cdot \frac{6}{30} + \frac{25}{30} \cdot \frac{23}{26} \cdot \frac{5}{30} = 0{,}1652.$$

Como puede observar el lector, la última fracción de cada uno de los cuatro sumandos anteriores corresponde a la probabilidad de que *el congresista elegido en la primera sala para moderar la sesión sea técnico*, condicionada por todos los sucesos que en cada una de las ramas se han ido produciendo sucesivamente.

6.37 El 60 por ciento de las pólizas que suscribe al mes una compañía corresponde a seguros de vida. Se eligen 10 pólizas al azar. ¿Cuál es la probabilidad de que exactamente 3 de ellas correspondan a seguros de vida?

SOLUCIÓN

Puede suceder, por ejemplo, que las 3 primeras pólizas elegidas sean las que corresponden a seguros de vida. Dada la independencia de los experimentos, la probabilidad de esa situación es

$$p(S_1 \ldots S_3 \overline{S}_4 \ldots \overline{S}_{10}) = p(S_1) \cdot \ldots \cdot p(S_3) \cdot p(\overline{S}_4) \cdot \ldots \cdot p(\overline{S}_{10}),$$

donde S_i es el suceso *la i-ésima póliza elegida corresponde a un seguro de vida*, cuya probabilidad es igual a 0,6, para todo valor de *i*.

Se tiene, por tanto, sin más que sustituir, que

$$p(S_1 \ldots S_3 \overline{S}_4 \ldots \overline{S}_{10}) = 0{,}6^3 \cdot 0{,}4^7 = 0{,}00035.$$

Sin embargo, aunque ésta no es la única situación en la cual se presentan 3 seguros de vida y 7 seguros que no son de vida, todas ellas tienen idéntica probabilidad, por lo que resulta suficiente con contabilizar su número y multiplicar por la probabilidad de una de ellas para obtener la probabilidad pedida.

Ahora bien, ¿de cuántas formas distintas se pueden elegir 10 pólizas de manera que exactamente 3 de ellas sean seguros de vida? Pues de tantas como posibles grupos de 3 unidades —pólizas de seguros de vida— se puedan hacer de entre un total de 10 unidades —total de pólizas—.

Este número es el número de combinaciones de 10 elementos tomadas de 3 en 3, $\binom{10}{3}$, que también coincide con el número de permutaciones con repetición de 10 elementos de los cuales se repiten 3 por un lado y 7 por otro, según vimos en el problema **6.28**.

En consecuencia,

$$p(\text{exactamente 3 pólizas sean de seguros de vida}) = \binom{10}{3} \cdot 0{,}00035 = 0{,}042.$$

6.38 En la tabla siguiente aparecen las probabilidades correspondientes al número de errores que un alumno de una academia comete en el examen teórico para obtener el carnet de conducir.

N.º errores	0	1	2	3 o más
Probabilidad	0,8	0,1	0,05	0,05

Se sabe, además, que el 40 por ciento de los alumnos de la academia que cometen algún error tiene estudios universitarios.

a) ¿Qué porcentaje de alumnos no tiene titulación superior y comete algún error?

b) Se eligen al azar 20 alumnos de la academia. ¿Cuál es la probabilidad de que exactamente 3 de ellos no cometan ningún error al realizar su examen?

SOLUCIÓN

a) Considerando los sucesos, A, *un alumno comete algún error*, y, U, *un alumno tiene estudios universitarios*, el enunciado proporciona el siguiente dato:

$$p(U/A) = 0{,}4.$$

Para hallar la probabilidad pedida, basta aplicar la regla de la multiplicación:

$$p(\overline{U}A) = p(A) \cdot p(\overline{U}/A).$$

Ahora bien, por un lado,

$$p(\overline{U}/A) = 1 - p(U/A) = 1 - 0{,}4 = 0{,}6,$$

y, por otro lado,

$$p(A) = 0{,}1 + 0{,}05 + 0{,}05 = 0{,}2,$$

ya que el suceso A puede descomponerse en la unión de tres sucesos: *un alumno comete un error*, *un alumno comete dos errores* y *un alumno comete tres o más errores*.

Por consiguiente,

$$p(\bar{U}A) = 0{,}2 \cdot 0{,}6 = 0{,}12,$$

es decir, el 12 por ciento de los alumnos no tiene titulación superior y comete algún error.

b) La probabilidad de *que un alumno no cometa error* es, según se ve en la tabla, igual a 0,8. Para hallar la probabilidad de *que exactamente 3 alumnos de los 20 no cometan error* puede considerarse, en primera instancia, que fueran los 3 primeros alumnos elegidos los que no hicieran fallos en su examen. Así, llamando N_i al suceso, *el alumno i-ésimo no comete error*, la probabilidad de la situación descrita dada la independencia de los experimentos es

$$p(\bar{N}_1 \ldots \bar{N}_3 N_4 \ldots N_{20}) = 0{,}8^3\,(1 - 0{,}8)^{17}.$$

Pero, como ya se ha comentado en repetidas ocasiones, éste no es el único caso en que exactamente 3 alumnos no tienen fallos: cualquier orden en la elección de los 3 alumnos que no cometen errores es válida. Ahora bien, puesto que todas esas reordenaciones tienen la misma probabilidad, es suficiente con conocer su número y multiplicarlo por la probabilidad anteriormente calculada.

En definitiva,

$$p(\text{exactamente 3 alumnos no cometan errores}) = \binom{20}{3} \cdot 0{,}8^3\,(1 - 0{,}8)^{17} \approx 0.$$

6.39 Una empresa jienense, dedicada al embotellado y comercialización de aceite de girasol y de oliva, posee dos máquinas de envasado. La máquina A, que envasa el 60 por ciento del total, se dedica al aceite de girasol, mientras que la máquina B, embotella aceite de oliva.

El porcentaje de botellas de cristal utilizado en el envasado es del 20 por ciento para el aceite de girasol y del 70 por ciento para el de oliva. El resto de los envases son de plástico.

Para su comercialización, las botellas se empaquetan en cajas de 12 unidades. Un error en el proceso de empaquetado ha hecho que se mezclen en cada caja unidades de todo tipo. ¿Cuál es la probabilidad de que en una caja haya exactamente 4 botellas de plástico?

SOLUCIÓN

Se sabe que la probabilidad de que *la botella sea envasada por la máquina A* es 0,6, esto es,

$$p(A) = 0{,}6,$$

y, por tanto, será igual a 0,4 la probabilidad de que *sea embotellada en la máquina B*.

Además, según se desprende del enunciado, el 80 por ciento de los envases utilizados por la máquina A son de plástico, siendo de este material el 30 por ciento de las botellas que provienen de la máquina B, con lo cual, si P es el suceso *ser de plástico*,

$$p(P/A) = 0{,}8 \text{ y } p(P/B) = 0{,}3.$$

Con esta información es posible calcular la probabilidad de que una botella cualquiera sea de plástico utilizando el teorema de la probabilidad total. De este modo,

$$p(P) = p(P/A) \cdot p(A) + p(P/B) \cdot p(B) = 0{,}8 \cdot 0{,}6 + 0{,}3 \cdot 0{,}4 = 0{,}6.$$

Si en una caja de 12 botellas tiene que haber exactamente 4 botellas de plástico —y, en consecuencia, 8 de cristal—, han de considerarse todas las situaciones en las que este hecho puede presentarse. Así, puede ocurrir, por ejemplo, que las 4 botellas de plástico sean las 4 primeras de la caja, siendo la probabilidad en ese caso:

$$p(P_1 \dots P_4 \overline{P}_5 \dots \overline{P}_{12}) = p(P_1) \dots p(P_4) \cdot p(\overline{P}_5) \dots p(\overline{P}_{12}) = 0{,}6^4 \cdot 0{,}4^8,$$

donde P_i es el suceso *la i-ésima botella es de plástico.*

Como el número de situaciones con exactamente 4 botellas de plástico coincide con el número de posibles elecciones de 4 —para las botellas de plástico— entre un total de 12 botellas que componen la caja, esto es, $\binom{12}{4}$, la probabilidad pedida es

$$\binom{12}{4} \cdot 0{,}6^4 \cdot 0{,}4^8 = \frac{12!}{4! \cdot 8!} \cdot 0{,}6^4 \cdot 0{,}4^8 = 0{,}04204.$$