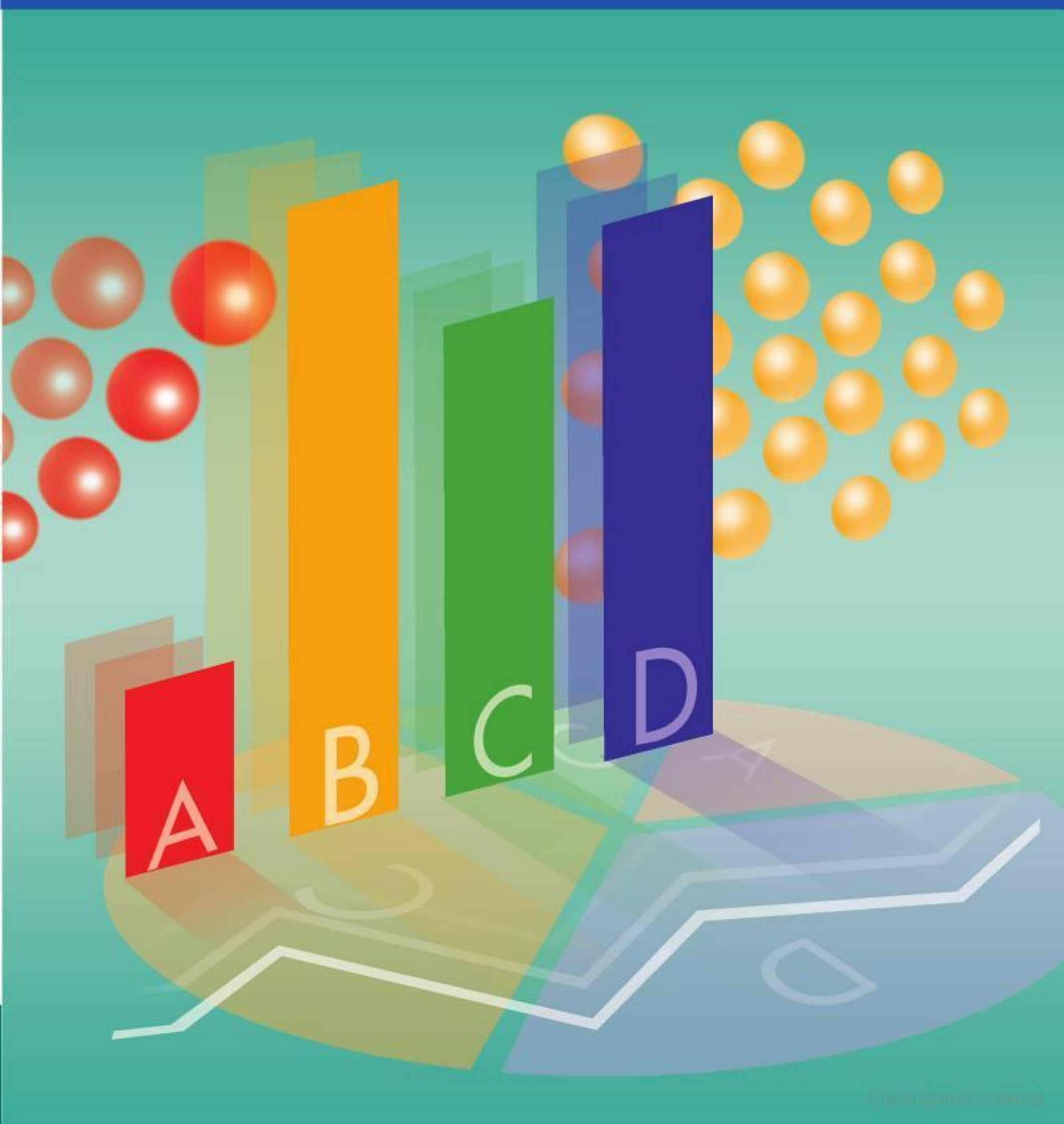
DELTA
PUBLICACIONES

CONCEPTOS BÁSICOS DE ESTADÍSTICA PARA CIENCIAS SOCIALES

José J. Cáceres Hernández



CONCEPTOS BÁSICOS DE ESTADÍSTICA PARA CIENCIAS SOCIALES

por JOSÉ JUAN CÁCERES HERNÁNDEZ

Editor gerente Fernando M. García Tomé
Diseño de cubierta Mizar Publicidad, S.L.
Preimpresión TXT. Servicios editoriales
Impresión Jacaryan, S.A.
Avda. Pedro Díez, 3. 28019 Madrid (España)

C/Luarca, 11
28230 Las Rozas (Madrid)
Dirección Web: www.deltapublicaciones.com

ISBN 84-96477-43-6
Depósito Legal

(0906-75)

Contenido

ESTADÍSTICA DESCRIPTIVA

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

A pesar de los trabajos de estos autores y aunque no es fácil señalar un momento concreto para el nacimiento de la estadística como ciencia, puede decirse que como tal disciplina inicia su camino en el siglo XX. La etapa de interacción progresiva entre análisis estadístico descriptivo y cálculo de probabilidades para conformar la estadística matemática, se consolida con los trabajos de Fisher de la década de los 20, en los que aparece formulada por primera vez una teoría general de la inferencia estadística paramétrica de base frecuencial. Fisher introdujo el paradigma de la verosimilitud a partir del cual efectuó interesantes aportaciones a la teoría de la estimación y el contraste de hipótesis. Neyman y Egon Pearson, a finales de la misma década, suministraron una teoría sistemática de contraste de hipótesis que terminaba de configurar las líneas básicas de la inferencia clásica como conjunto organizado de métodos. Por su parte, la inferencia bayesiana, representada, entre otros, por de Finetti y Savage, aparece como aproximación moderna a partir de finales de los años 40 e impulsa nuevos enfoques. Las ideas de Fisher, Neyman y Pearson, así como el redescubrimiento de los principios bayesianos, fueron la base de la teoría de la decisión, cuyo origen puede ubicarse en el trabajo de Wald en 1950.

A partir de mediados del siglo XX, debe destacarse el papel de la teoría de la información, que ha conducido al desarrollo de los ordenadores. En el terreno de la estadística, la revolución informática ha otorgado la posibilidad de manejar gran cantidad de información, lo que ha impulsado la aparición de diferentes procedimientos para el análisis de datos, cuya filosofía se basa en la búsqueda de patrones coherentes en los propios datos como fuente activa de información y no como base pasiva para contrastar hipótesis sugeridas por el investigador. Y aunque la propia ausencia de perspectiva histórica dificulta la descripción adecuada del avance científico de la estadística en las últimas décadas, es innegable que dos de los campos más potenciados por el progreso informático han sido los dedicados a los procesos estocásticos —con el desarrollo del moderno análisis de series temporales— y la denominada inteligencia artificial —cuyas técnicas pueden resultar útiles en el ámbito de las ciencias sociales, especialmente para predecir comportamientos evolutivos—.

1.2 UNA APROXIMACIÓN CONCEPTUAL. LAS FASES DEL PROCESO ESTADÍSTICO

Si se entiende por estadística un conjunto de métodos para tratar la información, será conveniente precisar qué se estudia y cómo se estudia, es decir, habrá que intentar aproximarse a su objeto material y su objeto formal. Dos de los elementos que caracterizan a la estadística son: 1) información acerca de un colectivo o universo, lo que constituye su objeto material; 2) un modo propio de razonamiento, el método estadístico, lo que constituye su objeto formal. Para entender la naturaleza de estos elementos es preciso introducir algunas nociones previas.

Los fenómenos que son objeto material de la estadística pueden clasificarse en causales o deterministas y aleatorios o estocásticos. En el primer caso, la observación proporciona una información cierta sobre el fenómeno estudiado, mientras que en el segundo existe un elemento de incertidumbre sobre las características reales del fenómeno bajo estudio a partir de las observaciones disponibles. Esta clasificación puede emplearse

para señalar que, en el ejercicio del análisis estadístico, cabe distinguir dos situaciones. Una, en la que se dispone de toda la información sobre el fenómeno o población bajo estudio, en cuyo caso, el objetivo puede ser extraer características de esa información que permitan obtener un conocimiento más fácilmente asimilable que el derivado del mero registro de observaciones (análisis descriptivo). Otra, en la que la información disponible se limita a la observación de una muestra de la población, y entonces, la finalidad de los métodos estadísticos puede consistir en extraer de esa información parcial, un conocimiento incierto sobre el conjunto de la población, pero con un grado de incertidumbre medido con ayuda de la teoría de la probabilidad (análisis inferencial). Cuando los fenómenos aleatorios que son objeto de tratamiento estadístico, dependen no sólo del azar, sino de posiciones o estrategias humanas, pueden incluirse como objeto material de la teoría estadística de la decisión, que constituye una herramienta formal diseñada para satisfacer una aspiración que fue el origen histórico del interés por la recopilación de datos: la obtención de información para orientar la toma de decisiones. En resumen, el objetivo de la estadística es describir, inferir y decidir. Y, dada la generalidad de su objeto material, puede considerarse sobre todo un método, lo que explica su carácter instrumental al servicio de las demás disciplinas.

La estadística descriptiva no pretende explicar, sino al contrario, separar lo esencial, resumir y medir con los medios apropiados un fenómeno colectivo que escapa por su extensión, diversidad e inconstancia a la comprensión directa e individual. Aunque a veces el fin de una investigación estadística es la mera descripción de los hechos, generalmente se trata sólo de la primera fase, el trabajo preliminar para la inferencia.

Por supuesto, el salto hacia la inferencia requiere necesariamente el paso por el cálculo de probabilidades, que supone un enfoque diferente al propio de la estadística descriptiva, en la medida en que sustituye el tratamiento de unas determinadas observaciones por la formulación de un modelo sobre los resultados posibles de un experimento aleatorio que permite evaluar las posibilidades de ocurrencia de éstos antes de observarlos. Ahora bien, el concepto de probabilidad no es único y, de hecho, por lo general, se distinguen cuatro tipos de probabilidad: clásica, planteada por Laplace; frecuencial, definida por von Mises; lógica, sostenida, entre otros, por Keynes; y subjetiva, defendida por autores como de Finetti, Ramsey o Savage. Ninguna de estas concepciones es plenamente satisfactoria ni es adecuada para todos los propósitos. Pero cualquiera de ellas es compatible con los axiomas que se precisan para construir el modelo matemático que permite tratar situaciones que encierran incertidumbre y, por ello, hace posible el salto hacia la inferencia estadística. El objetivo inferencial es el siguiente. A partir de los resultados obtenidos del análisis de una muestra de la población y utilizando como instrumento el cálculo de probabilidades, la inferencia estadística se encarga de generalizar estas leyes, es decir, infiere o estima las leyes generales del comportamiento de la población.

Las características poblacionales que interesa estudiar suelen formularse en términos de determinados parámetros —que se desea estimar o sobre los cuales se desea contrastar alguna hipótesis—, asumiendo como supuesto inicial que la variable que recoge la magnitud en cuestión sigue una determinada distribución conocida (métodos paramétricos); o bien, se trata de examinar características muy generales de la distribución de

dicha variable sin establecer ningún supuesto sobre la distribución de la misma (métodos no paramétricos).

Por otro lado, en función de la información considerada relevante, pueden distinguirse tres aproximaciones básicas: la inferencia clásica, que sólo admite la información muestral como base para el análisis, la inferencia bayesiana, cuyos representantes apuestan por la interacción entre información inicial probabilística de naturaleza subjetiva e información muestral, y la teoría de la decisión, que incorpora como información adicional la valoración que se hace de cada uno de los posibles resultados asociados a la toma de decisiones sobre aquellos aspectos en torno a los que se efectúa el análisis estadístico.

En la inferencia clásica el objetivo es formular inferencias inductivas que reduzcan la incertidumbre sobre algún parámetro a través de dos procedimientos básicos: la estimación y el contraste de hipótesis. Por su parte, en el enfoque bayesiano, la solución al problema inferencial consiste en proporcionar una distribución de probabilidad sobre el parámetro en cuestión. Mientras que en la teoría de la decisión se formula explícitamente un problema de elección entre diferentes cursos de acción en función de sus consecuencias, que dependen del estado de la naturaleza desconocido en el que se desarrolla dicha acción. El problema de decisión estadística consiste, entonces, en obtener una regla que permita elegir la acción óptima de acuerdo con algún criterio.

Ninguno de estos planteamientos está libre de crítica y quizás lo más acertado es concluir que la relevancia de uno u otro depende de las circunstancias particulares y la disponibilidad de distintos tipos de información. Pero, en cualquier caso, debería buscarse la reconciliación de las diferentes aproximaciones mostrándolas como distintas facetas de un objetivo común: el tratamiento de unas observaciones de acuerdo con una lógica y método propios: la metodología estadística. En este texto, los temas correspondientes a la inferencia estadística se abordarán desde la perspectiva clásica.

1.3 ESTADÍSTICA Y CIENCIAS SOCIALES

La estadística es actualmente una disciplina independiente de gran importancia, pero, dada la generalidad de su objeto material, puede considerarse sobre todo un método, ya que su objeto formal induce una metodología científica que puede ser utilizada por la totalidad de las ciencias empíricas, incorporándose como una parte más del objeto formal de éstas cuando los elementos o entes estudiados por ella sean de naturaleza incierta o aleatoria. Cualquiera que sea el origen de los datos que maneja, la estadística utiliza los mismos métodos y conceptos, lo que explica su carácter instrumental al servicio de las demás disciplinas. De ahí que la estadística haya recibido la denominación de tecnología del método científico o, de modo más ilustrativo, se haya dicho que la estadística es la «ancilla scientiarum», o esclava de las ciencias. Esta interpretación, que cuenta ya con una larga tradición, se ha mostrado especialmente cierta en el amplio campo de las ciencias sociales.

El papel actual de la estadística en la metodología científica ha sido posible gracias al cambio de dirección en la interpretación de las conclusiones inferidas por el razonamiento inductivo estadístico acaecido a principios del siglo XX, cuando se asume que este razonamiento puede hacerse preciso cuantificando la incertidumbre implicada

en las conclusiones inductivas; sin que, en ningún caso, deba interpretarse que esa pretendida precisión llegue más allá de lo que permite la naturaleza incierta de la inducción.

La teoría matemática de la probabilidad, en cuanto a disciplina matemática, no es susceptible de contrastación empírica: sus conclusiones teóricas, como las de todo sistema deductivo, son formalmente válidas siempre que estén lógicamente deducidas de los axiomas. Lo mismo puede decirse de la estadística cuando se identifique con esquemas probabilísticos deductivos; ahora bien, como ciencia inductiva, sus conclusiones no evitan de un modo tajante las dudas que se ciernen sobre la validez de la inducción y la contrastación empírica del conocimiento teórico elaborado. Y aunque la inferencia puede, de algún modo, ser todavía interpretada como un argumento de apoyo a la inducción, las conclusiones estadísticas no son por completo objetivas y exentas de crítica, sino que al contrario los elementos subjetivos y los diferentes enfoques metodológicos son enriquecedores.

Los métodos estadísticos han sido de gran valor en las ciencias sociales, dado que en situaciones caracterizadas por la presencia de un rango de variación en las observaciones a menudo notable y un número de observaciones frecuentemente limitado, sólo el análisis estadístico puede dar una estimación cuantitativa de la relevancia de los descubrimientos. Dada la complejidad propia de las ciencias sociales, éstas requieren algo más que simple deducción lógica y necesitan tratar adecuadamente una masa de información difícil de interpretar o explotar en el grado en que se precisa sin el recurso a métodos estadísticos.

Ahora bien, no toda la información sobre una realidad social se presta a un análisis estadístico. De hecho, no sería bueno que los científicos sociales renunciasen a aprovechar aquella información que no es reducible a tablas estadísticas. Precisamente la complejidad del objeto de estudio de los científicos sociales y la relativa escasez de información cuantificable que a veces presentan sus fuentes son razones que se invocan para discutir el interés que, para el estudio del comportamiento social, puedan presentar los métodos estadísticos, y para preferir un análisis no formal. Lo cierto es que, mientras que el análisis matemático es la gimnasia de la lógica deductiva, las ciencias sociales deben superar en múltiples ocasiones el rígido marco que impone dicha lógica, hasta el punto que resulta necesario sustituir el razonamiento deductivo puro por el establecimiento de criterios razonables desde el punto de vista del sentido común. Pero de aquí no debe deducirse que se está predicando el desamparo metodológico, sino que, por el contrario, se hace imprescindible la ayuda que herramientas como la estadística pueden dispensar como base para tales criterios, especialmente si se considera el elemento aleatorio. Cada vez es más necesario para el estudiante y estudioso de las ciencias sociales poseer unos conocimientos básicos y rigurosos sobre el contenido y alcance de la estadística, que le permitan comprender y evaluar apropiadamente esa realidad social que se presenta abrumadoramente cuantificada.

Desde una perspectiva racionalista, la ciencia se construye utilizando el método deductivo-matemático, en el que la realidad está esquematizada en modelos abstractos y la experimentación conceptual permite llegar a nuevas conclusiones. Pero las ciencias sociales se basan en la observación de fenómenos que están sometidos a las leyes

del azar, lo que exige acudir a modelos estocásticos y, por tanto, más que deducciones lógicas, se precisan leyes estadísticas.

Las ciencias sociales no disponen de laboratorios con condiciones controladas que permitan describir las leyes que guían el comportamiento de los individuos en todas sus vertientes. Por ello, la experimentación conceptual es de gran importancia en los modelos de comportamiento social donde no resulta fácil experimentar con hombres o instituciones. Pero también es necesario acudir a la observación, en la que inevitablemente participa el azar. En este sentido, la estadística proporciona un método que reemplaza el aislamiento de laboratorios, imposible de realizar en el campo de los fenómenos sociales, y ayuda a decidir sobre una base científica: la de la teoría de la probabilidad. El conocimiento sobre los comportamientos sociales tiene que basarse en la recogida continua de información, la propuesta de hipótesis de comportamiento a partir de lo observado y el contraste estadístico de dichas hipótesis, ya que uno de los rasgos más característicos de la sociedad es su carácter mutante.

Además, las ciencias sociales son frecuentemente acusadas de ser ciencias inexactas e incompletas a la hora de predecir comportamientos individuales, debido a la multiplicidad de factores que influyen en ellos. Generalmente, sin embargo, el comportamiento agregado de una población, resultante de la interacción entre diversos agentes, no es tan impredecible como los comportamientos individuales. La estadística proporciona una adecuada explicación a esta paradoja. En ello radica una de las principales utilidades de esta disciplina como herramienta para el análisis de los fenómenos sociales.

Y, por supuesto, además de su contribución al examen de la validez del conocimiento, la metodología estadística juega también un destacado papel en la construcción de los hechos que sustentan dicho conocimiento. La información estadística no sólo resulta necesaria para reflejar la realidad, sino que incluso es un instrumento para crearla. De hecho, existen ejemplos que muestran las interacciones entre la medida estadística y los procedimientos institucionales establecidos para identificar y codificar los objetos. Las herramientas estadísticas pueden ser conectadas a diferentes retóricas con el apoyo de varias construcciones intelectuales, sociales o políticas. Por tanto, no hay una forma simple de hacer que los números hablen y de usarlos como una base para el argumento. De modo que ciertos datos estadísticos como la inflación, el desempleo, el déficit público se convierten en fetiches aceptados como dogmas de fe que justifican las decisiones políticas. Y no es extraño que cifras básicas para entender la situación social de un país, sean utilizadas para exhibir logros políticos, aunque sea necesario acudir a un manejo poco riguroso de la información estadística.

Téngase en cuenta que la estadística permite la conexión entre conocimiento y poder, y que las decisiones dependen en muchos casos de la descripción estadística. Por ejemplo, la inflación no es un mero registro mental de subidas de precios, sino que su valor conlleva esencialmente repartos de rentas y salarios y, por tanto, se convierte en un elemento clave para la distribución del ingreso. En este sentido, la fijación de los incrementos salariales y otras rentas contratadas, de acuerdo con el índice del coste de la vida, es precisamente uno de los instrumentos fundamentales que pueden compensar la limitada influencia social en una economía de mercado. Por ello, la elección de las ponderaciones que se le conceden a los diferentes componentes del cesto de la compra adquiere una importancia que no es posible subrayar suficientemente. De igual

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

1.5 FUENTES DE OBTENCIÓN DE DATOS DE INTERÉS SOCIAL

Aunque finalmente la información disponible no dé pie a un análisis más sofisticado por medio de alguno de los procedimientos propios de la ciencia estadística, las fuentes de tales datos constituyen un recurso indispensable para adquirir conciencia de la realidad que se pretende estudiar o en la que se tienen que tomar decisiones. El interés por la recopilación de datos se pone de manifiesto por el hecho de que todos los países disponen de oficinas que se ocupan de recoger toda clase de información, que puede clasificarse en: a) datos estadísticos puros, es decir, información primaria, formada por encuestas (de carácter coyuntural o estructural) y censos, b) datos de valor añadido, es decir, información elaborada; y c) metainformación, es decir, información sobre la información. Múltiples datos son recopilados por la administración general del Estado y las Comunidades Autónomas españolas, así como por la Unión Europea, a través de las entidades correspondientes. Desde el punto de vista de la estadística oficial, es importante la labor desarrollada por el INE (España) o EUROSTAT (Unión Europea).

Pero no sólo los organismos públicos territoriales y supranacionales, sino también muchas organizaciones privadas y, en general, cualquier comunidad de individuos, parecen también impelidos por esta necesidad. A pesar de que la información oficial es cada vez más completa y desagregada, siempre existen ámbitos específicos en los que la información no ha sido registrada. De ahí que sea necesario utilizar mecanismos que permitan al investigador obtener por sí mismo la información que precisa para un estudio concreto. En este sentido, se puede distinguir entre métodos que permiten obtener información cualitativa (observación, experimentación, entrevista, proyección, ...) y otros que aportan información cuantitativa (encuesta personal, encuesta en establecimientos, encuesta telefónica, encuesta en medios de transporte, ...).

EJERCICIOS

1.1. Clasifique las siguientes características de acuerdo con su naturaleza cualitativa o cuantitativa y la escala de medida que puede utilizarse.

(a) Nivel de estudios de una persona.
(b) Porcentaje de parados en una población.
(c) Calidad de un producto.
(d) Número de hermanos de un individuo.
(e) Situación laboral de un individuo.
(f) Temperatura ambiente (en grados centígrados).
(g) Municipio de residencia de un individuo.
(h) Sueldo en euros de un trabajador.

1.2. Los tomates comercializados se clasifican de menor a mayor calibre en la siguiente escala: *MMM* , *MM* , *M* , *G* . Si se recodifica el atributo en una escala de uno a cuatro, se está midiendo ahora en una escala:

(a) nominal,
(b) ordinal,
(c) de intervalo,
(d) de razón.

1.3. ¿Cuáles de las siguientes variables son de naturaleza discreta?

(a) Peso exacto de una persona (kg).
(b) Tiempo de espera en una consulta médica.
(c) Rentabilidad anual de un activo financiero.
(d) Número de miembros de una unidad familiar.
(e) Proporción de trabajadores satisfechos con su trabajo en un grupo de 10.
(f) Tiempo que tarda un inmigrante en regularizar su situación.
(g) Número de habitaciones de una vivienda.
(h) Superficie de una vivienda en metros cuadrados.

2

Variable estadística unidimensional

Toda la estadística descriptiva está basada en el registro de observaciones de una magnitud para un conjunto de individuos sin que las conclusiones obtenidas superen el ámbito de lo observado. En este capítulo se introducen las nociones de variable estadística y distribución de frecuencias como mecanismo básico para expresar toda la información disponible sobre una magnitud de forma ordenada y sistemática, de modo que a partir de ella pueda efectuarse una adecuada descripción de los datos obtenidos. El análisis de los atributos se abordará más adelante.

2.1 CONCEPTO DE VARIABLE ESTADÍSTICA

Formalmente, una variable estadística es una función que asigna valores a la característica de la población analizada. Una variable estadística unidimensional X es un conjunto de observaciones de una magnitud para N individuos. En general, una variable estadística unidimensional X se denotará por

$$X:\left\{x_1^N,\ldots,x_N^N\right\}:\left\{x_k^N\right\}_{k=1,\ldots,N},$$

donde x_k^N representa el valor observado de la magnitud X para el k -ésimo individuo de la población.

La magnitud estudiada puede ser la edad de los N individuos de una población. Si se define X : "edad en años de los individuos de la población", entonces X es una variable estadística unidimensional. Si para cada individuo de la población en cuestión, se desea conocer su edad y su nivel anual de ingresos, será necesario superar el ámbito unidimensional. Si se define ahora Y : "nivel anual de ingresos en miles de euros de los individuos de la población", se tiene que (X,Y), es decir, el conjunto de pares de observaciones que indican la edad y nivel de ingresos de cada uno de los N individuos,

es una variable estadística bidimensional. En este capítulo se examinarán las variables estadísticas unidimensionales.

Ejemplo 2.1 Suponga que se han registrado las edades en años de 10 alumnos de una clase y han resultado ser las siguientes: $18,19,20,18,19,20,21,18,22,18$. Si se define la variable estadística X: "edad en años de los alumnos de la clase", resulta que $X:\{18,19,20,18,19,20,21,18,22,18\}$.

2.2 DISTRIBUCIÓN DE FRECUENCIAS DE UNA VARIABLE ESTADÍSTICA UNIDIMENSIONAL

Toda la información sobre determinada magnitud que contiene la variable estadística X puede recogerse de forma sistemática a través de la denominada distribución de frecuencias.

2.2.1. Frecuencias absolutas, relativas y acumuladas

A partir del conjunto de N observaciones registradas para los individuos de la población, es posible identificar el conjunto de valores distintos y ordenarlos de menor a mayor. Si existen n valores distintos, este conjunto puede denotarse por $\{x_1,\ldots,x_n\}:\{x_i\}_{i=1,\ldots,n}$. Y si cada valor x_i se ha observado n_i veces, entonces el conjunto de información puede representarse como

$$\{(x_1,n_1),\ldots,(x_n,n_n)\}:\{(x_i,n_i)\}_{i=1,\ldots,n},$$

es decir, indicando la frecuencia absoluta n_i con que se repite cada uno de los valores x_i. Este conjunto de pares define la distribución de frecuencias absolutas. Nótese que

$$\sum_{i=1}^{n} n_i = n_1 + \ldots + n_n = N .$$

En muchas ocasiones es interesante conocer, por ejemplo, cuántos individuos de la población poseen, en relación con la característica estudiada, un valor menor o igual que x_i. Para ello es útil definir las frecuencias absolutas acumuladas N_i, tales que

$$N_i = \sum_{j=1}^{i} n_j = n_1 + \ldots + n_i .$$

La distribución de frecuencias absolutas acumuladas puede definirse entonces como

$$\{(x_1,N_1),\ldots,(x_n,N_n)\}:\{(x_i,N_i)\}_{i=1,\ldots,n}.$$

Por supuesto, $N_n = N$.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

N individuos de la población original, $X:\{x_1^N,...,x_N^N\}:\{x_k^N\}_{k=1,...,N}$, se definen los intervalos $\{I_1,...,I_n\}:\{I_i\}_{i=1,...,n}$, tales que

$$I_i:(L_{i-1},L_i],$$

siendo L_{i-1} el extremo inferior del intervalo, que no pertenece a éste, y L_i el extremo superior de dicho intervalo, que sí se incluye en el mismo. Entonces, si se denota por n_i al número de valores observados que pertenecen al intervalo I_i, la distribución de frecuencias absolutas de la magnitud X puede expresarse como

$$\{(I_1,n_1),...,(I_n,n_n)\}:\{(I_i,n_i)\}_{i=1,...,n}.$$

Y de forma análoga a la explicada para el caso de distribuciones no agrupadas, pueden obtenerse las distribuciones de frecuencias acumuladas, frecuencias relativas y frecuencias relativas acumuladas.

La amplitud a_i del intervalo I_i viene dada por

$$a_i = L_i - L_{i-1}$$

y, en general, es recomendable que los diferentes intervalos tengan la misma amplitud, de modo que la distribución de frecuencias agrupadas no introduzca distorsiones significativas con respecto a la distribución de frecuencias no agrupadas. En cualquier caso, este criterio no siempre es válido para los intervalos extremos. Para el caso de intervalos cuya amplitud no sea constante, la frecuencia correspondiente al intervalo I_i puede ser engañosa y conviene comparar los intervalos en términos de su densidad de frecuencia d_i, definida como

$$d_i = \frac{n_i}{a_i},$$

que mide el número de valores registrados por unidad de longitud del intervalo.

Por otra parte, la información que proporciona la distribución agrupada no permite saber con qué frecuencia se registra cada uno de los valores individuales que pertenecen a un intervalo determinado. Esta inevitable pérdida de información puede atenuarse incrementando el número de intervalos y sacrificando, a cambio, el grado de simplificación que aporta la distribución agrupada. De todas formas, suele asumirse el supuesto de que los valores se distribuyen uniformemente en el interior del intervalo. Es decir, se asume que dentro de un subintervalo contenido en un intervalo determinado existe la misma proporción de valores que representa la longitud del subintervalo con respecto a la longitud del intervalo que lo contiene. Y, a partir de este supuesto, se representa el intervalo a través de la denominada marca de clase, definida como el punto medio del intervalo. Así, si se denota la marca de clase del intervalo I_i por c_i, resulta que

$$c_i = \frac{L_i + L_{i-1}}{2}.$$

Ejemplo 2.3 Suponga que el grado de satisfacción con el estilo propio de vida puede medirse en una escala de 0 a 10 y que los valores declarados en este sentido por un grupo de 30 personas se recogen en la siguiente tabla.

3	5	6	2	4	3	5	1	3	9
4	6	3	4	3	4	8	7	4	7
5	9	6	5	8	6	10	6	2	5

Entonces, si se denota la variable estadística anterior como X, la distribución de frecuencias de esta variable estadística viene dada por

$$\{(x_i,n_i)\}_{i=1,\ldots,10}:\{(1,1),(2,2),(3,5),(4,5),(5,5),(6,5),(7,2),(8,2),(9,2),(10,1)\}.$$

Y la distribución de frecuencias agrupada en intervalos definidos como $I_1:(0,2]$, $I_2:(2,4]$, $I_3:(4,6]$, $I_4:(6,8]$, $I_5:(8,10]$, es la que se representa a continuación.

$$\{(I_i,n_i)\}_{i=1,\ldots,5}:\{(I_1,3),(I_2,10),(I_3,10),(I_4,4),(I_5,3)\}.$$

2.3 REPRESENTACIONES GRÁFICAS DE VARIABLES ESTADÍSTICAS

Como se ha explicado, toda la información que una variable estadística proporciona sobre una determinada magnitud queda recogida en la distribución de frecuencias. Pero existen formas de representación gráfica de dicha información que permiten captarla de un modo más directo e intuitivo. Algunas de estas representaciones gráficas se explican a continuación.

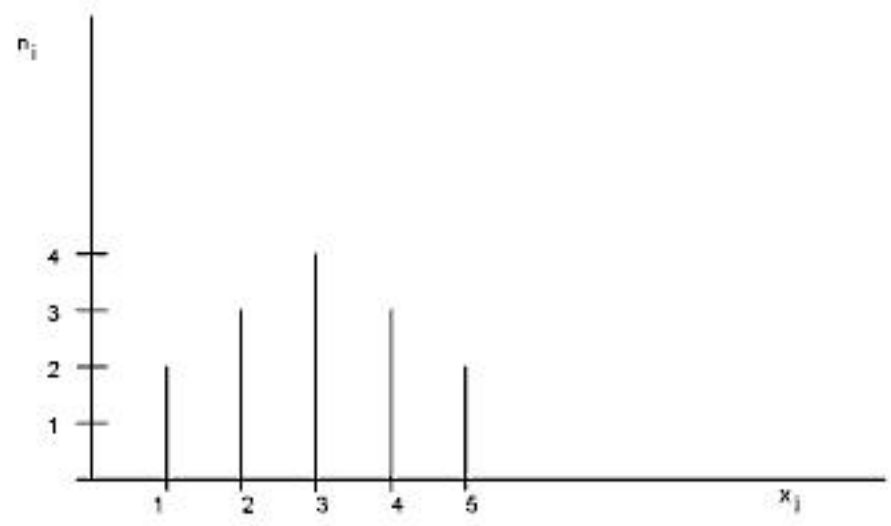
2.3.1. Distribuciones no agrupadas en intervalos

Sea la distribución de frecuencias $\{(x_i,n_i)\}_{i=1,\ldots,n}$.

(a) Diagrama de barras. En un sistema de coordenadas cartesianas en el plano, se representan los valores de la variable $\{x_i\}_{i=1,\ldots,n}$ en el eje de abscisas y sus frecuencias absolutas $\{n_i\}_{i=1,\ldots,n}$ en el eje de ordenadas. Si se han registrado cinco valores distintos y la distribución de frecuencias es tal que

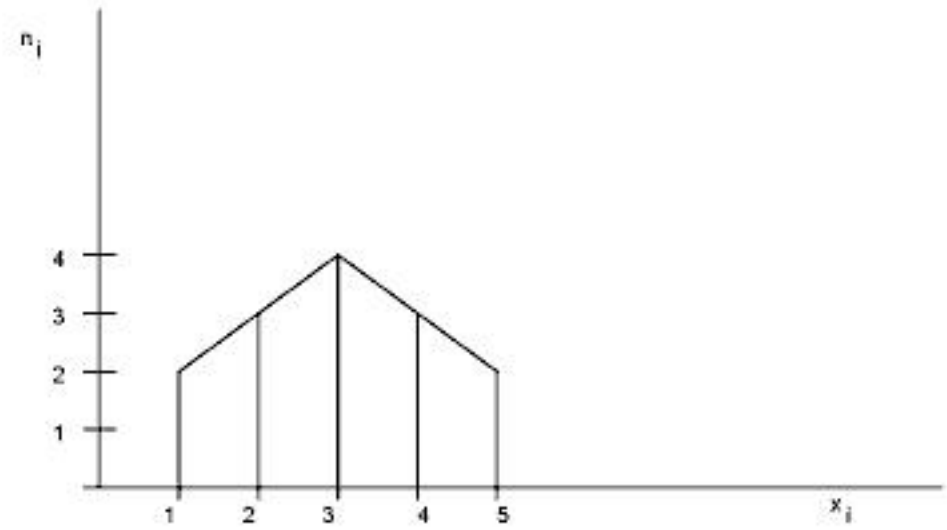
$$\{(x_i,n_i)\}_{i=1,\ldots,n}:\{(1,2),(2,3),(3,4),(4,3),(5,2)\},$$

el diagrama de barras es el que se indica.



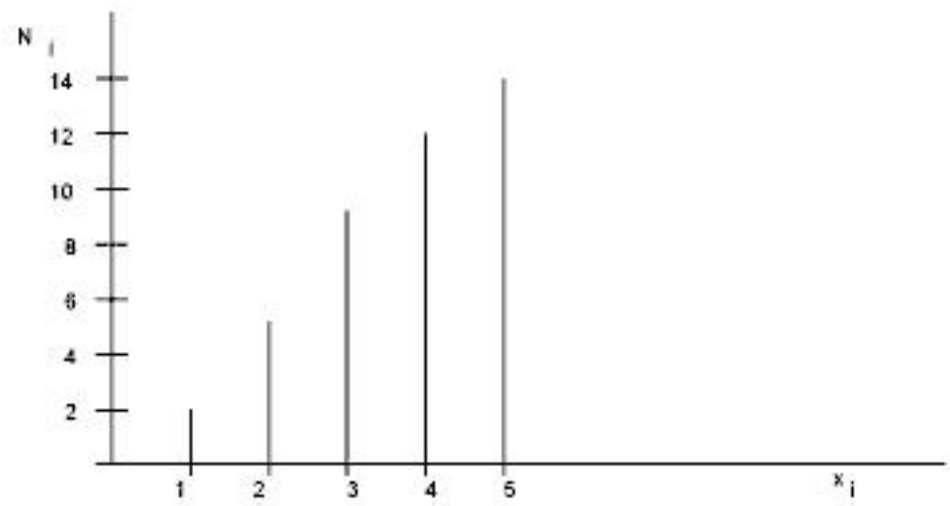
También se pueden utilizar las frecuencias relativas en lugar de las absolutas.

(b) Polígono de frecuencias. Se construye trazando una línea que une los vértices superiores de las líneas que aparecen en el diagrama de barras. Para el caso anterior, el polígono de frecuencias es la línea comentada y representada en el gráfico siguiente.

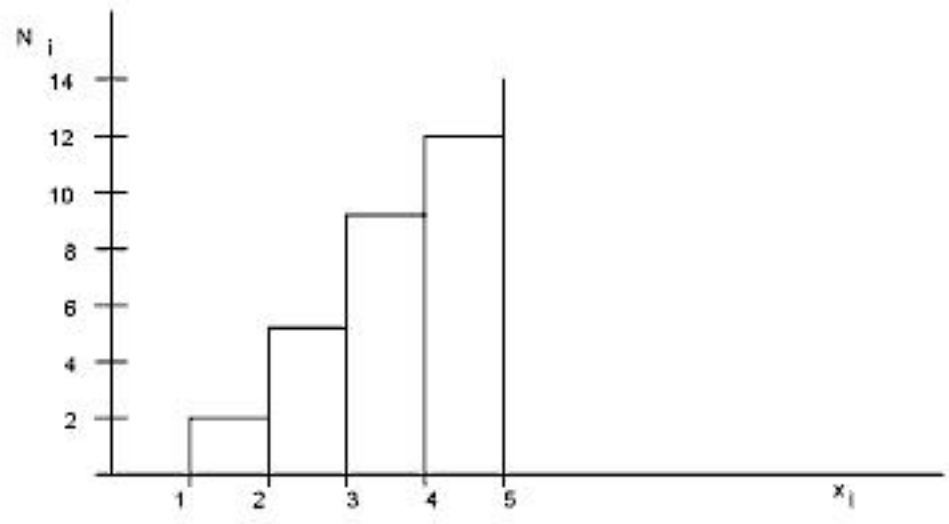


Como en el caso anterior, el gráfico proporciona información similar si se representan las frecuencias relativas.

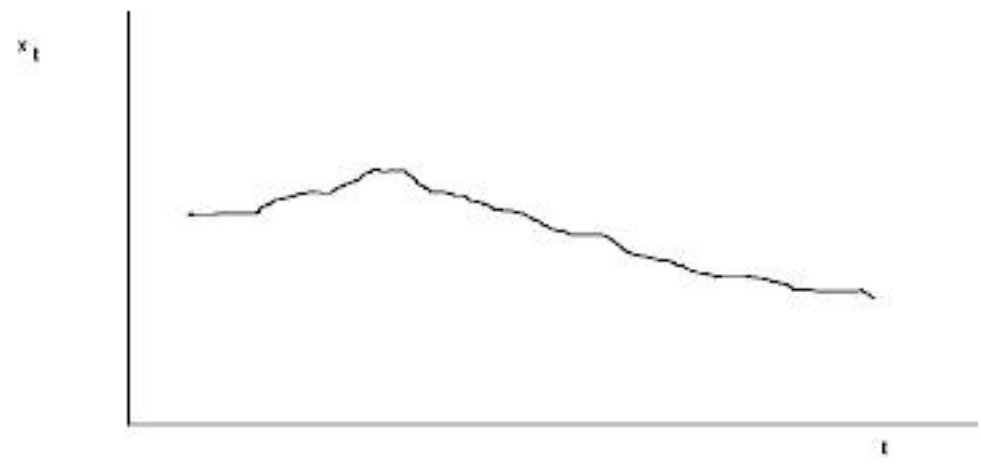
(c) Diagrama de frecuencias acumuladas. Se construye como el diagrama de barras, pero en el eje de ordenadas se representan las frecuencias absolutas acumuladas $\{N_i\}_{i=1,\ldots,n}$. Utilizando el mismo ejemplo, se obtiene el diagrama siguiente.



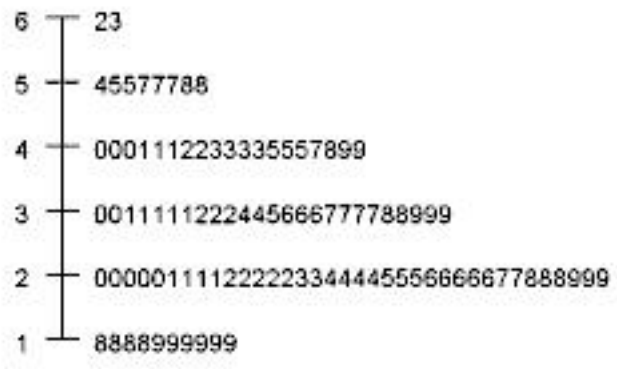
(d) Polígono de frecuencias acumuladas. Se construye trazando líneas horizontales desde cada vértice superior de las líneas que aparecen en el diagrama de frecuencias acumuladas hasta cortar con la línea vertical correspondiente al valor siguiente, de la manera que se indica.



(e) Gráficos de series temporales. Cuando los valores de la variable estadística corresponden a valores de una magnitud en diferentes instantes del tiempo, es útil representar estos valores en un sistema cartesiano. En este caso, conviene utilizar la distribución original, es decir, no ordenar los valores observados de acuerdo con su magnitud sino en función del instante del tiempo. Si los valores observados en los instantes del tiempo $\{1,...,N\}$ son, respectivamente, $\{x_1,...,x_N\}$, pueden indicarse los instantes del tiempo t en el eje de abscisas y los valores observados x_t en el eje de ordenadas. Por ejemplo, si la variable estadística X recoge la tasa de natalidad anual en un país europeo desde 1950 hasta 2000, y se representan los valores de dicha tasa frente al tiempo, se obtendrá un gráfico del tipo siguiente.



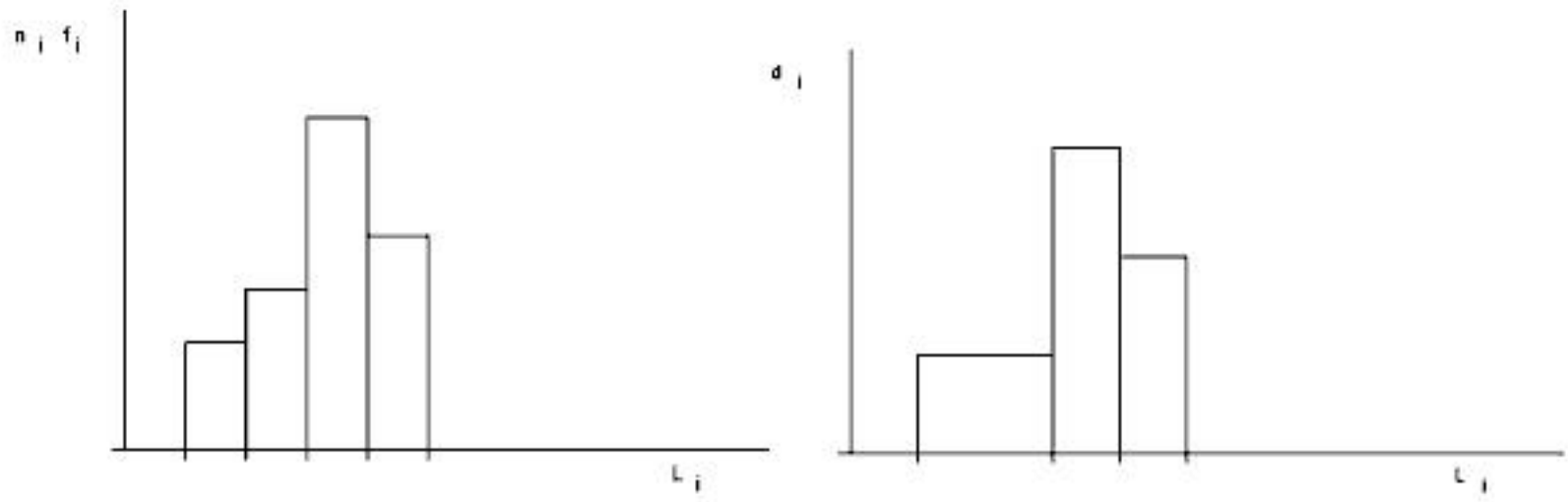
(f) Diagrama de tallos y hojas. Para conjuntos no muy numerosos de datos, la distribución de frecuencias puede expresarse en un gráfico de esta clase. Supóngase que la variable estadística X recoge la edad de los 100 trabajadores de una empresa, de modo que dicha variable es un conjunto de 100 valores ordenados de menor a mayor. Para construir el diagrama de tallos y hojas basta con escribir en una columna los primeros dígitos diferentes correspondientes a la edad de los trabajadores y en otra columna a la derecha indicar correlativamente el segundo dígito correspondiente a dichas edades. Se podría entonces obtener un gráfico como el siguiente.



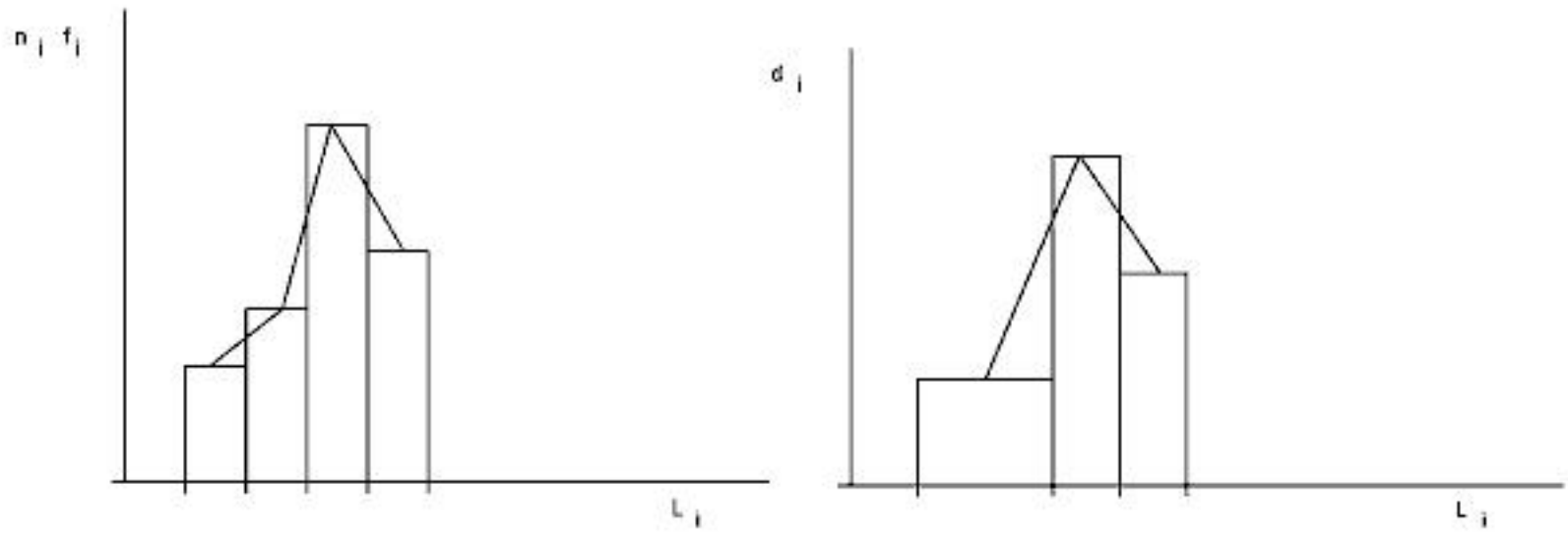
2.3.2. Distribuciones agrupadas en intervalos

Sea la distribución de frecuencias agrupadas en intervalos $\{(I_i, n_i)\}_{i=1,\ldots,n}$.

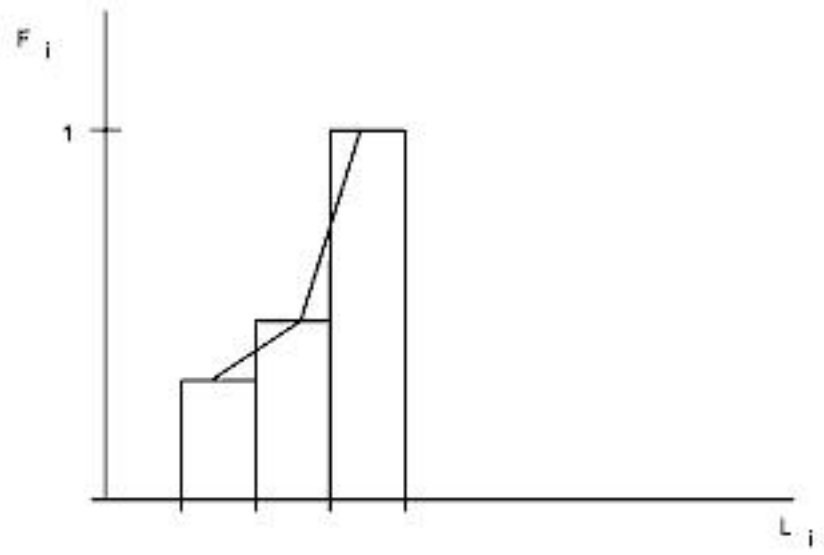
(a) Histograma. Se obtiene en un sistema de coordenadas cartesianas en el plano, representando en el eje de abscisas los extremos de cada intervalo y dibujando sobre cada uno de tales intervalos un rectángulo cuya área sea proporcional a la frecuencia absoluta o relativa del intervalo correspondiente. Si los intervalos son de amplitud constante, la altura de los rectángulos será también proporcional a dicha frecuencia y, de hecho, puede hacerse coincidir con ésta. Si la amplitud de los intervalos es variable, la altura de los rectángulos será proporcional a la densidad de frecuencia del intervalo. Es decir, los histogramas pueden tener las formas que se indican.



(b) Polígono de frecuencias. Se construye trazando una línea que une los centros de los segmentos superiores de los rectángulos dibujados en el histograma. Para los dos casos anteriores, los respectivos polígonos de frecuencias son las líneas que se superponen a los histogramas representados en los gráficos siguientes.



(c) Polígono de frecuencias acumuladas. Se obtiene del mismo modo que en el caso anterior, pero a partir de la representación de las frecuencias acumuladas. En el caso de intervalos de igual amplitud y utilizando las frecuencias relativas acumuladas, el polígono de frecuencias relativas acumuladas es la línea que une los puntos medios de la base superior de los rectángulos representados en el gráfico siguiente.



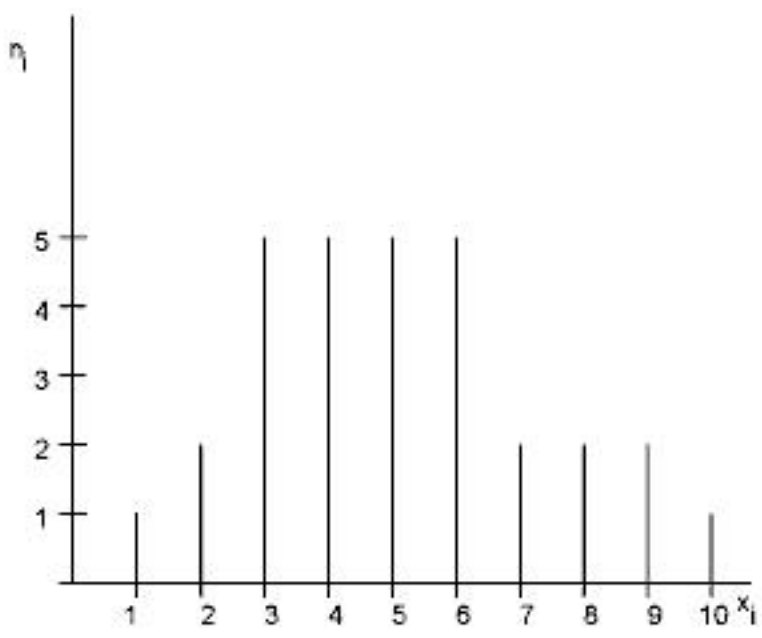
Ejemplo 2.4 Suponga de nuevo que el grado de satisfacción con el estilo propio de vida puede medirse en una escala de 0 a 10 y que los valores declarados en este sentido por un grupo de 30 personas se recogen en la siguiente tabla.

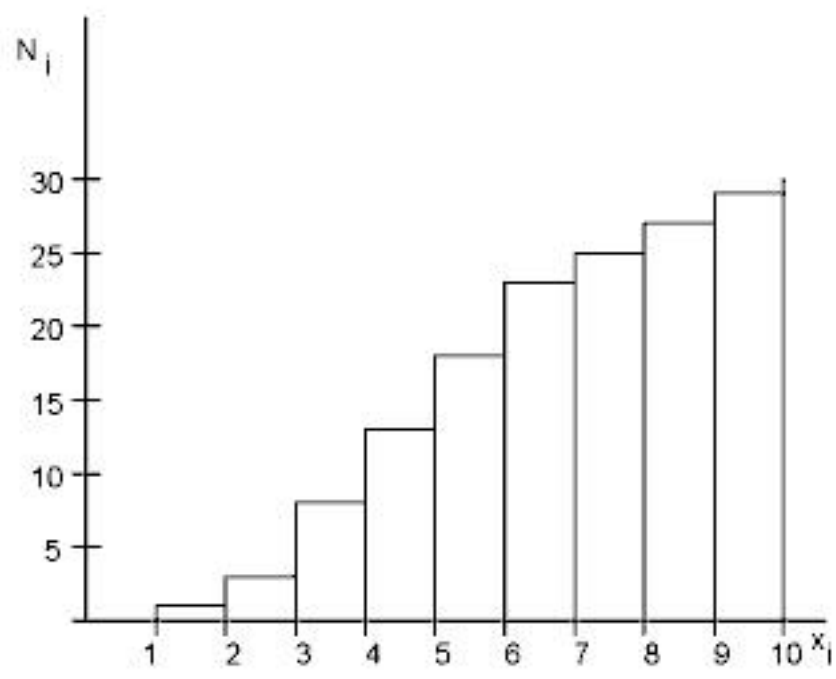
3	5	6	2	4	3	5	1	3	9
4	6	3	4	3	4	8	7	4	7
5	9	6	5	8	6	10	6	2	5

La distribución de frecuencias de esta variable estadística,

$$\{(x_i,n_i)\}_{i=1,\ldots,10}:\{(1,1),(2,2),(3,5),(4,5),(5,5),(6,5),(7,2),(8,2),(9,2),(10,1)\},$$

puede representarse mediante el diagrama de barras o el polígono de frecuencias acumuladas siguientes.

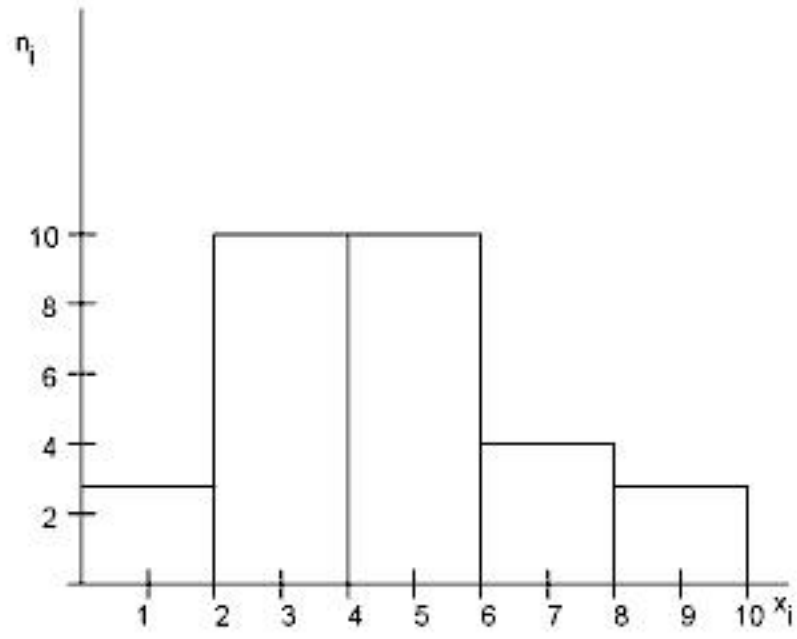


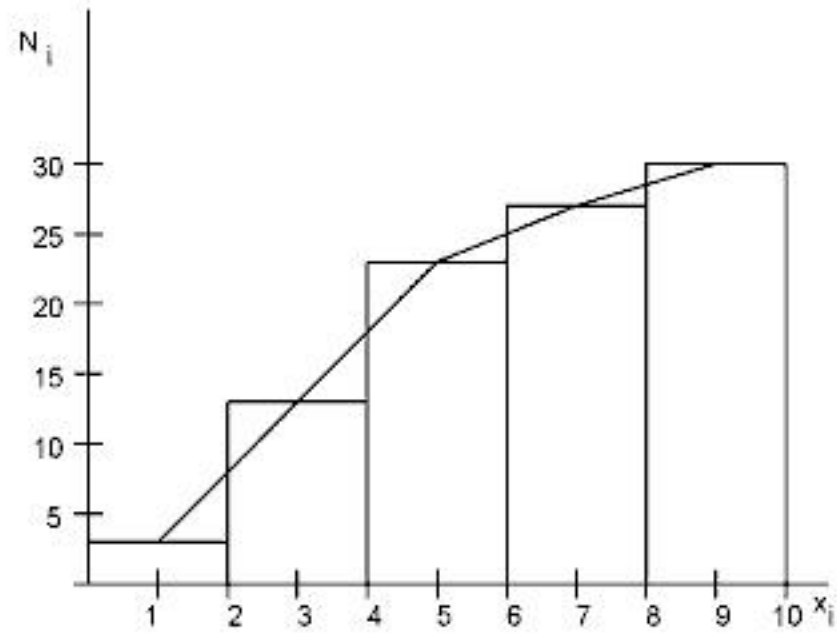


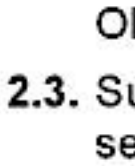
Mientras que la distribución de frecuencias agrupada en los intervalos $I_1:(0,2]$, $I_2:(2,4]$, $I_3:(4,6]$, $I_4:(6,8]$, $I_5:(8,10]$,

$$\{(I_i,n_i)\}_{i=1,\ldots,5}:\{(I_1,3),(I_2,10),(I_3,10),(I_4,4),(I_5,3)\},$$

puede representarse mediante un histograma o un polígono de frecuencias acumuladas, tal como se representa a continuación.







EJERCICIOS

2.1. Sea la variable X :"número de hijos de un conjunto de familias", cuya distribución de frecuencias es la que se muestra en la siguiente tabla. Suponga, además, que se sabe que el 35% de las familias tiene un hijo y que el 10% tiene cuatro hijos.

x_i	n_i	N_i	f_i	F_i
0		2		
1				
2	6			
3				
4				
N	20			

(a) Complete los datos que faltan en la tabla anterior.
(b) Obtenga el diagrama de barras, el polígono de frecuencias, el diagrama de frecuencias acumuladas y el polígono de frecuencias acumuladas.

2.2. Sea la variable X :"esperanza de vida de las mujeres de un conjunto de países", cuya distribución de frecuencias es la siguiente.

x_i	39	44	45	48	49	50	52	53	54	56	57	58	59	N
n_i	1	2	3	2	4	7	2	5	2	2	2	2	2	36

Obtenga el diagrama de tallos y hojas.

2.3. Suponga que los metros cuadrados de cada una de las 70 viviendas de una zona se recogen en la siguiente tabla.

70	71	53	91	66	68	54	78	150	75
99	138	87	100	88	61	87	103	95	108
105	66	97	136	119	65	96	88	200	100
115	90	78	93	185	120	92	205	95	68
75	120	143	106	106	86	110	66	80	135
96	117	84	76	45	100	85	89	87	72
93	118	75	87	140	82	100	140	78	175

(a) Construya la distribución de frecuencias de esta variable estadística y represéntela a través de un diagrama de barras y un polígono de frecuencias.
(b) Construya la distribución de frecuencias agrupada en intervalos definidos de la siguiente forma: $I_1:(40,70]$, $I_2:(70,100]$, $I_3:(100,130]$, $I_4:(130,160]$, $I_5:(160,190]$, $I_6:(190,220]$. Represente la distribución obtenida mediante un histograma, un polígono de frecuencias y un polígono de frecuencias acumuladas.
(c) Proponga otra agrupación en intervalos que pueda tener interés y represente el histograma, el polígono de frecuencias y el polígono de frecuencias acumuladas.

2.4. Suponga que el grado de satisfacción con el estilo propio de vida puede medirse en una escala de 0 a 10. Sea la variable X que recoge dicho grado de satisfacción para un grupo de personas, cuya distribución de frecuencias agrupadas en intervalos es la siguiente.

I_i	n_i
$(0,2]$	4
$(2,4]$	8
$(4,6]$	6
$(6,10]$	12

(a) Calcule la amplitud y la densidad de frecuencia de cada intervalo.
(b) Represente el histograma y el polígono de frecuencias.
(c) Proponga otra agrupación en intervalos que pueda tener interés y obtenga su distribución de frecuencias. Represente esta distribución mediante el polígono de frecuencias.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Ejemplo 3.1 Suponga una variable estadística que recoge las edades de 10 alumnos de una clase y, denotando esa variable por X, resulta que $X:\{18,18,18,18,19,19,20,20,21,22\}$. La distribución de frecuencias de esta variable es $\{(x_i,n_i)\}_{i=1,\ldots,5}:\{(18,4),(19,2),(20,2),(21,1),(22,1)\}$, de modo que

$$\overline{x}=m_1=\sum_{i=1}^{5}x_i\frac{n_i}{N}=19.3$$

y

$$m_2=\sum_{i=1}^{5}x_i^2\frac{n_i}{N}=374.3\,.$$

Por otra parte,

$$\mu_2=\sum_{i=1}^{5}(x_i-\overline{x})^2\frac{n_i}{N}=1.81$$

o bien,

$$\mu_2=m_2-m_1^2=1.81\,.$$

En el caso de distribuciones agrupadas en intervalos, los momentos respecto al origen pueden evaluarse utilizando las marcas de clase de los intervalos. Y una vez obtenida la media aritmética, los momentos centrales pueden calcularse promediando las diferencias entre las marcas de clase de los intervalos y la media aritmética previamente obtenida elevadas a la potencia correspondiente.

3.2 MEDIDAS DE POSICIÓN

Las medidas de posición constituyen puntos de referencia que permiten ubicar la situación de una variable estadística en la recta real y ofrecen, de este modo, una síntesis de toda la información contenida en la distribución de frecuencias.

3.2.1. Moda

Una de las medidas de posición es la moda, que identifica el valor o intervalo que más se repite. Formalmente, puede definirse la moda de la variable estadística X, que se denotará por Mo_X, como

$$Mo_X=x_j\,/\,f_j=\max_i\{f_i\}_{i=1,\ldots,n}\,.$$

Es decir, la moda o valor modal es el valor con mayor frecuencia relativa. Evidentemente, también se podría definir como el valor con mayor frecuencia absoluta, es decir, el valor que se ha observado un mayor número de veces. Desde este punto de vista, cabe distinguir entre distribuciones unimodales, en las que la máxima frecuencia corresponde a un solo valor, y distribuciones multimodales, en las que la máxima frecuencia se registra en varios valores distintos de la variable.

Ejemplo 3.2 Suponga de nuevo que la variable estadística $X:\{18,18,18,18,19,19,20,20,21,22\}$ recoge las edades de 10 alumnos de una clase. La distribución de frecuencias de esta variable es $\{(x_i,n_i)\}_{i=1,\dots,5}:\{(18,4),(19,2),(20,2),(21,1),(22,1)\}$ y se tiene que $Mo_X=18$.

En el caso de distribuciones agrupadas y si los intervalos son todos de la misma amplitud, el intervalo modal es aquél que posee mayor frecuencia absoluta o relativa. Y en el caso de distribuciones unimodales, puede asumirse que el valor modal será alguno de los valores que pertenecen a dicho intervalo, es decir,

$$Mo_X \in \left(L_{j-1},L_j\right] \Leftrightarrow f_j = \max_i \{f_i\}_{i=1,\dots,n}.$$

La moda pertenece al intervalo $\left(L_{j-1},L_j\right]$ si y sólo si a dicho intervalo corresponde la frecuencia relativa máxima. Y dentro del intervalo modal, el valor modal puede elegirse de acuerdo con uno de los dos criterios siguientes.

Puede considerarse que el valor modal estará más próximo a uno u otro extremo del intervalo modal en función de las frecuencias absolutas de los intervalos anterior y posterior a éste. Es decir,

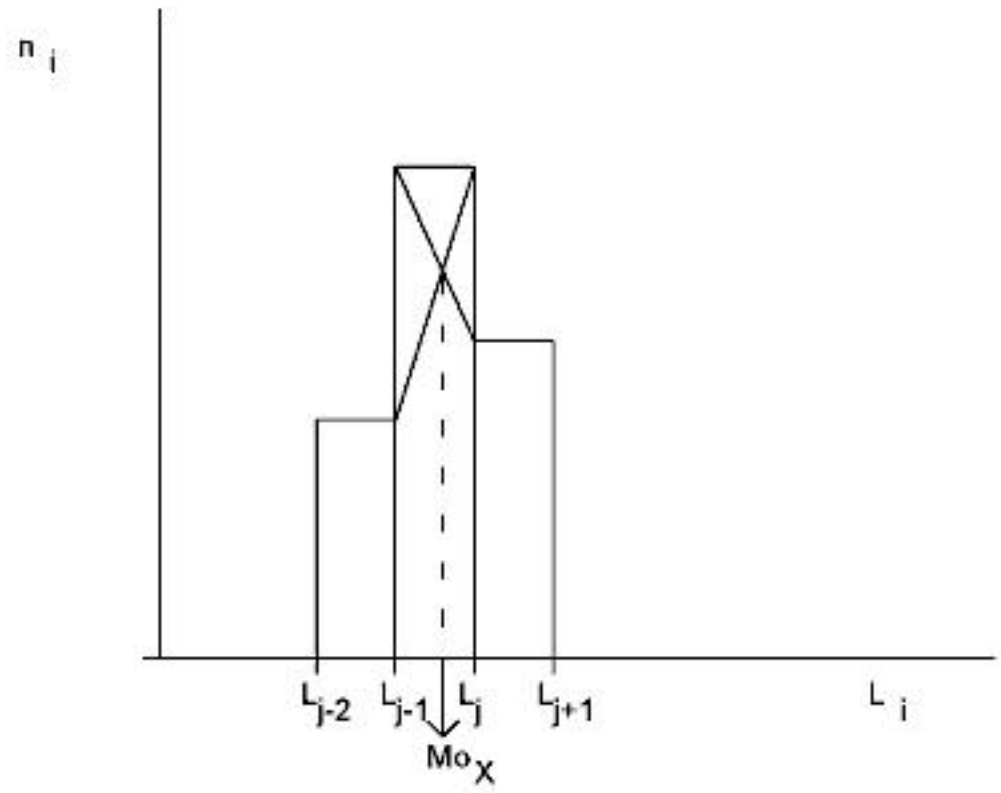
$$Mo_X = L_{j-1} + \frac{n_{j+1}}{n_{j+1}+n_{j-1}}a,$$

siendo a la amplitud de los intervalos. Nótese que si la frecuencia del intervalo anterior, n_{j-1}, coincide con la del intervalo posterior, n_{j+1}, entonces el valor modal se situará en el punto medio del intervalo modal. Si $n_{j-1}<n_{j+1}$, el valor modal estará más próximo al intervalo posterior; y si $n_{j-1}>n_{j+1}$, el valor modal estará más próximo al intervalo anterior.

También puede considerarse que el valor modal estará más próximo a uno u otro extremo del intervalo modal en función de las diferencias entre la frecuencia absoluta de dicho intervalo y las frecuencias absolutas de los intervalos contiguos. Es decir,

$$Mo_X = L_{j-1} + \frac{(n_j - n_{j-1})}{(n_j - n_{j-1})+(n_j - n_{j+1})}a,$$

siendo de nuevo a la amplitud de los intervalos. Nótese que, como ocurre con el criterio anterior, si la frecuencia del intervalo anterior, n_{j-1}, coincide con la del intervalo posterior, n_{j+1}, entonces el valor modal se situará en el punto medio del intervalo modal. Si $n_{j-1}<n_{j+1}$, el valor modal estará más próximo al intervalo posterior; y si $n_{j-1}>n_{j+1}$, el valor modal estará más próximo al intervalo anterior.



Ejemplo 3.3 Suponga que la variable estadística X recoge el número de miembros de 84 familias. La distribución de frecuencias absolutas agrupadas en los intervalos $I_1:(0.5,2.5]$, $I_2:(2.5,4.5]$, $I_3:(4.5,6.5]$, $I_4:(6.5,8.5]$, $I_5:(8.5,10.5]$, es

$$\{(I_i,n_i)\}_{i=1,\ldots,5}:\{(I_1,17),(I_2,40),(I_3,16),(I_4,8),(I_5,3)\}.$$

Por tanto, el intervalo modal es $I_2:(2.5,4.5]$. Y de acuerdo con los criterios propuestos, el valor modal es

$$Mo_X = L_1 + \frac{n_3}{n_3+n_1}a = 2.5+\frac{16}{16+17}2 = 3.4696,$$

o bien,

$$Mo_X = L_1 + \frac{(n_2-n_1)}{(n_2-n_1)+(n_2-n_3)}a = 2.5+\frac{23}{23+24}2 = 3.4786.$$

Dada la naturaleza discreta de la magnitud considerada, estos valores modales son poco informativos y resulta más útil la información que proporciona la identificación del intervalo modal.

Cuando los intervalos poseen amplitud variable, parece más lógico asumir que el valor modal pertenecerá a aquél intervalo con mayor densidad de frecuencia. Es decir,

$$Mo_X \in (L_{j-1},L_j] \Leftrightarrow d_j = \max_i\{d_i\}_{i=1,\ldots,n}.$$

La moda pertenece al intervalo $(L_{j-1},L_j]$ si y sólo si a dicho intervalo corresponde la densidad de frecuencia máxima. Y dentro del intervalo $(L_{j-1},L_j]$, el valor modal puede elegirse de acuerdo con uno de los dos criterios siguientes.

Puede considerarse que el valor modal estará más próximo a uno u otro extremo del intervalo $\left(L_{j-1},L_j\right]$ en función de las densidades de frecuencias de los intervalos anterior y posterior a éste. Es decir,

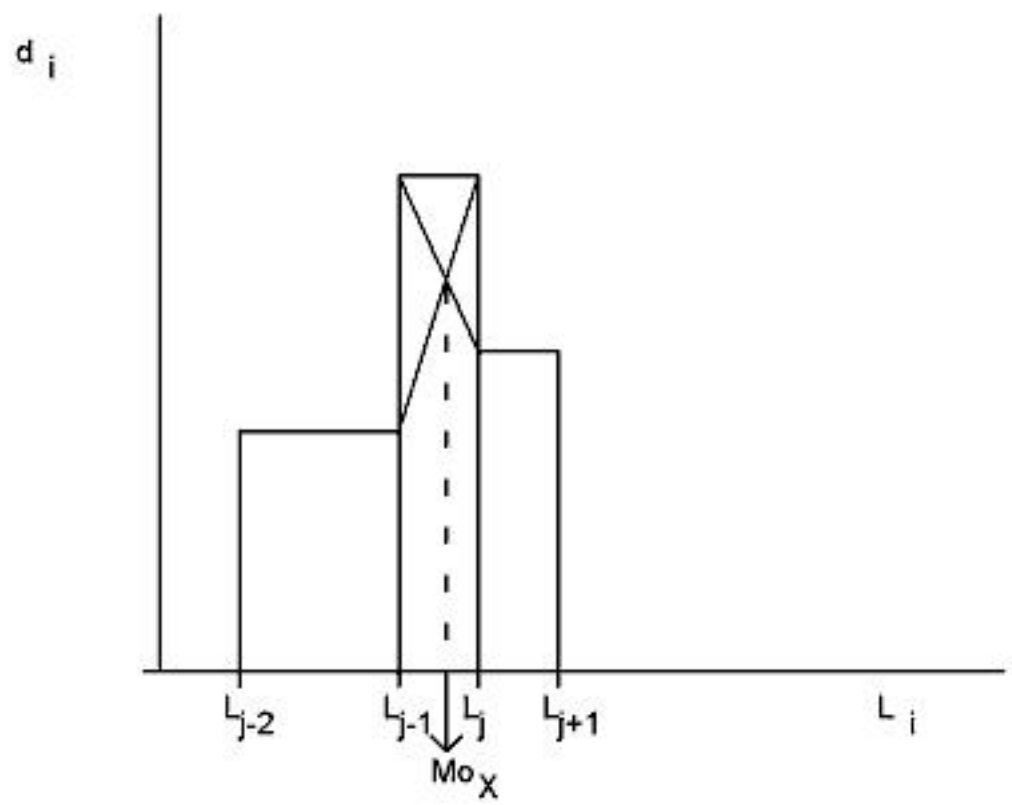
$$Mo_X = L_{j-1} + \frac{d_{j+1}}{d_{j+1}+d_{j-1}} a_j ,$$

siendo a_j la amplitud del intervalo $\left(L_{j-1},L_j\right]$. Nótese que si la densidad de frecuencia del intervalo anterior, d_{j-1}, coincide con la del intervalo posterior, d_{j+1}, entonces el valor modal se situará en el punto medio del intervalo $\left(L_{j-1},L_j\right]$. Si $d_{j-1} < d_{j+1}$, el valor modal estará más próximo al intervalo posterior; y si $d_{j-1} > d_{j+1}$, el valor modal estará más próximo al intervalo anterior.

También puede considerarse que el valor modal estará más próximo a uno u otro extremo del intervalo $\left(L_{j-1},L_j\right]$ en función de las diferencias entre la densidad de frecuencia de dicho intervalo y las densidades de frecuencia de los intervalos contiguos. Es decir,

$$Mo_X = L_{j-1} + \frac{(d_j - d_{j-1})}{(d_j - d_{j-1}) + (d_j - d_{j+1})} a_j ,$$

siendo de nuevo a_j la amplitud del intervalo $\left(L_{j-1},L_j\right]$. Nótese que, como ocurre con el criterio anterior, si la densidad de frecuencia del intervalo anterior, d_{j-1}, coincide con la del intervalo posterior, d_{j+1}, entonces el valor modal se situará en el punto medio del intervalo $\left(L_{j-1},L_j\right]$. Si $d_{j-1} < d_{j+1}$, el valor modal estará más próximo al intervalo posterior; y si $d_{j-1} > d_{j+1}$, el valor modal estará más próximo al intervalo anterior.



El valor modal puede ser engañoso, ya que no necesariamente indica dónde se ubica la mayoría de los valores de la distribución en su conjunto. Por ello, conviene buscar otros indicadores que señalen dónde se sitúa el *centro* de la distribución. Tales indicadores son las medidas de posición central.

3.2.2. Medidas de tendencia central: los promedios y la mediana

Las medidas de tendencia central por excelencia son los promedios y la mediana. Al grupo de los promedios pertenecen, entre otros, la media aritmética, la media geométrica y la media armónica. Aunque existen situaciones específicas en que uno de estos promedios puede ser más apropiado que otros, para buena parte de los fenómenos sociales la media aritmética es el promedio más apropiado y en determinados casos puede ser conveniente recurrir a la media geométrica. Estos dos son, de hecho, los únicos promedios que se explican a continuación.

Como ya se adelantó, la media aritmética de la variable estadística X, denotada por $\overline{x}$, es el momento respecto al origen de primer orden, es decir,

$$\overline{x} = \sum_{i=1}^{n} x_i \frac{n_i}{N}.$$

Por tanto, la media aritmética es un promedio de los valores de la variable. Pero no todos los valores intervienen en el promedio en el mismo grado o con la misma fuerza, sino que se trata de un promedio ponderado. El peso con que cada valor observado incide en el promedio evaluado viene dado por su frecuencia relativa. Así, puede interpretarse que cada valor observado tira hacia sí mismo del promedio con una fuerza proporcional a su frecuencia relativa. De este modo, la media aritmética es una especie de centro de gravedad de la distribución resultante de la compensación de tales fuerzas.

Ejemplo 3.4 Sea una variable estadística X tal que $X:\{1,2,3\}$. Entonces la distribución de frecuencias es unitaria y viene dada por $\{(x_i,n_i)\}_{i=1,\dots,3}:\{(1,1),(2,1),(3,1)\}$, de modo que

$$\overline{x} = \sum_{i=1}^{3} x_i \frac{1}{3} = 1\frac{1}{3} + 2\frac{1}{3} + 3\frac{1}{3} = 2.$$

Pero si la variable estadística es $X:\{1,2,3,1\}$, entonces

$$\{(x_i,n_i)\}_{i=1,\dots,3}:\{(1,2),(2,1),(3,1)\}$$

y el valor 1 tendrá más peso en el cálculo del promedio, de manera que el resultado se acercará más a este valor; concretamente, se tiene que

$$\overline{x} = 1\frac{2}{4} + 2\frac{1}{4} + 3\frac{1}{4} = \frac{7}{4}.$$

Gráficamente, los resultados obtenidos pueden observarse en los siguientes diagramas.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$\{(I_i, n_i)\}_{i=1,\ldots,6} : \{(I_1,11),(I_2,38),(I_3,12),(I_4,7),(I_5,2),(I_6,2)\}.$$

Por tanto, la media aritmética de esta variable puede calcularse como

$$\overline{x} = \sum_{i=1}^{6} c_i \frac{n_i}{72} = 97.0833 .$$

Por otra parte, se define la media geométrica de la variable estadística X, denotada por $\overline{x}_G$, como

$$\overline{x}_G = \sqrt[N]{\prod_{i=1}^{n} x_i^{n_i}} .$$

Y para el caso de distribuciones agrupadas en intervalos, se tiene que

$$\overline{x}_G = \sqrt[N]{\prod_{i=1}^{n} c_i^{n_i}} .$$

El uso de este promedio resulta especialmente apropiado cuando se desea promediar tasas de cambio de alguna magnitud, como refleja el ejemplo siguiente.

Ejemplo 3.6 Sea una variable estadística X que recoge las cifras anuales de parados de un país desde el año 1 hasta el año 10, es decir, $X : \{x_t\}_{t=1,\ldots,10}$. Si el número de parados (en miles de personas) en estos 10 años es el que figura en la siguiente tabla

t	1	2	3	4	5	6	7	8	9	10
x_t	1000	1100	1210	1340	1470	1620	1780	1950	2150	2358

entonces puede evaluarse la tasa media de crecimiento anual del número de parados. Nótese que la tasa de crecimiento anual, en tantos por uno, es

$$\frac{x_t - x_{t-1}}{x_{t-1}} = r_t ,$$

de modo que

t	1	2	3	4	5	6	7	8	9	10
r_t	-	0.1000	0.1000	0.1074	0.0970	0.1020	0.0988	0.0955	0.1026	0.0967

Por tanto,

$$x_t = (1 + r_t) x_{t-1}, \; t = 2,\ldots,10 ,$$

así que

$$x_2 = (1 + r_2) x_1 ,$$

$$x_3 = (1+r_3)x_2 = (1+r_3)(1+r_2)x_1,$$

$$x_4 = (1+r_4)x_3 = (1+r_4)(1+r_3)(1+r_2)x_1.$$

y así sucesivamente, resulta que

$$x_t = (1+r_t)\ldots(1+r_3)(1+r_2)x_1, \quad t = 2,\ldots,10,$$

de modo que

$$x_{10} = (1+r_{10})\ldots(1+r_3)(1+r_2)x_1.$$

Pues bien, la tasa media de crecimiento anual debe ser una magnitud r tal que

$$x_{10} = (1+r)^9 x_1.$$

Es decir,

$$(1+r)^9 = (1+r_{10})\ldots(1+r_3)(1+r_2),$$

lo que significa que el valor $1+r$ es la media geométrica de los valores $1+r_t$, es decir,

$$1+r = \sqrt[9]{(1+r_2)(1+r_3)\ldots(1+r_{10})},$$

de modo que la tasa media de crecimiento anual del número de parados es

$$r = \sqrt[9]{(1+r_2)(1+r_3)\ldots(1+r_{10})} - 1.$$

Por tanto,

$$r = 0.100$$

Nótese que

$$x_{10} = (1+r)^9 x_1 = (1+0.100)^9 1000 = 2358.$$

Otra medida de posición central es la mediana, que concede más peso a las frecuencias que a los valores en sí mismos y resulta, por tanto, menos sensible a los valores extremos. Si se observa el valor de una magnitud para un conjunto de N individuos y se disponen de menor a mayor los valores observados, se obtiene la variable estadística X tal que $X:\{x_1^N,\ldots,x_N^N\}$. Entonces, el valor mediano, que se denotará por Me_X, es aquél de los valores observados pertenecientes al conjunto $\{x_1^N,\ldots,x_N^N\}$ tal que exista el mismo número de observaciones a la izquierda y a la derecha de dicho valor. De acuerdo con esta definición, si el número de individuos es impar, está garantizado que la mediana será uno de los valores observados. Pero si el número de individuos es par, la mediana puede definirse como promedio de los dos valores centrales del conjunto $\{x_1^N,\ldots,x_N^N\}$.

Ejemplo 3.7 Suponga una variable estadística que recoge las edades de 11 alumnos de una clase y, denotando esa variable por X, resulta que $X:\{18,18,18,18,19,\mathbf{19},19,20,20,21,22\}$. En este caso la mediana es el valor 19 subrayado en negrita. Suponga ahora que la variable estadística X recoge las edades de 10 alumnos tales que $X:\{18,18,18,18,\mathbf{19},\mathbf{19},20,20,21,22\}$. En este caso la mediana es 19, que es el promedio de los dos valores centrales subrayados en negrita. Y si las edades observadas fueran tales que $X:\{18,18,18,19,\mathbf{19},\mathbf{20},20,20,21,22\}$, el valor mediano es 19.5, que no es un valor observado.

Sobre todo cuando el número de observaciones es elevado, la identificación del valor mediano es más inmediata utilizando la distribución de frecuencias. Sea $\{(x_i,n_i)\}_{i=1,\ldots,n}$ la distribución de frecuencias de la variable estadística X. Entonces el valor mediano es un valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ que verifica ciertas condiciones. Si no existe ningún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{1}{2}$, entonces

$$Me_X = x_j \Big/ x_j = min\left\{x_i \Big/ F_i > \frac{1}{2}\right\},$$

es decir, el valor mediano es el mínimo de los valores del conjunto $\{x_i\}_{i=1,\ldots,n}$ con frecuencia acumulada mayor que $\frac{1}{2}$. Por el contrario, si existe algún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{1}{2}$, entonces

$$Me_X = \frac{x_j + x_{j+1}}{2}.$$

En el caso de distribuciones agrupadas, el valor mediano pertenecerá al primero de los intervalos cuya frecuencia relativa acumulada sea mayor o igual que $\frac{1}{2}$. Es decir,

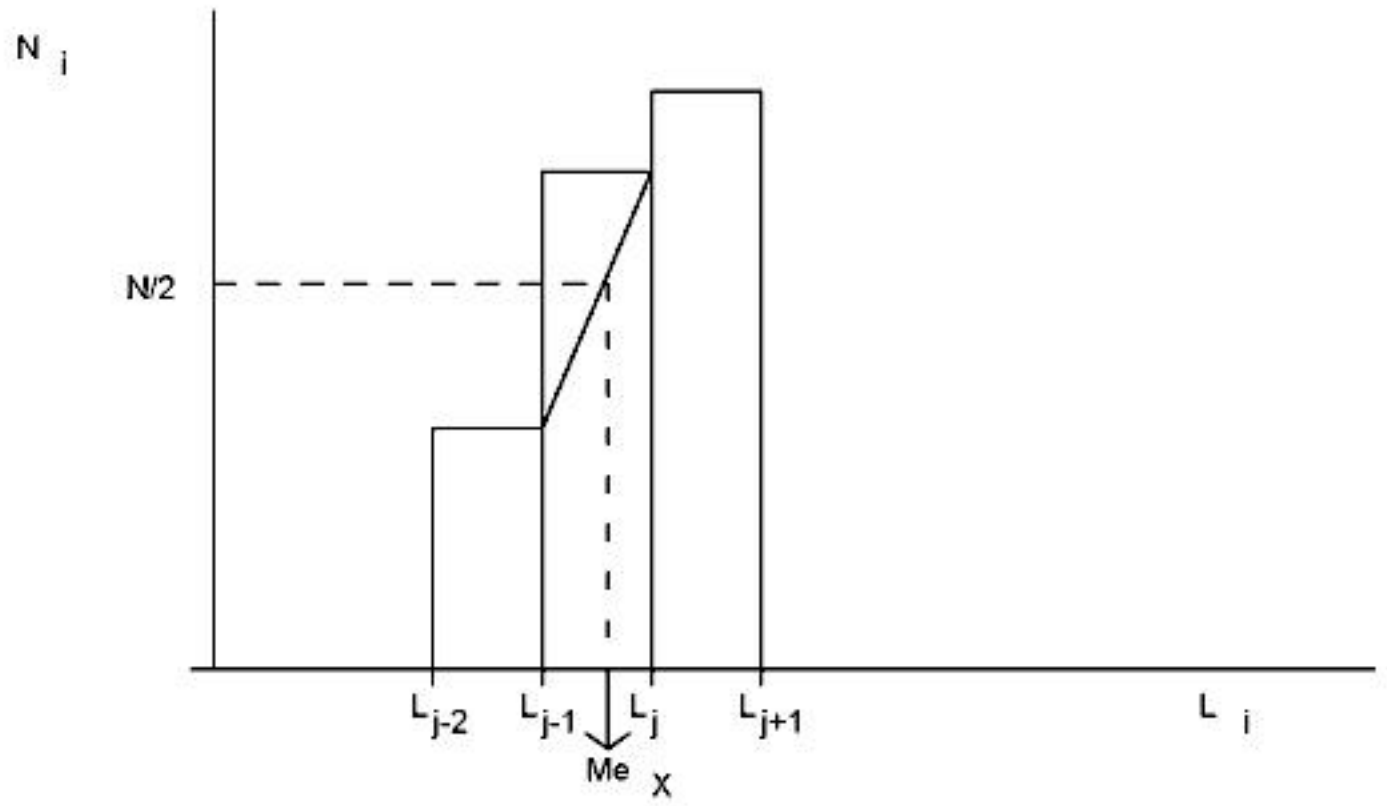
$$Me_X \in \left(L_{j-1},L_j\right] \Leftrightarrow j = min\left\{i \Big/ F_i \geq \frac{1}{2}\right\}.$$

Y el valor mediano dentro de dicho intervalo puede obtenerse asumiendo que la frecuencia registrada en un segmento del intervalo es proporcional a la longitud del segmento, de modo que la frecuencia acumulada en el interior del intervalo va creciendo

linealmente desde el extremo inferior al extremo superior. De acuerdo con este supuesto, el valor mediano dentro del intervalo $\left(L_{j-1}, L_j\right]$ puede definirse como

$$Me_X = L_{j-1} + \frac{\frac{N}{2} - N_{j-1}}{n_j} a_j ,$$

siendo a_j la amplitud del intervalo $\left(L_{j-1}, L_j\right]$. Este criterio se aprecia con claridad en el siguiente gráfico.



Ejemplo 3.8 Suponga de nuevo que la variable estadística X recoge el número de miembros de 84 familias. La distribución de frecuencias absolutas de esta variable es $\{(x_i, n_i)\}_{i=1,\ldots,10} : \{(1,6),(2,11),(3,21),(4,19),(5,8),(6,8),(7,7),(8,1),(9,2),(10,1)\}$, mientras que la correspondiente distribución de frecuencias relativas acumuladas viene dada por el conjunto

$\{(x_i, F_i)\}_{i=1,\ldots,10}$:

$\{(1,0.07),(2,0.202),(3,0.45),(4,0.68),(5,0.77),(6,0.87),(7,0.952),(8,0.96),(9,0.988),(10,1)\}$

Por tanto, el valor mediano es

$$Me_X = x_j \Big/ x_j = min\left\{ x_i \Big/ F_i > \frac{1}{2} \right\} = 4 .$$

Agrupando los valores observados en los intervalos, $I_1 : (0.5, 2.5]$, $I_2 : (2.5, 4.5]$, $I_3 : (4.5, 6.5]$, $I_4 : (6.5, 8.5]$, $I_5 : (8.5, 10.5]$, las distribuciones de frecuencias absolutas, acumuladas y relativas acumuladas son, respectivamente,

$$\{(I_i, n_i)\}_{i=1,\ldots,5} : \{(I_1,17),(I_2,40),(I_3,16),(I_4,8),(I_5,3)\}.$$

$$\{(I_i,N_i)\}_{i=1,\ldots,5}:\{(I_1,17),(I_2,57),(I_3,73),(I_4,81),(I_5,84)\},$$

y

$$\{(I_i,F_i)\}_{i=1,\ldots,5}:\{(I_1,0.202),(I_2,0.679),(I_3,0.869),(I_4,0.964),(I_5,1)\}.$$

Por tanto, si sólo se dispusiera de la distribución agrupada, el valor mediano pertenecería al intervalo segundo, es decir,

$$Me_X \in (2.5,4.5].$$

Y dentro de este intervalo, dicho valor mediano puede calcularse como

$$Me_X = L_1 + \frac{\frac{N}{2}-N_1}{n_2}a_2 = 2.5+\frac{42-17}{40}2 = 3.75.$$

3.2.3. Cuantiles

Además, existen medidas de posición no necesariamente centrales denominadas cuantiles, que permiten ubicar partes de la distribución. Las más utilizadas son los cuartiles, los deciles y los percentiles.

Si se define la variable estadística X como $X:\{x_1^N,\ldots,x_N^N\}$, donde $\{x_1^N,\ldots,x_N^N\}$ es el conjunto de valores observados de una magnitud ordenados de menor a mayor para un conjunto de N individuos, entonces los cuartiles, que se denotarán por C_1, C_2 y C_3, son aquellos valores que, si es posible, dividen el conjunto $\{x_1^N,\ldots,x_N^N\}$ en cuatro partes con igual número de observaciones.

Como se comentó para el caso de la mediana, la identificación de los cuartiles puede efectuarse a partir de la distribución de frecuencias. Sea $\{(x_i,n_i)\}_{i=1,\ldots,n}$ la distribución de frecuencias de la variable estadística X. Si no existe ningún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{k}{4}$, $k=1,2,3$, entonces

$$C_k = x_j \Big/ x_j = min\left\{x_i \Big/ F_i > \frac{k}{4}\right\},\quad k=1,2,3,$$

es decir, el cuartil C_k es el mínimo de los valores del conjunto $\{x_i\}_{i=1,\ldots,n}$ con frecuencia relativa acumulada mayor que $\frac{k}{4}$. Por el contrario, si existe algún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{k}{4}$, entonces

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

perteneciente al conjunto $\{x_i\}_{i=1,\dots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{k}{100}$, entonces

$$P_k = \frac{x_j + x_{j+1}}{2}, \quad k = 1,\dots,99.$$

Nótese que el percentil veinticinco coincide con el primer cuartil, es decir, $C_1 = P_{25}$; el percentil cincuenta coincide con el quinto decil y, por tanto con el segundo cuartil y con la mediana, es decir, $Me_X = C_2 = D_5 = P_{50}$; y el percentil setenta y cinco coincide con el tercer cuartil, es decir, $C_3 = P_{75}$.

En el caso de distribuciones agrupadas, el percentil P_k pertenecerá al primero de los intervalos cuya frecuencia relativa acumulada sea mayor o igual que $\frac{k}{100}$. Es decir,

$$P_k \in \left(L_{j-1}, L_j\right] \Leftrightarrow j = mín\left\{i \middle/ F_i \geq \frac{k}{100}\right\}.$$

Y asumiendo que la frecuencia acumulada en el interior del intervalo va creciendo linealmente desde el extremo inferior al extremo superior, el valor P_k dentro del intervalo $\left(L_{j-1}, L_j\right]$ puede identificarse como

$$P_k = L_{j-1} + \frac{\frac{k}{100}N - N_{j-1}}{n_j} a_j, \quad k = 1,\dots,99,$$

siendo a_j la amplitud del intervalo $\left(L_{j-1}, L_j\right]$.

También pueden definirse medidas de posición más generales, denominadas cuantiles, que dividan la distribución en las partes que se desee. Si se desea dividir el conjunto $\{x_1^N,\dots,x_N^N\}$ de N valores observados y ordenados de menor a mayor de la variable estadística X en h partes con igual número de observaciones, pueden definirse los cuantiles k/h, $k = 1,\dots,h-1$, como aquellos $h-1$ valores que, si es posible, dividen el conjunto $\{x_1^N,\dots,x_N^N\}$ en las h partes señaladas.

La identificación de los cuantiles k/h, que se denotarán por $Q_{k/h}$, puede efectuarse también a partir de la distribución de frecuencias. Sea $\{(x_i, n_i)\}_{i=1,\dots,n}$ la distribución de frecuencias de la variable estadística X. Si no existe ningún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\dots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{k}{h}$, $k = 1,\dots,h-1$, entonces

$$Q_{k/h} = x_j \Big/ \; x_j = min\left\{ x_i \Big/ F_i > \frac{k}{h} \right\}, \quad k = 1,\ldots,h-1,$$

es decir, el cuantil $Q_{k/h}$ es el mínimo de los valores del conjunto $\{x_i\}_{i=1,\ldots,n}$ con frecuencia relativa acumulada mayor que $\frac{k}{h}$. Por el contrario, si existe algún valor x_j perteneciente al conjunto $\{x_i\}_{i=1,\ldots,n}$ tal que su frecuencia relativa acumulada sea igual a $\frac{k}{h}$, entonces

$$Q_{k/h} = \frac{x_j + x_{j+1}}{2}, \quad k = 1,\ldots,h-1.$$

En el caso de distribuciones agrupadas, el cuantil $Q_{k/h}$ pertenecerá al primero de los intervalos cuya frecuencia relativa acumulada sea mayor o igual que $\frac{k}{h}$. Es decir,

$$Q_{k/h} \in \left(L_{j-1}, L_j\right] \Leftrightarrow j = min\left\{ i \Big/ F_i \geq \frac{k}{h} \right\}.$$

Y asumiendo que la frecuencia acumulada en el interior del intervalo va creciendo linealmente desde el extremo inferior al extremo superior, el valor $Q_{k/h}$ dentro del intervalo $\left(L_{j-1}, L_j\right]$ puede identificarse como

$$Q_{k/h} = L_{j-1} + \frac{\frac{k}{h}N - N_{j-1}}{n_j} a_j, \quad k = 1,\ldots,h-1,$$

siendo a_j la amplitud del intervalo $\left(L_{j-1}, L_j\right]$.

Ejemplo 3.9 Suponga otra vez que la variable estadística X recoge el número de miembros de 84 familias. La distribución de frecuencias absolutas de esta variable es $\{(x_i, n_i)\}_{i=1,\ldots,10} : \{(1,6),(2,11),(3,21),(4,19),(5,8),(6,8),(7,7),(8,1),(9,2),(10,1)\}$, mientras que la correspondiente distribución de frecuencias relativas acumuladas viene dada por el conjunto

$\{(x_i, F_i)\}_{i=1,\ldots,10}$:

$\{(1,0.07),(2,0.202),(3,0.45),(4,0.68),(5,0.77),(6,0.87),(7,0.952),(8,0.96),(9,0.988),(10,1)\}$

Por tanto, los cuartiles son $C_1 = 3$, $C_2 = 4$ y $C_3 = 5$. Se tiene también que los deciles son $D_1 = 2$, $D_2 = 2$, $D_3 = 3$, $D_4 = 3$, $D_5 = 4$, $D_6 = 4$, $D_7 = 5$, $D_8 = 6$ y $D_9 = 7$. Por último, algunos percentiles son $P_1 = 1$, $P_5 = 1$, $P_{10} = 2$, $P_{25} = 3$, $P_{50} = 4$, $P_{75} = 5$, $P_{90} = 7$, $P_{95} = 7$ o $P_{99} = 10$.

Ejemplo 3.10 Suponga ahora que la variable estadística X recoge el precio de venta de un conjunto de 72 pisos ubicados en una zona urbana (en miles de euros). Agrupando los valores observados en los intervalos $I_1:(40,70]$, $I_2:(70,100]$, $I_3:(100,130]$, $I_4:(130,160]$, $I_5:(160,190]$, $I_6:(190,220]$, las distribuciones de frecuencias absolutas, acumuladas y relativas acumuladas son, respectivamente,

$$\{(I_i,n_i)\}_{i=1,\ldots,6}:\{(I_1,11),(I_2,38),(I_3,12),(I_4,7),(I_5,2),(I_6,2)\},$$

$$\{(I_i,N_i)\}_{i=1,\ldots,6}:\{(I_1,11),(I_2,49),(I_3,61),(I_4,68),(I_5,70),(I_6,72)\}$$

y

$$\{(I_i,F_i)\}_{i=1,\ldots,6}:\{(I_1,0.1528),(I_2,0.6806),(I_3,0.8472),(I_4,0.9444),(I_5,0.9722),(I_6,1)\}.$$

Por tanto, el primer cuartil pertenece al intervalo segundo, es decir, $C_1\in(70,100]$. Y dentro de este intervalo, dicho cuartil puede calcularse como

$$C_1=L_1+\frac{\frac{N}{4}-N_1}{n_2}a_2=70+\frac{18-11}{38}30=75.5263.$$

Por supuesto, el segundo cuartil coincidirá con la mediana. Nótese que $C_2\in(70,100]$ y, concretamente su valor es

$$C_2=L_1+\frac{\frac{2}{4}N-N_1}{n_2}a_2=70+\frac{36-11}{38}30=89.7368.$$

Finalmente, el tercer cuartil pertenece al intervalo tercero, es decir, $C_3\in(100,130]$, y puede identificarse como

$$C_3=L_2+\frac{\frac{3}{4}N-N_2}{n_3}a_3=100+\frac{54-49}{12}30=112.5.$$

Por otra parte, el primer decil pertenece al intervalo primero, es decir, $D_1\in(40,70]$. Y dentro de este intervalo, dicho decil puede calcularse como

$$D_1=L_0+\frac{\frac{N}{10}}{n_1}a_1=40+\frac{7.2}{11}30=59.6363.$$

De igual manera, se tiene, por ejemplo, que

$$D_5=L_1+\frac{\frac{5}{10}N-N_1}{n_2}a_2=70+\frac{36-11}{38}30=89.7368,$$

que, obviamente, coincide con la mediana.

Finalmente, algunos percentiles son

$$P_5 = L_0 + \frac{\frac{5}{100}N}{n_1} a_1 = 40 + \frac{5 \bullet 0.72}{11} 30 = 49.8181$$

o

$$P_{95} = L_4 + \frac{\frac{95}{100}N - N_4}{n_5} a_5 = 160 + \frac{68.4 - 68}{2} 30 = 166 .$$

3.3 MEDIDAS DE DISPERSIÓN

Una vez resumida la información contenida en la distribución de frecuencias, es posible preguntarse por la representatividad de las medidas de posición. Esta labor puede cumplirse con éxito acudiendo a las medidas de dispersión. Se distingue entre medidas de dispersión absolutas, que miden de una u otra forma la distancia entre los valores de la distribución y alguna medida de posición central, y medidas de dispersión relativas, que tratan de compensar el hecho de que tal distancia se ve afectada por las unidades en que se mide la variable.

3.3.1. Medidas de dispersión absolutas

Sea una variable estadística X con distribución de frecuencias. $\{(x_i, n_i)\}_{i=1,\ldots,n}$. Uno de los mecanismos más sencillos para evaluar el grado de dispersión de la distribución consiste en observar los valores extremos. Así, se define el recorrido de la variable estadística X, que se denotará por Re_X, como

$$Re_X = \max_i \{x_i\} - \min_i \{x_i\},$$

es decir, como la distancia entre el máximo y el mínimo de los valores observados. En un sentido similar, se define el recorrido intercuartílico, que se denotará por RI_X, como

$$RI_X = C_3 - C_1 ,$$

es decir, como la distancia entre el primer y el tercer cuartil. Pero las medidas de dispersión más utilizadas son la varianza y su raíz cuadrada positiva, la desviación típica. A continuación se explica el sentido de estos indicadores y la forma de obtenerlos.

Si se pretende medir la dispersión de la distribución, es decir, la lejanía entre los valores observados y alguna medida de posición central, puede pensarse en evaluar las magnitudes $x_i - \bar{x}$ y promediarlas. Pero, como ya se demostró, el promedio de las desviaciones entre los valores observados y la media aritmética es nulo, de modo que no puede aportar información que ayude a calibrar el grado de dispersión de diferentes distribuciones. En este sentido, podría optarse por evaluar las desviaciones absolutas

entre los valores observados y la media aritmética, $|x_i - \bar{x}|$. Y otra solución consiste en calcular las desviaciones al cuadrado, $(x_i - \bar{x})^2$, de forma que las desviaciones extremas tengan un peso relevante en el cálculo del promedio. Debido sobre todo a sus mejores propiedades para el cálculo matemático, el indicador generalmente adoptado para medir la dispersión es un promedio de las desviaciones al cuadrado entre los valores observados y la media aritmética. Así, se define la varianza de la variable estadística X, que se denotará por S_X^2, como

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

Como ya se adelantó, la varianza es el momento central de orden 2 y puede expresarse en términos de momentos respecto al origen. En concreto, se tiene que

$$S_X^2 = \mu_2 = m_2 - m_1^2 = \sum_{i=1}^{n} x_i^{\,2} \frac{n_i}{N} - \bar{x}^2.$$

Dado que la varianza está definida como un promedio de cantidades al cuadrado, nunca podrá ser negativa. Cuanto más alejados estén los valores observados con respecto a la media aritmética, mayor será la varianza, que reflejará por tanto el grado de dispersión. Sólo podrá considerarse que no existe dispersión cuando el conjunto de valores $\{x_i\}_{i=1,\ldots,n}$ contiene un solo valor. En ese caso, dicho valor coincidirá con la media aritmética y la varianza será nula.

Por otra parte, teniendo en cuenta que la varianza es una medida de dispersión y no de posición, no debería verse afectada por cambios de origen. Supóngase que se ha obtenido la distribución de frecuencias de una variable estadística Y, $\{(y_i, n_i)\}_{i=1,\ldots,n}$, sustituyendo los valores x_i de la distribución original por otros valores y_i tales que $y_i = x_i + k$, siendo k una constante. Entonces

$$S_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \frac{n_i}{N} = \sum_{i=1}^{n} (x_i + k - (\bar{x} + k))^2 \frac{n_i}{N} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{n_i}{N} = S_X^2,$$

de manera que, efectivamente, si todos los valores observados se incrementan en una cantidad constante k, el promedio se incrementa en la misma cantidad y, por tanto, el grado de dispersión con respecto a dicho promedio no se modifica.

Si se efectúa un cambio de escala, es decir, si la transformación es tal que $y_i = kx_i$, siendo k una constante, entonces

$$S_Y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \frac{n_i}{N} = \sum_{i=1}^{n} (kx_i - k\bar{x})^2 \frac{n_i}{N} = \sum_{i=1}^{n} k^2 (x_i - \bar{x})^2 \frac{n_i}{N} = k^2 S_X^2,$$

de modo que, si todos los valores observados se multiplican por una cantidad constante k, el promedio queda multiplicado por esa misma cantidad y, por tanto, lo mismo ocurre con las desviaciones entre los valores observados y el promedio, resultando que

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Ejemplo 3.11 Suponga de nuevo que la variable estadística X recoge el número de miembros de 84 familias. La distribución de frecuencias absolutas es

$$\{(x_i,n_i)\}_{i=1,\dots,10}:\{(1,6),(2,11),(3,21),(4,19),(5,8),(6,8),(7,7),(8,1),(9,2),(10,1)\}$$

Por tanto, la media aritmética de esta variable es

$$\overline{x}=\sum_{i=1}^{10}x_i\frac{n_i}{84}=4.0476\,.$$

La varianza es entonces

$$S_X^2=\sum_{i=1}^{10}x_i^{\,2}\frac{n_i}{N}-\overline{x}^2=3.8549\,.$$

Y la desviación típica es

$$S_X=+\sqrt{S_X^2}=1.9634\,.$$

El coeficiente de variación de Pearson es igual a

$$CV_X=\frac{S_X}{|\overline{x}|}=0.4851$$

y revela que la desviación típica es casi la mitad de la media aritmética.

Ejemplo 3.12 Suponga ahora que la variable estadística X recoge el precio de venta de un conjunto de 72 pisos ubicados en una zona urbana (en miles de euros). Definidos los intervalos $I_1:(40,70]$, $I_2:(70,100]$, $I_3:(100,130]$, $I_4:(130,160]$, $I_5:(160,190]$, $I_6:(190,220]$, la distribución de frecuencias absolutas es

$$\{(I_i,n_i)\}_{i=1,\dots,6}:\{(I_1,11),(I_2,38),(I_3,12),(I_4,7),(I_5,2),(I_6,2)\}$$

y la media aritmética de esta distribución es $\overline{x}=97.0833$. Por tanto, la varianza puede calcularse como

$$S_X^2=\sum_{i=1}^{6}c_i^{\,2}\frac{n_i}{N}-\overline{x}^2=1116.4995$$

De modo que la desviación típica y el coeficiente de variación de Pearson ascienden a

$$S_X=+\sqrt{S_X^2}=33.4141\,.$$

y

$$CV_X=\frac{S_X}{|\overline{x}|}=0.3442\,.$$

3.3.3. Variable tipificada

Una importante cuestión es la posible carencia de sentido de las comparaciones entre variables que se expresan en distintas unidades o tienen diferente grado de dispersión; por lo que es aconsejable, como paso previo, proceder a su estandarización, que permite obtener nuevas variables con media nula y varianza unitaria.

Sea una variable estadística X con distribución de frecuencias. $\{(x_i, n_i)\}_{i=1,\dots,n}$. Se define la variable estandarizada o tipificada Z con distribución de frecuencias $\{(z_i, n_i)\}_{i=1,\dots,n}$ como aquélla que resulta de efectuar la transformación de los valores originales de acuerdo con la relación $z_i = \dfrac{x_i - \overline{x}}{S_X}$. Es decir, los valores de la variable estadística Z son los valores originales a los que se ha restado la media aritmética y se ha dividido por la desviación típica. El resultado de este cambio de origen y escala es que la variable resultante tiene media nula y desviación típica unitaria. Nótese que

$$\overline{z} = \frac{1}{S_Y}(\overline{x} - \overline{x}) = 0$$

y

$$S_Z = \frac{1}{S_X} S_X = 1.$$

Mediante esta operación pueden compararse los valores estandarizados de dos variables estadísticas X e Y con diferentes medias y desviaciones típicas. Si se denotan los valores estandarizados de estas dos variables por z_i^X y z_i^Y, su signo positivo o negativo denota que el valor original está por encima o por debajo de la media correspondiente. Y su magnitud indica la posición del valor en términos de un número de desviaciones típicas por encima o por debajo de dicha media.

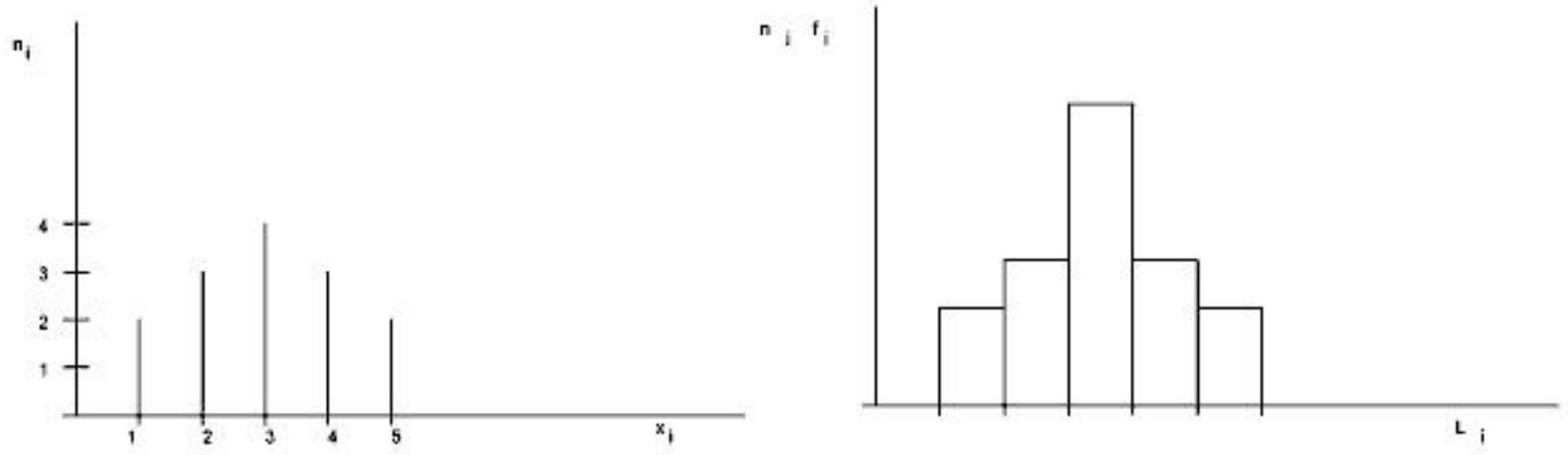
Ejemplo 3.13 Sea X el salario mensual en euros de un conjunto de trabajadores españoles e Y el salario mensual en libras de un conjunto de trabajadores ingleses. Se ha observado que el salario de uno de los españoles es $x = 900$ y el salario de uno de los ingleses es $y = 600$. A partir de estos dos valores no puede decirse que el español esté, dentro del conjunto de trabajadores españoles considerados, en mejor posición que la que tiene el inglés dentro del conjunto de trabajadores ingleses. Ahora bien, si los valores tipificados son $z^X = -1$ y $z^Y = 1$, deberia concluirse que el trabajador español posee un salario que está una desviación típica por debajo del salario medio del conjunto de trabajadores españoles, mientras que el trabajador inglés posee un salario que está una desviación típica por encima del salario medio del conjunto de trabajadores ingleses. Es decir, la posición relativa del trabajador inglés es mejor que la del español.

3.4 MEDIDAS DE FORMA

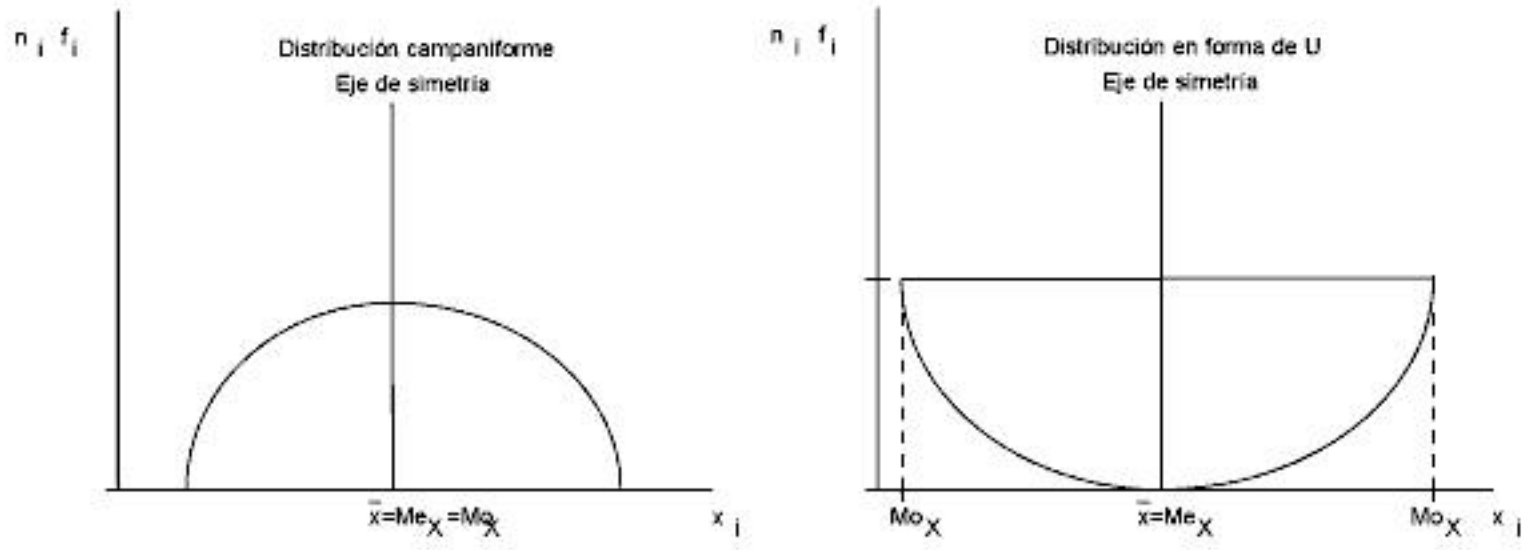
El análisis de la distribución de frecuencias se completa con las medidas de forma, que recogen características referidas, no ya a la magnitud de la dispersión respecto a las medidas de posición central, sino a la manera en la que los datos se distribuyen en torno a ellas. Tales características son la existencia o no de simetría y el grado de apuntamiento de la distribución.

3.4.1. Medidas de asimetría

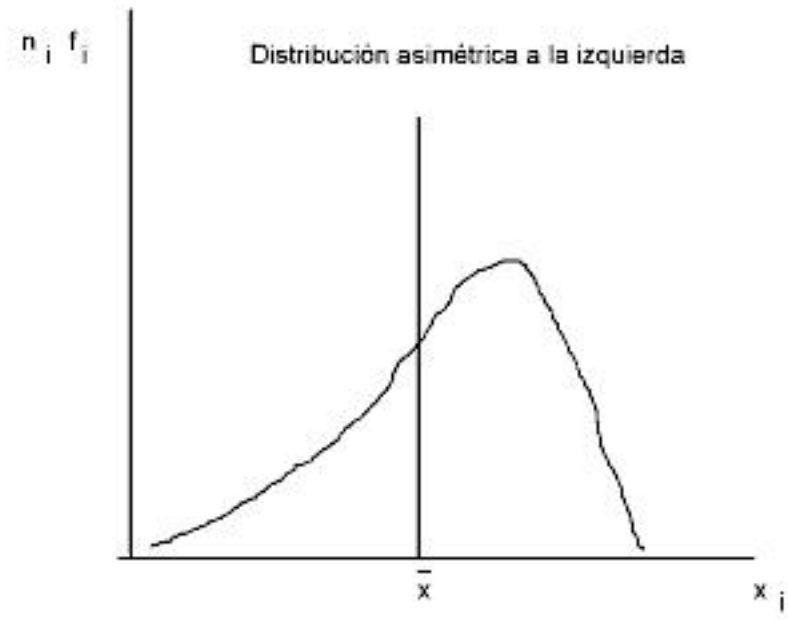
Se dice que una distribución es simétrica cuando su comportamiento a ambos lados de alguna medida de posición central es el mismo. Si se toma como punto de referencia la media aritmética, la distribución de una variable estadística será simétrica si el diagrama de barras presenta la misma altura a igual distancia de la media por la derecha y por la izquierda. Es decir, a la misma distancia de la media por la derecha y por la izquierda, y sea cual sea dicha distancia, los valores correspondientes deben haberse observado el mismo número de veces. En el caso de distribuciones agrupadas, existirá simetría cuando la media aritmética actúe como eje de simetría de forma que el área comprendida bajo el histograma a la izquierda de la media es la imagen del área comprendida a la derecha de ésta. Los gráficos siguientes ilustran este concepto.

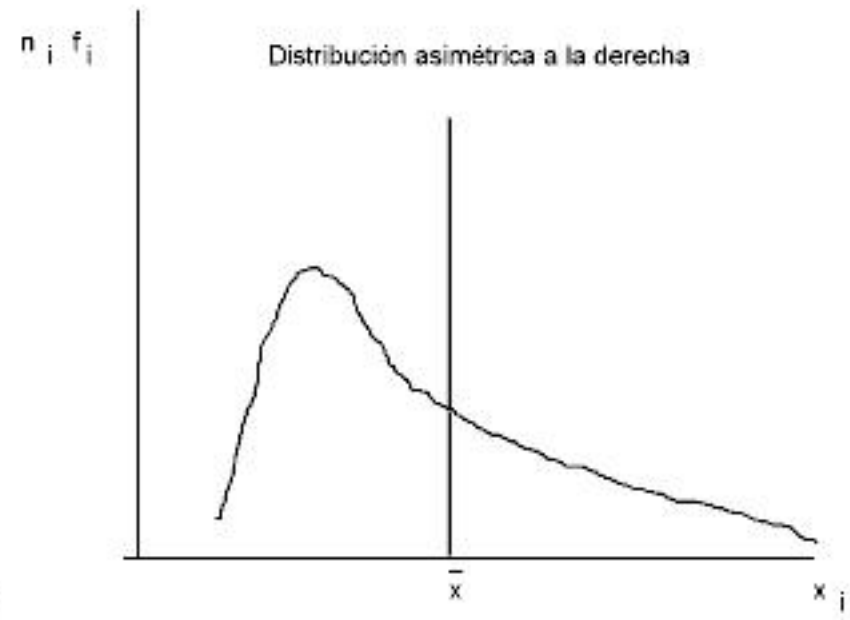


En el caso de distribuciones simétricas, la media aritmética coincide con la mediana. Y si la distribución es unimodal, coinciden media, mediana y moda. Este es el caso de las distribuciones campaniformes, en las que el polígono de frecuencias muestra una curva en forma de campana. Sin embargo, en distribuciones en forma de U, que serán bimodales, sólo coincidirán la media aritmética y la mediana.



Se dice que la distribución es asimétrica a la izquierda o que presenta asimetría negativa si existen muchos valores observados pequeños con frecuencias bajas, mientras que el número de valores altos y distintos es menor pero sus frecuencias son altas. De este modo, la rama del polígono de frecuencias a la izquierda de la media es más larga. Cuando existen valores altos con bajas frecuencias y los valores más bajos tienen frecuencias más altas, se dice que la distribución es asimétrica a la derecha o que presenta asimetría positiva. En este caso, la rama del polígono de frecuencias a la derecha de la media es más larga. Los dos gráficos siguientes muestran las dos situaciones para distribuciones campaniformes.





En distribuciones campaniformes asimétricas a la izquierda se tiene que $\bar{x} < Me_X < Mo_X$. Y en distribuciones campaniformes asimétricas a la derecha ocurre lo contrario, es decir, $Mo_X < Me_X < \bar{x}$.

Sobre el grado de asimetría informan varios coeficientes. Uno de ellos es el coeficiente de asimetría de Fisher, que se denotará por g_1, definido como

$$g_1 = \frac{\mu_3}{(S_X)^3},$$

es decir, como el cociente entre el momento central de orden 3 y el cubo de la desviación típica. El sentido de la asimetría lo proporciona el signo del numerador. Recuérdese que

$$\mu_3 = \sum_{i=1}^{n} (x_i - \bar{x})^3 \frac{n_i}{N},$$

de modo que se trata de un promedio de las desviaciones entre los valores observados y la media aritmética elevadas al cubo. La magnitud y el signo de estas desviaciones determinan la forma en que los datos se distribuyen alrededor de la media. Para conservar la información sobre el signo y la magnitud de las desviaciones es necesario elevar tales desviaciones a una potencia impar, y teniendo en cuenta que la suma de estas desviaciones es nula, el momento de tercer orden es la opción más sencilla.

Cuando existe simetría, este coeficiente será nulo. Mientras que si existe asimetría a la izquierda, el promedio está dominado por las desviaciones $x_i - \bar{x}$ de signo negativo

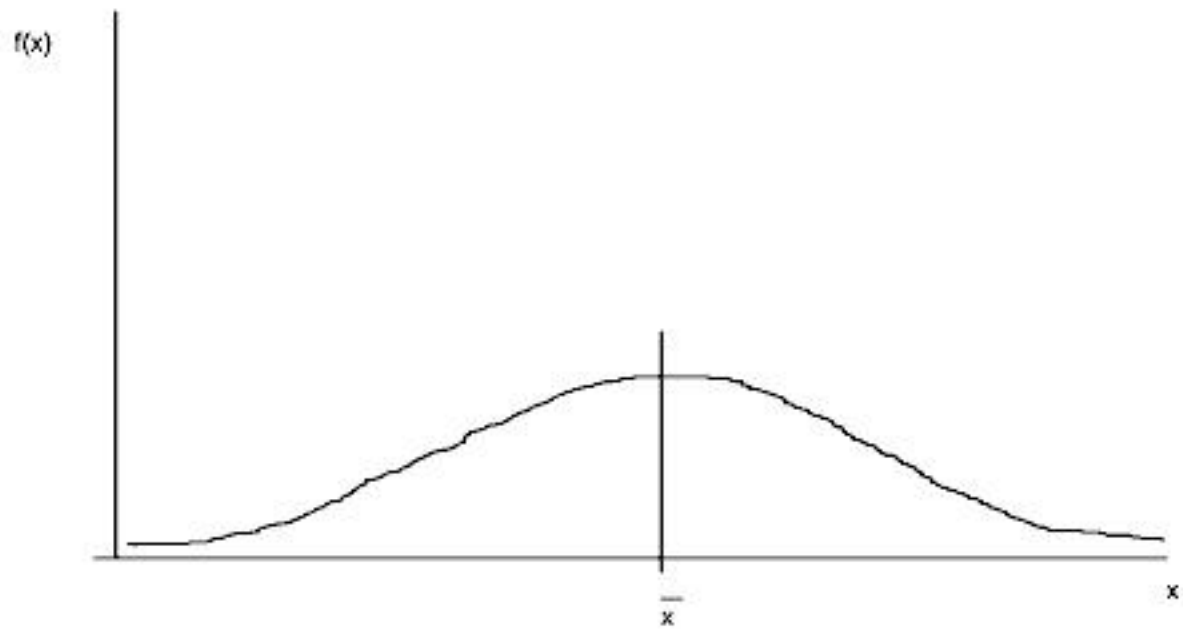
y el coeficiente será negativo. Por el contrario, si existe asimetría a la derecha, el promedio estará dominado por las desviaciones $x_i - \bar{x}$ de signo positivo y el coeficiente será también positivo. El denominador que interviene en el coeficiente de asimetría convierte a este último en un coeficiente adimensional, en el sentido de que no le afectan los cambios de escala. Por supuesto, el coeficiente tampoco se ve afectado por cambios de origen.

3.4.2. Medidas de apuntamiento o curtosis

Estas medidas resultan apropiadas para medir el grado de apuntamiento en distribuciones campaniformes moderadamente asimétricas. El apuntamiento o curtosis es un concepto que trata de reflejar la concentración de observaciones en la parte central de la distribución. Cuanto mayor sea esta concentración central, mayor será el apuntamiento observado en el polígono de frecuencias. Ahora bien la evaluación del grado de apuntamiento exige tomar algún punto de referencia. En este sentido, se considera que el grado de apuntamiento de la distribución de una variable estadística X, con media $\bar{x} = \mu$ y desviación típica $S_X = \sigma$, es normal si la línea descrita por el polígono de frecuencias se aproxima a la función

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Esta función corresponde a la denominada función de densidad de la distribución normal, que se explicará más adelante cuando se aborde la teoría matemática de la probabilidad. De momento, basta tener en cuenta que la forma de esta función es parecida a una campana, de forma que los valores más frecuentes deben ser los más próximos a la media, y la frecuencia de observación va disminuyendo simétricamente a medida que los valores observados se alejan de la media. Es decir,



De forma que una variable estadística cuya distribución presente una forma similar a la función de densidad normal, se dirá que posee un grado de apuntamiento normal o que es mesocúrtica. Si la concentración de frecuencias en la zona central es mayor, la distribución será más apuntada y se dice entonces que la distribución es leptocúrtica.

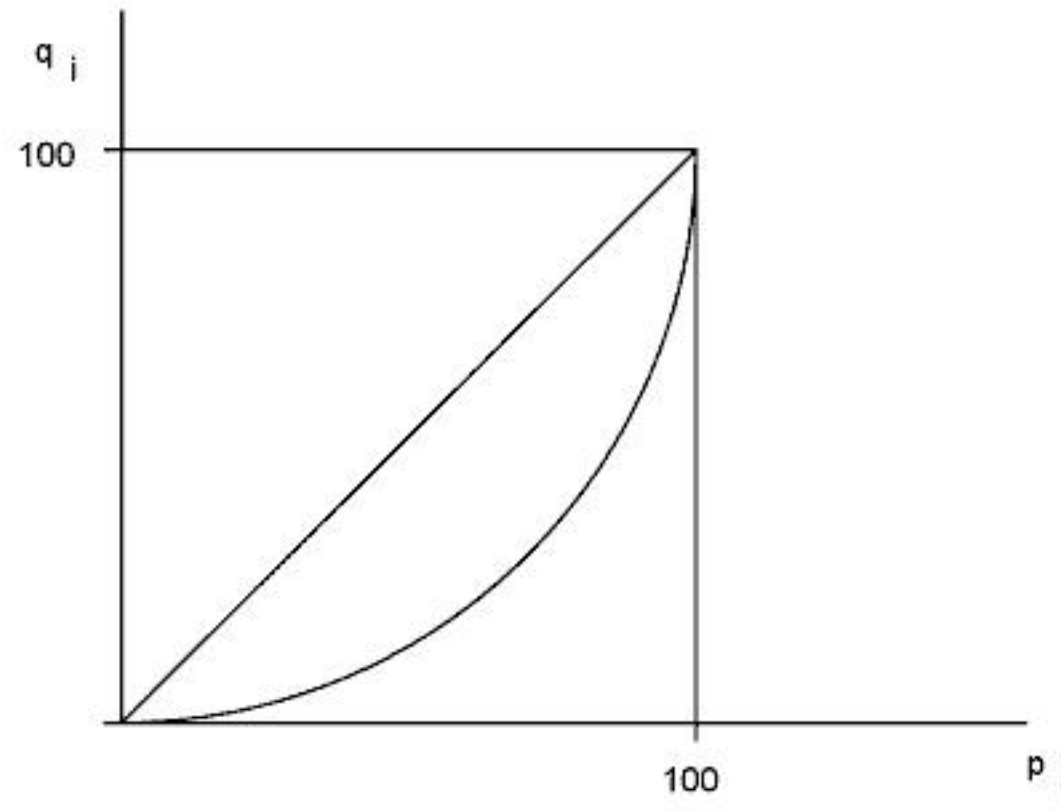
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

más bajos. La distribución será más equitativa cuanto más se aproximen estos porcentajes a los porcentajes correspondientes de trabajadores, definidos como $p_i = \frac{N_i}{N} \bullet 100$. En definitiva el grado de concentración o equidistribución en el reparto de la masa salarial puede calibrarse a partir del cálculo y comparación de estos porcentajes, tal como se indica en la tabla siguiente.

x_i	n_i	N_i	$x_i n_i$	u_i	$p_i = \frac{N_i}{N} \bullet 100$	$q_i = \frac{u_i}{u_n} \bullet 100$
x_1	n_1	N_1	$x_1 n_1$	u_1	p_1	q_1
x_2	n_2	N_2	$x_2 n_2$	u_2	p_2	q_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	n_n	N	$x_n n_n$	u_n	$p_n = 100$	$q_n = 100$

La curva de Lorenz no es más que la representación de los pares (p_i, q_i), $i = 1,\ldots,n$, en un gráfico como el siguiente.



Los pares (p_i, q_i) representados en la curva de Lorenz estarán por debajo de la diagonal principal, ya que q_i representa el porcentaje de salario que absorben los trabajadores con salarios más bajos y, por tanto, $q_i < p_i$, $i = 1,\ldots,n-1$. De manera que cuanto mayor sea el área comprendida entre la curva de Lorenz y la diagonal principal, mayor es el grado de concentración en la distribución salarial.

Una aproximación al área comprendida entre la curva de Lorenz y la diagonal principal puede obtenerse a través de un índice denominado índice de Gini (Gini, 1953:210-218), que se denotará por I_G, definido como

$$I_G = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}.$$

Si la concentración es mínima, los valores p_i y q_i estarán muy próximos. Y en la situación límite, podría considerarse que p_i tiende a igualarse a q_i, de modo que I_G tiende a 0. Por el contrario, en las situaciones próximas a la máxima concentración, los porcentajes q_i, $i = 1,...,n-1$, tienden a 0, de modo que I_G tiende a 1. En resumen, cuanto mayor sea el índice de Gini, mayor es el grado de concentración, aunque un mismo valor de este índice puede corresponder a repartos diferentes.

En el caso de distribuciones agrupadas, el cálculo de los q_i puede efectuarse utilizando las marcas de clase correspondientes a los diferentes intervalos. Ahora bien, debe tenerse en cuenta que tanto el número de intervalos como la definición de éstos tienen consecuencias sobre los estadísticos empleados para evaluar el grado de concentración.

Ejemplo 3.15 Suponga que la variable estadística X recoge los salarios mensuales de un conjunto de 72 trabajadores de una empresa (en euros). Agrupando los valores observados en los intervalos $I_1 : (1000,1250]$, $I_2 : (1250,1500]$, $I_3 : (1500,1750]$, $I_4 : (1750,2000]$, $I_5 : (2000,2250]$, $I_6 : (2250,2500]$, las distribuciones de frecuencias absolutas, acumuladas y relativas acumuladas son, respectivamente,

$$\{(I_i,n_i)\}_{i=1,...,6} : \{(I_1,11),(I_2,38),(I_3,12),(I_4,7),(I_5,2),(I_6,2)\},$$

$$\{(I_i,N_i)\}_{i=1,...,6} : \{(I_1,11),(I_2,49),(I_3,61),(I_4,68),(I_5,70),(I_6,72)\}$$

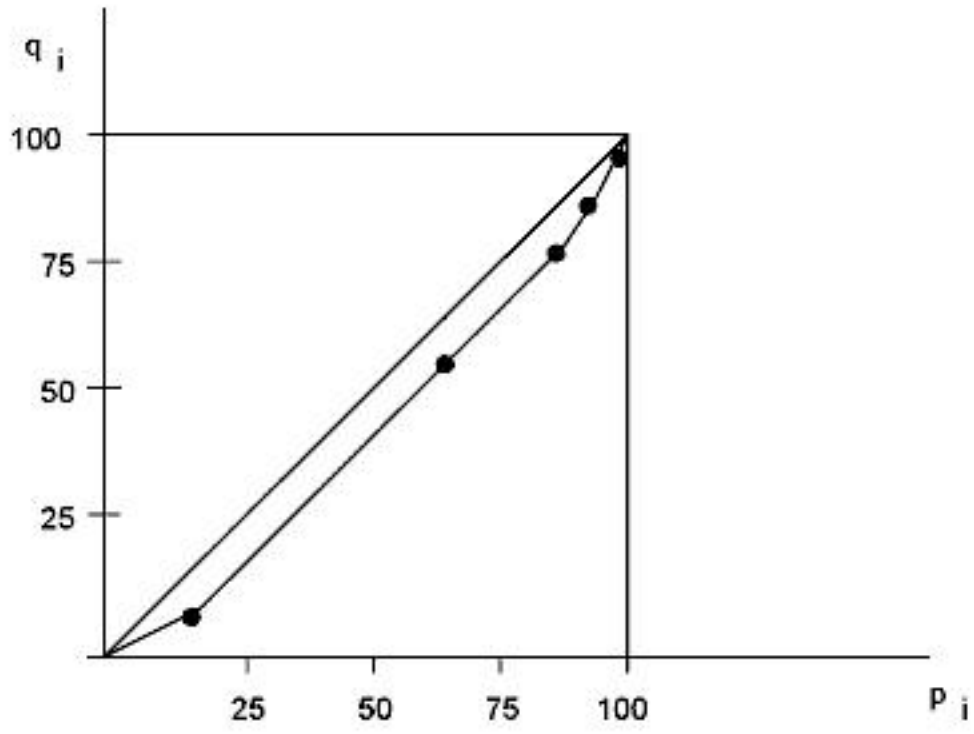
y

$$\{(I_i,F_i)\}_{i=1,...,6} : \{(I_1,0.1528),(I_2,0.6806),(I_3,0.8472),(I_4,0.9444),(I_5,0.9722),(I_6,1)\}.$$

Para evaluar la equidad en la distribución de salarios en la empresa, puede representarse la curva de Lorenz y calcular el índice de Gini.

c_i	n_i	N_i	$c_i n_i$	u_i	$p_i = \frac{N_i}{N} \bullet 100$	$q_i = \frac{u_i}{u_n} \bullet 100$
1125	11	11	12375	12375	15.28	11.65
1375	38	49	52250	64625	68.06	60.82
1625	12	61	19500	84125	84.72	79.18
1875	7	68	13125	97250	94.44	91.53
2125	2	70	4250	101500	97.22	95.53
2375	2	72	4750	106250	100	100

La curva de Lorenz es la siguiente.



Obsérvese que la curva está bastante próxima a la diagonal, lo que significa que la distribución de la masa salarial es bastante equitativa. Este resultado debe ser confirmado por el índice de Gini, cuyo valor es

$$I_G = \frac{\sum_{i=1}^{5}(p_i - q_i)}{\sum_{i=1}^{5} p_i} = \frac{3.63+7.23+5.55+2.92+1.69}{15.28+68.06+84.72+94.44+97.22} = 0.0584\,,$$

que resulta bastante cercano a cero.

EJERCICIOS

3.1. Sea una distribución agrupada en intervalos de amplitud constante tal que, aplicando los criterios convencionales, resulta que el valor modal se sitúa justo en el punto medio del intervalo modal. ¿Puede asegurarse que las frecuencias de los intervalos contiguos al modal son iguales?

3.2. Suponga que se realiza un estudio sobre el número de automóviles, X, de un grupo de familias residentes en una zona rural, obteniéndose los siguientes resultados.

x_i	0	1	2	3	N
n_i	3	14	7	6	30

(a) Calcule los cuartiles y los deciles primero, quinto, séptimo y octavo.
(b) Calcule la varianza y la desviación típica.

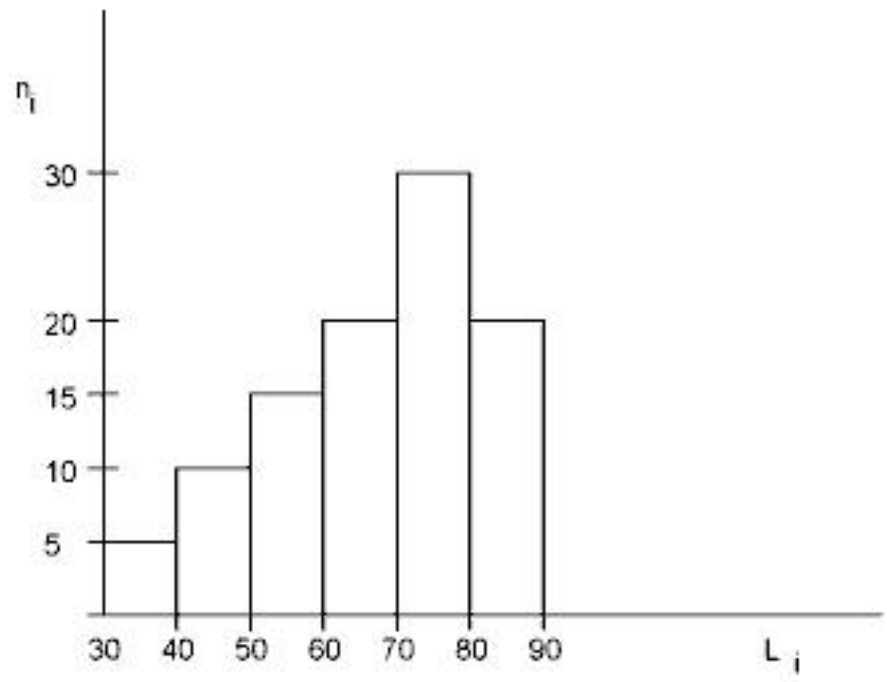
3.3. Se han recogido las puntuaciones en dos escalas de satisfacción, X e Y, de dos grupos de 6 personas, obteniéndose los resultados que a continuación se describen.

x_i	n_i
60	1
63	1
65	2
68	2

y_i	n_i
50	1
68	1
70	2
75	1
80	1

Determine cuál de las dos distribuciones es más dispersa.

3.4. El siguiente histograma muestra la distribución de las edades de 100 personas entre 30 y 90 años agrupadas en los intervalos $(30,40]$, $(40,50]$, $(50,60]$, $(60,70]$, $(70,80]$ y $(80,90]$.



(a) Determine:
  (a.1) el número de personas con edad no superior a 60 años;
  (a.2) la edad que supera el 25% de las personas;
  (a.3) la edad que no supera el 20% de las personas;
  (a.4) la edad media, modal y mediana.

(b) Evalúe la dispersión de la distribución.

3.5. La siguiente tabla muestra las distribuciones de frecuencias de tres variables estadísticas que recogen datos relativos al número de trabajos desempeñados a lo largo de su vida profesional por tres grupos de individuos: jóvenes, de edades intermedias y de edades más avanzadas, que se denotan, respectivamente, por A, B y C.

	A	B	C
x_i	n_i	n_i	n_i
1	6	3	1
2	3	4	3
3	1	3	6
N	10	10	10

(a) Represente los polígonos de frecuencias.
(b) Calcule el coeficiente de asimetría.

3.6. Un estudio sobre el número de habitaciones de la residencia habitual de 50 personas de un determinado colectivo arroja los resultados que se muestran en la siguiente tabla.

x_i	1	3	4	6	10	N
n_i	5	12	20	8	5	50

(a) Calcule la media aritmética, la moda, la mediana y la desviación típica.
(b) Calcule los coeficientes de asimetría y apuntamiento.

3.7. Las puntuaciones sobre 20 puntos totales relativas a la valoración de la vivienda por parte de 60 personas están recogidas en la siguiente tabla.

8	9.5	10.5	9.5	10.5	4.5	6.5	13.5	13	7	13.5	12
9	10.5	18	11.5	14.5	12	5.5	10	10	14.5	6	10
11	13	13	11.5	12	10.5	10	17.5	11.5	6	9	8
8.5	11.5	17	12.5	8	11.5	9	11	13	7.5	9	15
7	11.5	11.5	13	5.5	11.5	12	10	12	9	12	11.5

Agrupe la variable en intervalos de dos puntos de amplitud y calcule:
(a) moda, media y mediana;
(b) recorrido intercuartílico y desviación típica;
(c) coeficientes de asimetría y apuntamiento.

3.8. Suponga que se han registrado los salarios de los 100 trabajadores de cada una de las empresas *A* y *B* y, agrupándolos en los intervalos que se indican, se han obtenido las distribuciones de frecuencias que se muestran en la tabla siguiente.

Empresa A

I_i	n_i
(1000,1500]	20
(1500,2000]	35
(2000,2500]	15
(2500,3000]	10
(3000,3500]	20
N	100

Empresa B

I_i	n_i
(1000,1500]	20
(1500,2000]	10
(2000,2500]	10
(2500,3000]	20
(3000,3500]	40
N	100

(a) Calcule el índice de Gini. ¿En qué empresa están los salarios más equitativamente distribuidos?
(b) Represente la curva de Lorenz. ¿En qué empresa está en mejor posición relativa el 20% de trabajadores con salarios más bajos?

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

4.1 VARIABLE ESTADÍSTICA MULTIDIMENSIONAL Y DISTRIBUCIÓN DE FRECUENCIAS

Una variable estadística multidimensional $(X_1,...,X_m)$ es un conjunto de observaciones referidas a m magnitudes o características para un conjunto de N individuos. En el ejemplo anterior, la población en cuestión es el conjunto de trabajadores de la actividad y, para cada individuo de la población, se desea conocer su edad y su salario. Si se definen las variables X :"edad en años de los trabajadores" e Y :"salario anual en miles de euros de los trabajadores", se tiene que (X,Y), es decir, el conjunto de pares de observaciones que indican la edad y salario de cada uno de los N trabajadores, es una variable estadística bidimensional. Una variable estadística bidimensional (X,Y) se denotará por

$$(X,Y):\{(x_1^N,y_1^N),...,(x_N^N,y_N^N)\}:\{(x_k^N,y_k^N)\}_{k=1,...,N},$$

donde (x_k^N,y_k^N) representa el par de valores observados de las magnitudes X e Y para el k -ésimo individuo de la población.

Toda la información sobre determinada magnitud que contiene la variable estadística bidimensional (X,Y) puede recogerse de forma sistemática a través de la denominada distribución bidimensional de frecuencias. Es decir, a partir del conjunto de N pares de observaciones registradas para los individuos de la población, es posible identificar el conjunto de pares de valores distintos y ordenarlos, primero, en orden creciente de los valores de una característica y, segundo, en orden creciente de los valores de la otra característica. Si existen n pares de valores distintos, este conjunto puede denotarse por $\{(x_1,y_1),...,(x_n,y_n)\}:\{(x_i,y_i)\}_{i=1,...,n}$. Y si cada par de valores (x_i,y_i) se ha observado n_i veces, entonces el conjunto de información puede representarse como

$$\{(x_1,y_1;n_1),...,(x_n,y_n;n_n)\}:\{(x_i,y_i;n_i)\}_{i=1,...,n},$$

es decir, indicando la frecuencia absoluta n_i con que se repite cada par de valores (x_i,y_i). Este conjunto de pares define la distribución bidimensional de frecuencias absolutas. Por supuesto, se tiene que $\sum_{i=1}^{n} n_i = n_1 + ... + n_n = N$.

Pero es más frecuente definir esta distribución en términos de todos los pares que pueden formarse combinando los valores diferentes de ambas características. Suponga que el conjunto de valores diferentes observados para la característica X es $\{x_i\}_{i=1,...,r}:\{x_1,...,x_r\}$, mientras que en el caso de la característica Y el conjunto anterior es $\{y_j\}_{i=1,...,s}:\{y_1,...,y_s\}$. Entonces, la distribución bidimensional de frecuencias puede escribirse en términos del conjunto de pares $\{(x_i,y_j)\}_{i=1,...,r;\,j=1,...,s}$. Toda la

información que contiene la variable estadística bidimensional (X,Y) puede recogerse en el conjunto

$$\{(x_i, y_j; n_{i,j})\}_{i=1,\dots,r;\, j=1,\dots,s},$$

donde $n_{i,j}$ es el número de veces que se repite el par (x_i, y_j), es decir, $n_{i,j}$ es la frecuencia absoluta del par de valores observados (x_i, y_j). Ahora se tiene que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j} = \sum_{i=1}^{r}\left(n_{i,1} + \dots + n_{i,s}\right) = \left(n_{1,1} + \dots + n_{1,s}\right) + \dots + \left(n_{r,1} + \dots + n_{r,s}\right) = N .$$

La distribución de frecuencias relativas de la variable estadística bidimensional (X,Y) puede expresarse como

$$\{(x_i, y_j; f_{i,j})\}_{i=1,\dots,r;\, j=1,\dots,s},$$

donde $f_{i,j}$ es la proporción de casos en los que se observa el par (x_i, y_j), es decir, $f_{i,j}$ es la frecuencia relativa del par de valores observados (x_i, y_j). Nótese que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{i,j} = \sum_{i=1}^{r}\left(f_{i,1} + \dots + f_{i,s}\right) = \left(f_{1,1} + \dots + f_{1,s}\right) + \dots + \left(f_{r,1} + \dots + f_{r,s}\right) = 1 .$$

Esta información puede escribirse en tablas de doble entrada del modo siguiente.

$X \setminus Y$	y_1	$\cdots$	y_j	$\cdots$	y_s
x_1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_r	$n_{r,1}$	$\cdots$	$n_{r,j}$	$\cdots$	$n_{r,s}$

$X \setminus Y$	y_1	$\cdots$	y_j	$\cdots$	y_s
x_1	$f_{1,1}$	$\cdots$	$f_{1,j}$	$\cdots$	$f_{1,s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_i	$f_{i,1}$	$\cdots$	$f_{i,j}$	$\cdots$	$f_{i,s}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
x_r	$f_{r,1}$	$\cdots$	$f_{r,j}$	$\cdots$	$f_{r,s}$

Ejemplo 4.1 Suponga que se han registrado las edades en años y salarios anuales en miles de euros de 10 trabajadores. Indicando primero la edad y después el salario, los valores observados fueron: $(20,8)$, $(20,8)$, $(26,14)$, $(26,14)$, $(33,21)$, $(33,21)$, $(20,8)$, $(20,8)$, $(26,14)$, $(33,21)$. Si se definen las variables X :"edad en años de los 10 trabajadores" e Y :"salario anual en miles de euros de los 10 trabajadores", se tiene que

$$(X,Y):\left\{\left(x_k^{10},y_k^{10}\right)\right\}_{k=1,\ldots,10}$$

$$:\{(20,8),(20,8),(26,14),(26,14),(33,21),(33,21),(20,8),(20,8),(26,14),(33,21)\}$$

La distribución bidimensional de frecuencias absolutas de la variable estadística (X,Y) puede escribirse como

$$\{(x_i,y_i;n_i)\}_{i=1,2,3}:\{(20,8;4),(26,14;3),(33,21;3)\}$$

o también como

$$\{(x_i,y_j;n_{i,j})\}_{\substack{i=1,2,3\\j=1,2,3}}:$$

$$:\{(20,8;4),(20,14;0),(20,21;0),(26,8;0),(26,14;3),(26,21;0),(33,8;0),(33,14;0),(33,21;3)\}$$

Esta última distribución de frecuencias puede expresarse en la tabla siguiente.

$X \setminus Y$	8	14	21
20	4	0	0
26	0	3	0
33	0	0	3

Y la distribución de frecuencias relativas de la variable estadística (X,Y) puede escribirse como

$$\{(x_i,y_i;f_i)\}_{i=1,2,3}:\left\{\left(20,8;\frac{4}{10}\right),\left(26,14;\frac{3}{10}\right),\left(33,21;\frac{3}{10}\right)\right\}$$

o también como

$$\{(x_i,y_j;f_{i,j})\}_{\substack{i=1,2,3\\j=1,2,3}}:$$

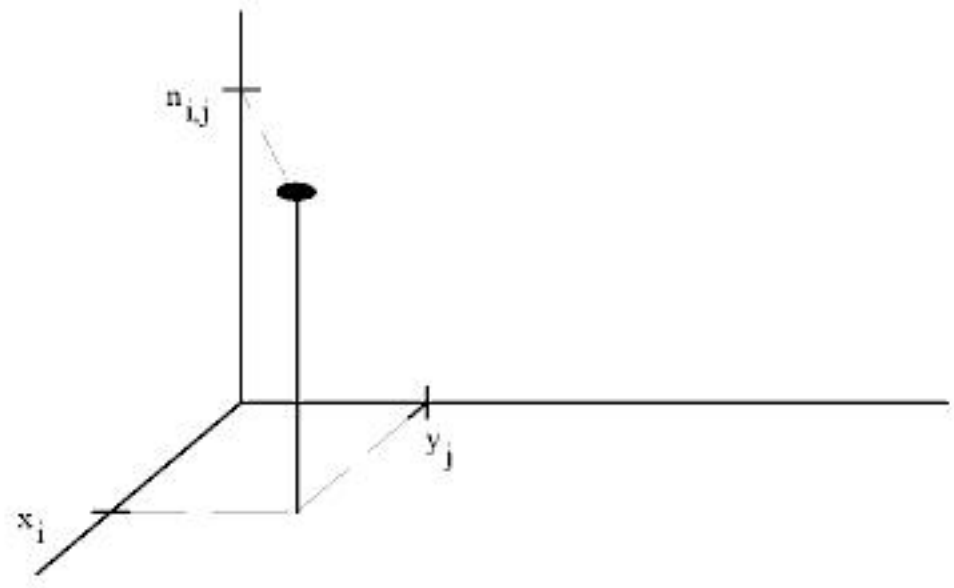
$$\left\{\left(20,8;\frac{4}{10}\right),(20,14;0),(20,21;0),(26,8;0),\left(26,14;\frac{3}{10}\right),(26,21;0),(33,8;0),(33,14;0),\left(33,21;\frac{3}{10}\right)\right\}$$

Y esta última distribución de frecuencias relativas puede expresarse en la tabla siguiente.

$X \setminus Y$	8	14	21
20	0.4	0	0
26	0	0.3	0
33	0	0	0.3

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

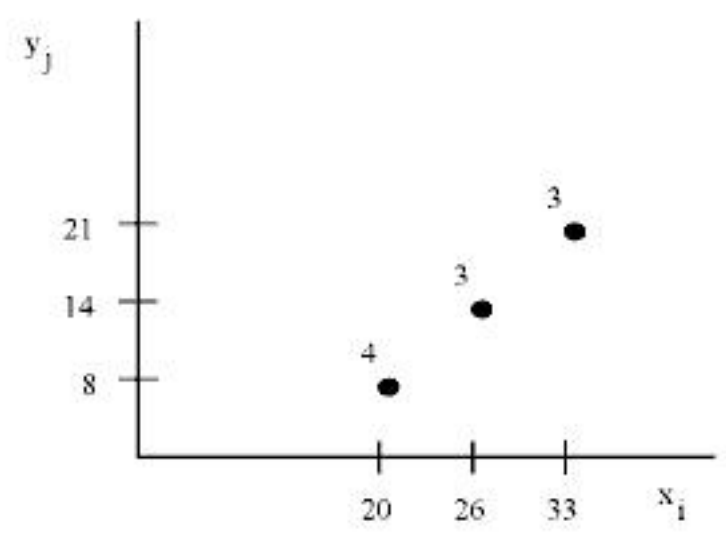
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.



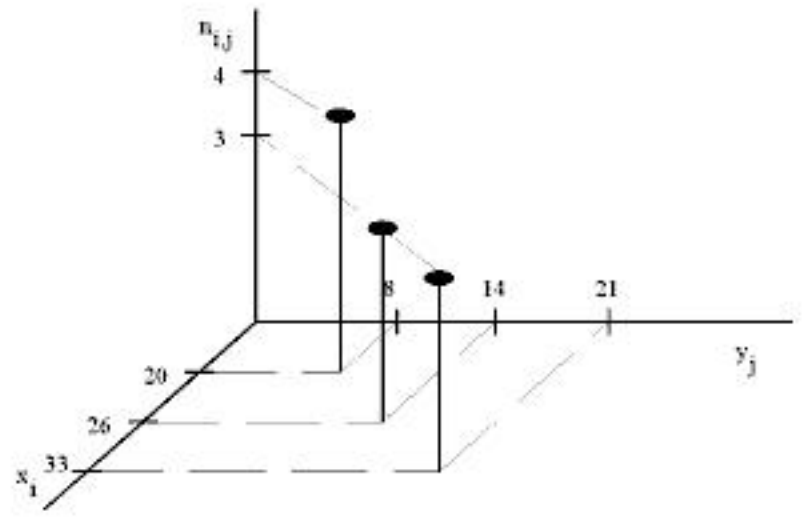
Ejemplo 4.3 Sea la variable estadística (X,Y), donde X :"edad en años de 10 trabajadores" e Y :"salario anual en miles de euros de los 10 trabajadores", tal que la distribución de frecuencias absolutas es la que se indica en la tabla siguiente.

$X \setminus Y$	8	14	21
20	4	0	0
26	0	3	0
33	0	0	3

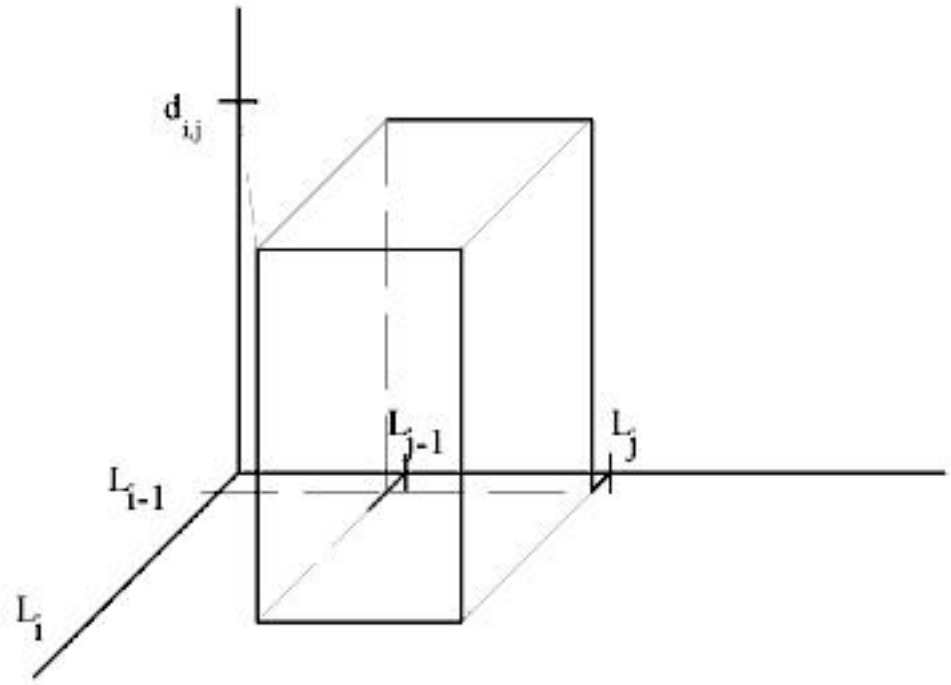
Esta distribución puede representarse mediante la nube de puntos que se indica.



O también a través del siguiente diagrama de dispersión tridimensional.



En el caso de distribuciones agrupadas, el diagrama de dispersión tridimensional puede construirse dibujando, sobre cada uno de los rectángulos que resulta de cruzar el segmento que definen los extremos del intervalo $I_i : (L_{i-1}, L_i]$ para la variable estadística X con el que definen los extremos del intervalo $I_j : (L_{j-1}, L_j]$ para la variable estadística Y, un cubo cuyo volumen sea proporcional a la frecuencia $n_{i,j}$. Si los intervalos $I_i : (L_{i-1}, L_i]$, $i = 1,...,r$, son todos de la misma amplitud y lo mismo sucede con los intervalos $I_j : (L_{j-1}, L_j]$, $j = 1,...,s$, entonces la proporcionalidad entre volúmenes y frecuencias puede conseguirse tomando como altura de los cubos las propias frecuencias $n_{i,j}$ o las frecuencias relativas $f_{i,j}$. Sin embargo, si todos los intervalos $I_i : (L_{i-1}, L_i]$, $i = 1,...,r$, no son de la misma amplitud, o bien, no todos los intervalos $I_j : (L_{j-1}, L_j]$, $j = 1,...,s$, poseen la misma amplitud, entonces la proporcionalidad entre volúmenes y frecuencias puede conseguirse tomando como altura de los cubos las densidades de frecuencia $d_{i,j}$ definidas como $d_{i,j} = \dfrac{n_{i,j}}{a_i a_j}$, siendo a_i la amplitud del intervalo $I_i : (L_{i-1}, L_i]$ y a_j la amplitud del intervalo $I_j : (L_{j-1}, L_j]$.



4.3 DISTRIBUCIONES MARGINALES

A partir de la distribución bidimensional de frecuencias, es posible determinar el número de veces que se observa cada uno de los valores de una u otra de las dos magnitudes que definen la variable estadística bidimensional. Esta información sobre cada una de las dos variables unidimensionales queda recogida en las denominadas distribuciones marginales.

Sea una variable estadística bidimensional (X, Y) cuya distribución de frecuencias es

$$\{(x_i, y_j; n_{i,j})\}_{i=1,...,r;\, j=1,...,s}.$$

Si se representan estas frecuencias $\{n_{i,j}\}_{i=1,\dots,r;\,j=1,\dots,s}$ en una tabla de doble entrada, las distribuciones marginales pueden obtenerse sumando las filas o las columnas. Es decir, las frecuencias $\{n_{i,\bullet}\}_{i=1,\dots,r}$ correspondientes a los valores $\{x_i\}_{i=1,\dots,r}$ de la variable estadística X pueden obtenerse como

$$n_{i,\bullet} = \sum_{j=1}^{s} n_{i,j}\,,\ i = 1,\dots,r\,.$$

Del mismo modo, las frecuencias $\{n_{\bullet,j}\}_{j=1,\dots,s}$ correspondientes a los valores $\{y_j\}_{j=1,\dots,s}$ de la variable estadística Y pueden obtenerse como

$$n_{\bullet,j} = \sum_{i=1}^{r} n_{i,j}\,,\ j = 1,\dots,s\,.$$

$X \setminus Y$	y_1	$\cdots$	y_j	$\cdots$	y_s	
x_1	$n_{1,1}$	$\cdots$	$n_{1,j}$	$\cdots$	$n_{1,s}$	$n_{1,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	$n_{i,1}$	$\cdots$	$n_{i,j}$	$\cdots$	$n_{i,s}$	$n_{i,\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	$n_{r,1}$	$\cdots$	$n_{r,j}$	$\cdots$	$n_{r,s}$	$n_{r,\bullet}$
	$n_{\bullet,1}$	$\cdots$	$n_{\bullet,j}$	$\cdots$	$n_{\bullet,s}$	N

Así, la distribución marginal de la variable estadística X viene dada por

$$\{(x_i, n_{i,\bullet})\}_{i=1,\dots,r}\,,$$

mientras que la distribución marginal de la variable estadística Y es

$$\{(y_j, n_{\bullet,j})\}_{j=1,\dots,s}\,.$$

Una vez obtenidas las distribuciones marginales de frecuencias absolutas se pueden obtener las distribuciones de frecuencias relativas. En el caso de la variable estadística X, la distribución de frecuencias relativas es

$$\{(x_i, f_{i,\bullet})\}_{i=1,\dots,r}\,,$$

donde

$$f_{i,\bullet} = \frac{n_{i,\bullet}}{N}\,,\ i = 1,\dots,r\,.$$

La distribución de frecuencias relativas de la variable estadística Y es

$$\left\{\left(y_j, f_{\bullet j}\right)\right\}_{j=1,\ldots,s},$$

donde

$$f_{\bullet j} = \frac{n_{\bullet j}}{N},\ j = 1,\ldots,s\ .$$

4.4 DISTRIBUCIONES CONDICIONADAS

Las distribuciones condicionadas recogen la distribución de frecuencias de una variable cuando la otra toma un valor dado. Sea una variable estadística bidimensional (X,Y) cuya distribución de frecuencias es

$$\left\{\left(x_i, y_j; n_{ij}\right)\right\}_{i=1,\ldots,r;j=1,\ldots,s}.$$

La distribución de frecuencias de la variable estadística Y condicionada a que la variable estadística X tome el valor x_i, que se denotará por $Y/X = x_i$, recoge el número de veces que se observa cada uno de los valores de la variable estadística Y en el conjunto de observaciones tales que $X = x_i$. Es decir, para cada uno de los valores x_i, $i = 1,\ldots,r$, que toma la variable estadística X, se obtiene una variable estadística $Y/X = x_i$ cuya distribución de frecuencias absolutas es

$$\left\{\left(y_j, n_{ij}\right)\right\}_{j=1,\ldots,s},$$

mientras que la distribución de frecuencias relativas de esta variable viene dada por

$$\left\{\left(y_j, f_{j/i}\right)\right\}_{j=1,\ldots,s},$$

donde

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}},\ j = 1,\ldots,s\ .$$

De la misma manera, la distribución de frecuencias de la variable estadística X condicionada a que la variable estadística Y tome el valor y_j, que se denotará por $X/Y = y_j$, recoge el número de veces que se observa cada uno de los valores de la variable estadística X en el conjunto de observaciones tales que $Y = y_j$. Es decir, para cada uno de los valores y_j, $j = 1,\ldots,s$, que toma la variable estadística Y, se obtiene una variable estadística $X/Y = y_j$ cuya distribución de frecuencias absolutas es

$$\left\{\left(x_i, n_{ij}\right)\right\}_{i=1,\ldots,r},$$

mientras que la distribución de frecuencias relativas de esta variable viene dada por

$$\left\{\left(x_i, f_{i/j}\right)\right\}_{i=1,\ldots,r},$$

donde

$$f_{i/j} = \frac{n_{i,j}}{n_{\bullet,j}},\ i = 1,\ldots,r.$$

Ejemplo 4.4 Sea la variable estadística (X,Y), donde X :"edad en años de 20 trabajadores" e Y : "salario anual en miles de euros de los 20 trabajadores", tal que la distribución de frecuencias absolutas es la que se indica en la tabla siguiente.

$X \setminus Y$	8	14	21	
20	5	1	0	6
26	1	5	2	8
33	1	1	4	6
	7	7	6	20

La distribución marginal de la variable estadística X viene dada por

$$\left\{\left(x_i, n_{i\bullet}\right)\right\}_{i=1,\ldots,3} : \{(20,6),(26,8),(33,6)\},$$

mientras que la distribución marginal de la variable estadística Y es

$$\left\{\left(y_j, n_{\bullet,j}\right)\right\}_{j=1,\ldots,3} : \{(8,7),(14,7),(21,6)\}.$$

Y las distribuciones marginales de frecuencias relativas son

$$\left\{\left(x_i, f_{i\bullet}\right)\right\}_{i=1,\ldots,3} : \left\{\left(20,\frac{6}{20}\right),\left(26,\frac{8}{20}\right),\left(33,\frac{6}{20}\right)\right\}$$

y

$$\left\{\left(y_j, f_{\bullet,j}\right)\right\}_{j=1,\ldots,3} : \left\{\left(8,\frac{7}{20}\right),\left(14,\frac{7}{20}\right),\left(21,\frac{6}{20}\right)\right\}.$$

La distribución de frecuencias absolutas de la variable estadística $Y/X = 20$ es

$$\left\{\left(y_j, n_{1,j}\right)\right\}_{j=1,\ldots,3} : \{(8,5),(14,1),(21,0)\},$$

y su distribución de frecuencias relativas es

$$\left\{\left(y_j, f_{j/1}\right)\right\}_{j=1,\ldots,3} : \left\{\left(8,\frac{5}{6}\right),\left(14,\frac{1}{6}\right),(21,0)\right\}.$$

Para la variable estadística $Y/X = 26$ estas distribuciones son

$$\left\{\left(y_j, n_{2,j}\right)\right\}_{j=1,\ldots,3} : \{(8,1),(14,5),(21,2)\}$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$f_{j/i} = \frac{n_{i,j}}{n_{i,\bullet}} = \frac{\frac{n_{i,j}}{N}}{\frac{n_{i,\bullet}}{N}} = \frac{f_{i,\bullet} f_{\bullet,j}}{f_{i,\bullet}} = f_{\bullet,j}, \quad j = 1,\ldots,s, \quad i = 1,\ldots,r,$$

y también

$$f_{i/j} = \frac{n_{i,j}}{n_{\bullet,j}} = \frac{\frac{n_{i,j}}{N}}{\frac{n_{\bullet,j}}{N}} = \frac{f_{i,\bullet} f_{\bullet,j}}{f_{\bullet,j}} = f_{i,\bullet}, \quad i = 1,\ldots,r, \quad j = 1,\ldots,s.$$

Estas igualdades significan que la distribución de frecuencias relativas de una variable no cambia conforme varia el valor de la otra.

Ejemplo 4.5 Sea la variable estadística (X,Y), donde X :"edad en años de 30 trabajadores" e Y :"salario anual en miles de euros de los 30 trabajadores". Suponga que la distribución de frecuencias absolutas es una de las tres que se indican en las tablas siguientes.

(a)

$X \setminus Y$	8	14	21	
20	5	0	0	5
26	0	20	0	20
33	0	0	5	5
	5	20	5	30

(b)

$X \setminus Y$	8	14	21	
20	3	1	0	4
26	0	18	2	20
33	1	1	4	6
	4	20	6	30

(c)

$X \setminus Y$	8	14	21	
20	3	3	3	9
26	4	4	4	12
33	3	3	3	9
	10	10	10	30

En el caso a), existe una relación funcional perfecta entre X e Y. Se tiene que la distribución de frecuencias absolutas de la variable estadística $Y/X = 20$ es

$$\{(y_j, n_{1,j})\}_{j=1,\ldots,3} : \{(8,5),(14,0),(21,0)\}.$$

Para la variable estadística $Y/X = 26$ esta distribución es

$$\{(y_j, n_{2,j})\}_{j=1,\ldots,3} : \{(8,0),(14,20),(21,0)\}.$$

Y en el caso de la variable estadística $Y/X=33$ se tiene que

$$\{(y_j,n_{3,j})\}_{j=1,\dots,3}:\{(8,0),(14,0),(21,5)\}.$$

Es decir: si $X=20$, siempre ocurre que $Y=8$; si $X=26$, siempre ocurre que $Y=14$; si $X=33$, siempre ocurre que $Y=21$. El salario crece en la misma magnitud que la edad de acuerdo con una relación lineal. Por otro lado, las distribuciones de frecuencias absolutas de las variables estadísticas $X/Y=8$, $X/Y=14$ y $X/Y=21$ son, respectivamente,

$$\{(x_i,n_{i,1})\}_{i=1,\dots,3}:\{(20,5),(26,0),(33,0)\},$$

$$\{(x_i,n_{i,2})\}_{i=1,\dots,3}:\{(20,0),(26,20),(33,0)\}$$

y

$$\{(x_i,n_{i,3})\}_{i=1,\dots,3}:\{(20,0),(26,0),(33,5)\},$$

lo que confirma la relación funcional perfecta entre X e Y.

En el caso b), el aumento de la edad produce generalmente un incremento salarial, pero la variación entre ambas variables no se ajusta a una relación funcional perfecta. Ahora bien, tampoco existe independencia estadística entre X e Y. De hecho, las distribuciones de frecuencias relativas de las variables estadísticas $Y/X=20$, $Y/X=26$ e $Y/X=33$ son

$$\{(y_j,f_{j/1})\}_{j=1,\dots,3}:\left\{\left(8,\frac{3}{4}\right),\left(14,\frac{1}{4}\right),(21,0)\right\},$$

$$\{(y_j,f_{j/2})\}_{j=1,\dots,3}:\left\{(8,0),\left(14,\frac{18}{20}\right),\left(21,\frac{2}{20}\right)\right\}$$

y

$$\{(y_j,f_{j/3})\}_{j=1,\dots,3}:\left\{\left(8,\frac{1}{6}\right),\left(14,\frac{1}{6}\right),\left(21,\frac{4}{6}\right)\right\},$$

respectivamente. De modo que el valor que toma la variable estadística X modifica la distribución de frecuencias de la variable estadística Y. En el mismo sentido, las distribuciones de frecuencias relativas de las variables estadísticas $X/Y=8$, $X/Y=14$ y $X/Y=21$, son

$$\{(x_i,f_{i/1})\}_{i=1,\dots,3}:\left\{\left(20,\frac{3}{4}\right),(26,0),\left(33,\frac{1}{4}\right)\right\},$$

$$\{(x_i,f_{i/2})\}_{i=1,\dots,3}:\left\{\left(20,\frac{1}{20}\right),\left(26,\frac{18}{20}\right),\left(33,\frac{1}{20}\right)\right\}$$

y

$$\{(x_i,f_{i/3})\}_{i=1,\dots,3}:\left\{(20,0),\left(26,\frac{2}{6}\right),\left(33,\frac{4}{6}\right)\right\}.$$

Por tanto, se vuelve a confirmar el sentido de la relación entre X e Y.

En el caso c), no se aprecia un sentido dominante de la relación entre ambas magnitudes. Se tiene que las distribuciones de frecuencias relativas de las variables $Y/X=20$, $Y/X=26$ e $Y/X=33$ son

$$\{(y_j, f_{j/1})\}_{j=1,\ldots,3} : \left\{\left(8,\frac{1}{3}\right),\left(14,\frac{1}{3}\right),\left(21,\frac{1}{3}\right)\right\},$$

$$\{(y_j, f_{j/2})\}_{j=1,\ldots,3} : \left\{\left(8,\frac{1}{3}\right),\left(14,\frac{1}{3}\right),\left(21,\frac{1}{3}\right)\right\}$$

y

$$\{(y_j, f_{j/3})\}_{j=1,\ldots,3} : \left\{\left(8,\frac{1}{3}\right),\left(14,\frac{1}{3}\right),\left(21,\frac{1}{3}\right)\right\}.$$

Por tanto, el cambio en el valor de la variable estadística X no modifica la distribución de frecuencias de la variable estadística Y, es decir, existe independencia estadística entre X e Y. La misma conclusión se obtiene si se observa que las distribuciones de frecuencias relativas de las variables $X/Y=8$, $X/Y=14$ y $X/Y=21$ son

$$\{(x_i, f_{i/1})\}_{i=1,\ldots,3} : \left\{\left(20,\frac{3}{10}\right),\left(26,\frac{4}{10}\right),\left(33,\frac{3}{10}\right)\right\},$$

$$\{(x_i, f_{i/2})\}_{i=1,\ldots,3} : \left\{\left(20,\frac{3}{10}\right),\left(26,\frac{4}{10}\right),\left(33,\frac{3}{10}\right)\right\}$$

y

$$\{(x_i, f_{i/3})\}_{i=1,\ldots,3} : \left\{\left(20,\frac{3}{10}\right),\left(26,\frac{4}{10}\right),\left(33,\frac{3}{10}\right)\right\}.$$

En definitiva, el grado de relación entre los componentes de una variable estadística bidimensional puede oscilar entre la independencia estadística y la relación funcional perfecta. En cualquier caso, es importante advertir la distinción entre dependencia estadística y causalidad, de modo que la relación estadística observada sea interpretada en sus justos términos y no se derive de ella una relación causa-efecto que, en todo caso, podría justificarse si existe una teoría que la sustente.

EJERCICIOS

4.1. Sea la variable estadística (X,Y), donde X :"edad en años de 100 trabajadores" e Y :"salario anual en miles de euros de los 100 trabajadores". Suponga que las edades se han agrupado en los intervalos $\{I_i\}_{i=1,2,3}$, tales que $I_1:(20,30]$, $I_2:(30,40]$ y $I_3:(40,50]$, mientras que para los salarios se han definido los intervalos $\{I_j\}_{j=1,2}$, tales que $I_1:(10,20]$ y $I_2:(20,30]$. Si la distribución de frecuencias absolutas es

$$\{(I_i,I_j;n_{i,j})\}_{i=1,2,3;j=1,2}:$$

$$:\{(I_1,I_1;20),(I_1,I_2;20),(I_2,I_1;20),(I_2,I_2;15),(I_3,I_1;15),(I_3,I_2;10)\}$$

exprese esta información en una tabla de doble entrada y represéntela mediante un diagrama de dispersión tridimensional.

4.2. Sea la variable estadística (X,Y), donde X:"horas semanales de estudio de 30 alumnos universitarios" e Y:"horas semanales de televisión de los 30 alumnos". Suponga que la distribución de frecuencias absolutas de esta variable es una de las que se expresa en las tablas siguientes:

$X \setminus Y$	7	14	21
7	0	0	10
14	0	12	0
21	8	0	0

$X \setminus Y$	7	14	21
7	0	2	8
14	4	8	0
21	4	2	2

(a) Obtenga las distribuciones marginales correspondientes a los dos casos. ¿Puede determinarse la distribución conjunta a partir de las marginales?

(b) Utilizando la segunda tabla de distribución de frecuencias de (X,Y), obtenga la distribución de frecuencias relativas de las variables $X/_{Y=7}$, $X/_{Y=14}$ y $X/_{Y=21}$. ¿Diría usted que los alumnos que ven más horas de televisión estudian menos?

4.3. Sea la variable estadística (X,Y), donde X:"porcentaje de hombres residentes en 100 municipios" e Y:"% de mujeres residentes en los 100 municipios". Suponga que los porcentajes de hombres y mujeres se han agrupado, respectivamente, en los intervalos $\{I_i\}_{i=1,2}$, tales que $I_1:(0,50]$ y $I_2:(50,100]$, y en los intervalos $\{I_j\}_{j=1,2}$, tales que $I_1:(0,50]$ y $I_2:(50,100]$. Suponga, además, que la distribución de frecuencias absolutas es

$$\{(I_i,I_j;n_{i,j})\}_{i=1,2;j=1,2}:\{(I_1,I_1;0),(I_1,I_2;60),(I_2,I_1;40),(I_2,I_2;0)\}.$$

¿Diría usted que X e Y son independientes? ¿Qué grado de dependencia estadística existe entre estas dos variables? ¿Existe sólo dependencia estadística?

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

5

Medidas características de distribuciones multidimensionales

En el capítulo precedente se señaló que la necesidad de superar el ámbito unidimensional quedaba justificada porque las variables estadísticas multidimensionales aportan información que no se obtiene a partir de las distribuciones marginales. Esa información añadida es la referida a la relación entre los componentes de la variable multidimensional. En este capítulo se estudian el grado y la forma de las relaciones entre dichos componentes. De nuevo, se prestará especial atención al caso bidimensional.

Por ejemplo, para obtener una conclusión más fundada acerca de la impresión existente sobre el efecto de la edad sobre el salario entre los trabajadores de determinada actividad, es preciso obtener una serie de medidas que caractericen la información conjunta sobre edades y salarios de estos individuos. Tales medidas se expresan, por lo general, como función de momentos, cuya definición se aporta en el primer epígrafe del capítulo. De este modo se obtiene información referida a las características que como variables estadísticas unidimensionales poseen cada uno de los componentes de una variable estadística bidimensional o multidimensional. Pero además existen determinadas medidas que tratan de reflejar el grado, e incluso el tipo, de dependencia estadística entre dichos componentes. Estas otras medidas serán objeto de estudio en los restantes epígrafes del capítulo.

5.1 MOMENTOS

La mayoría de las medidas características de una variable estadística multidimensional se definen, como en el caso unidimensional, en términos de momentos, que son funciones de los valores observados de la variable estadística. Sea una variable estadística bidimensional (X,Y) cuya distribución de frecuencias viene dada por

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

y

$$\mu_{0,2} = \sum_{i=1}^{r}\sum_{j=1}^{s}(y_j - \overline{y})^2 \frac{n_{i,j}}{N} = \sum_{j=1}^{s}(y_j - \overline{y})^2 \frac{n_{\bullet,j}}{N} = S_Y^2 .$$

Y, como se definirá más adelante, el momento central de órdenes 1 y 1, definido como

$$\mu_{1,1} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \overline{x})(y_j - \overline{y})\frac{n_{i,j}}{N},$$

es la covarianza de la variable estadística (X,Y), que se denotará por $S_{X,Y}$. Este momento puede expresarse en términos de los momentos respecto al origen. En concreto, se tiene que

$$\mu_{1,1} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \overline{x})(y_j - \overline{y})\frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i y_j - \overline{x}y_j - x_i\overline{y} + \overline{x}\,\overline{y})\frac{n_{i,j}}{N} =$$

$$= m_{1,1} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{x}\,\overline{y} = m_{1,1} - m_{1,0}m_{0,1}$$

Ejemplo 5.1 Sea la variable estadística (X,Y), donde X:"edad en años de 30 trabajadores" e Y:"salario anual en miles de euros de los 30 trabajadores". Suponga que la distribución de frecuencias absolutas es la que se indica en la tabla siguiente.

$X \setminus Y$	8	14	20	
20	3	3	3	9
26	4	4	4	12
32	3	3	3	9
	10	10	10	30

Entonces,

$$m_{1,0} = \sum_{i=1}^{3}\sum_{j=1}^{3} x_i \frac{n_{i,j}}{30} = \sum_{i=1}^{3} x_i \sum_{j=1}^{3}\frac{n_{i,j}}{30} = \sum_{i=1}^{3} x_i \frac{n_{i,\bullet}}{30} = 20\frac{9}{30} + 26\frac{12}{30} + 32\frac{9}{30} = 26 = \overline{x}.$$

Del mismo modo,

$$m_{0,1} = \sum_{j=1}^{3}\sum_{i=1}^{3} y_j \frac{n_{i,j}}{30} = \sum_{j=1}^{3} y_j \sum_{i=1}^{3}\frac{n_{i,j}}{30} = \sum_{j=1}^{3} y_j \frac{n_{\bullet,j}}{30} = 8\frac{10}{30} + 14\frac{10}{30} + 20\frac{10}{30} = 14 = \overline{y}.$$

También se tiene que

$$m_{2,0} = \sum_{i=1}^{r}\sum_{j=1}^{s} x_i^{\,2}\frac{n_{i,j}}{N} = 20^2\frac{9}{30} + 26^2\frac{12}{30} + 32^2\frac{9}{30} = 697.6$$

y

$$m_{0,2} = \sum_{i=1}^{r}\sum_{j=1}^{s} y_j^{\,2}\frac{n_{i,j}}{N} = 8^2\frac{10}{30} + 14^2\frac{10}{30} + 20^2\frac{10}{30} = 220,$$

mientras que

$$m_{1,1} = \sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j \frac{n_{i,j}}{N} = 364 .$$

También se tiene que

$$S_X^2 = \mu_{2,0} = \sum_{i=1}^{r} (x_i - 26)^2 \frac{n_{i,\bullet}}{30} = 21.6 ,$$

$$S_Y^2 = \mu_{0,2} = \sum_{j=1}^{s} (y_j - 14)^2 \frac{n_{\bullet,j}}{30} = 24$$

y

$$S_{X,Y} = \mu_{1,1} = m_{1,1} - m_{1,0} m_{0,1} = 364 - 364 = 0 .$$

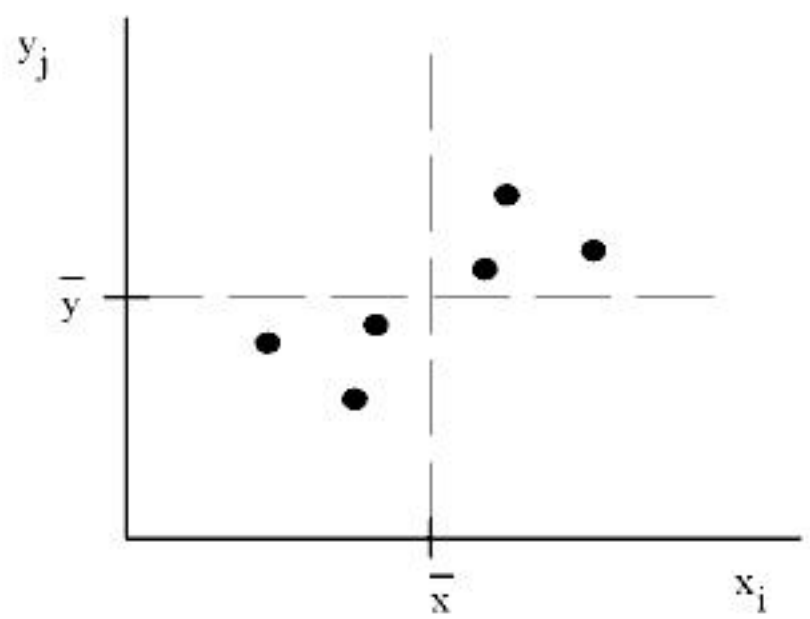
5.2 COVARIANZA Y COEFICIENTE DE CORRELACIÓN LINEAL

La covarianza y el coeficiente de correlación lineal son dos importantes indicadores del grado de relación lineal entre los componentes de una variable estadística bidimensional. Sea una variable estadística bidimensional (X,Y) cuya distribución de frecuencias viene dada por $\{(x_i, y_j; n_{i,j})\}_{i=1,\ldots,r; j=1,\ldots,s}$. Como ya se ha indicado, la covarianza se define como

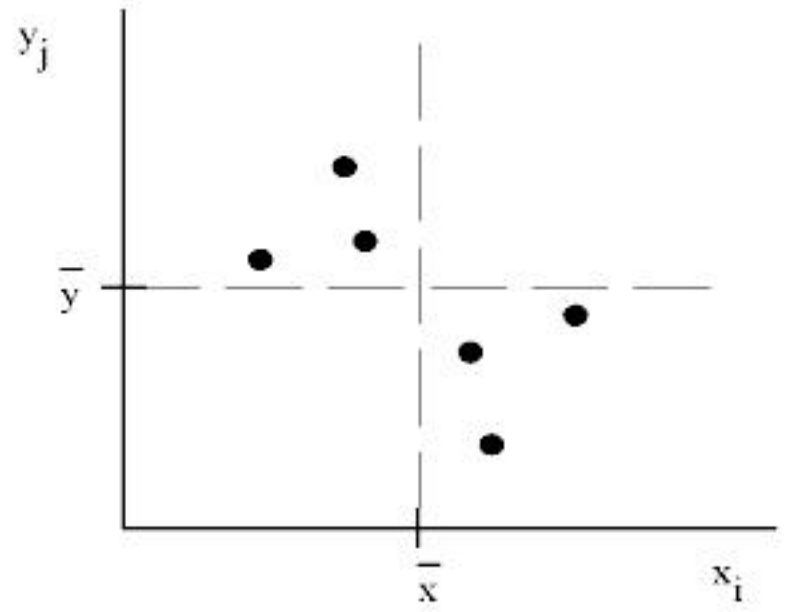
$$S_{X,Y} = \sum_{i=1}^{r}\sum_{j=1}^{s} (x_i - \bar{x})(y_j - \bar{y}) \frac{n_{i,j}}{N} .$$

Se trata, por tanto, de un promedio ponderado de los valores de la función $(x_i - \bar{x})(y_j - \bar{y})$ en cada uno de los pares de valores observados (x_i, y_j).

De este modo, si los valores observados de la variable estadística (X,Y) se distribuyen como se representa en el gráfico siguiente

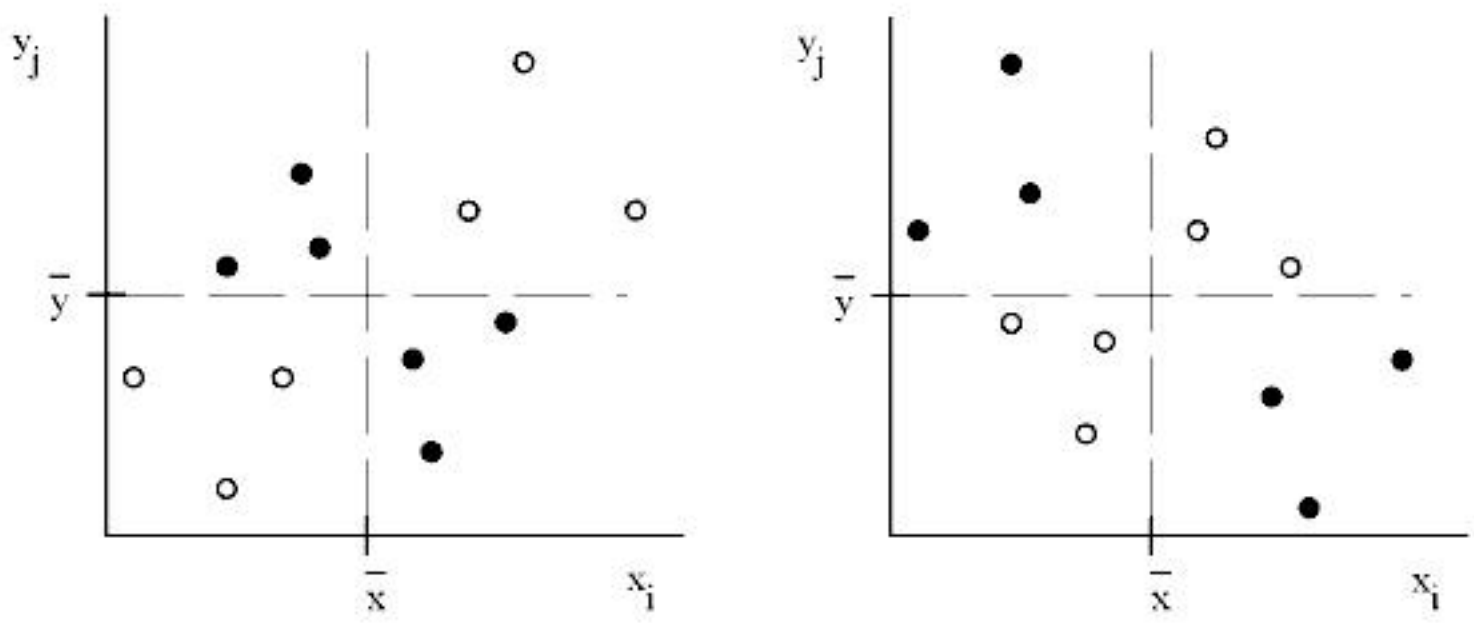


el signo de las diferencias $(x_i - \overline{x})$ e $(y_j - \overline{y})$ es siempre el mismo, de tal manera que la covarianza es un promedio de cantidades positivas y será entonces positiva. Nótese que, en este caso, cada vez que el valor de la variable estadística X está por encima de su media, el valor de la variable estadística Y también está por encima de su media. Del mismo modo, cada vez que el valor de la variable estadística X está por debajo de su media, el valor de la variable estadística Y también está por debajo de su media. Las dos variables tienden a covariar en el mismo sentido, es decir, existe una relación lineal directa entre X e Y. Sin embargo, si los valores observados de la variable estadística (X,Y) se distribuyen como se representa en el gráfico siguiente



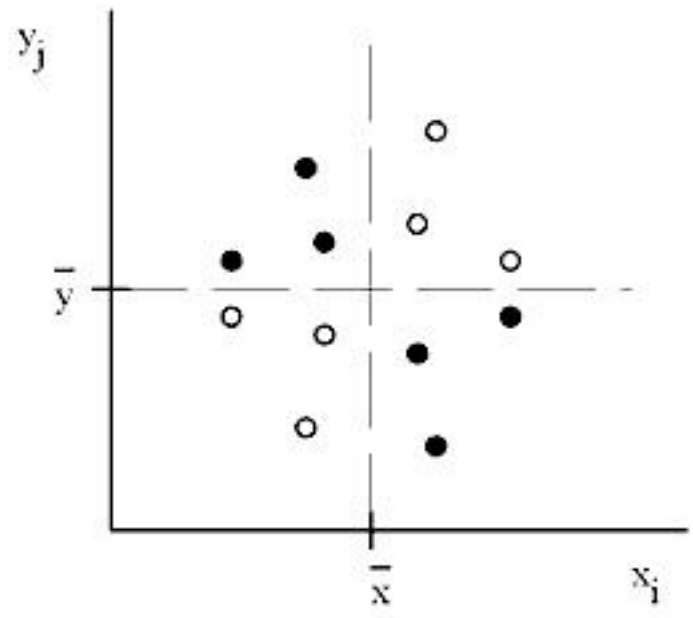
las diferencias $(x_i - \overline{x})$ e $(y_j - \overline{y})$ son siempre de signo contrario, de tal manera que la covarianza es un promedio de cantidades negativas y será entonces negativa. Nótese que, en este caso, cada vez que el valor de la variable estadística X está por encima de su media, el valor de la variable estadística Y está por debajo de la suya. Y cada vez que el valor de la variable estadística X está por debajo de su media, el valor de la variable estadística Y está, en cambio, por encima de su media. Ahora las dos variables tienden a covariar en sentido inverso, es decir, existe una relación lineal inversa entre X e Y.

Suponga que la distribución de frecuencias es una de las dos siguientes



y que todos los pares observados tengan la misma frecuencia. En el caso de la izquierda, la covarianza será positiva, ya que se están promediando, con igual ponderación,

valores positivos y negativos, pero la dimensión de los valores positivos es mayor. En cambio, en el caso de la derecha, la covarianza será negativa. Finalmente, si la distribución de frecuencias es la siguiente



y resulta que los pares correspondientes a los puntos no sombreados poseen mayor frecuencia, entonces la covarianza será positiva. Lo contrario ocurrirá si los puntos sombreados son los que poseen mayor frecuencia. En definitiva, el signo de la covarianza indica el sentido dominante de la relación lineal entre X e Y. Por otro lado, la magnitud de la covarianza será mayor cuando todos los pares son tales que los productos $(x_i - \bar{x})(y_j - \bar{y})$ son del mismo signo. En cambio, si ocurre que cuando el valor de la variable estadística X está por encima de su media, el valor de la variable estadística Y está en ocasiones por debajo de la suya y en otras por encima, la covarianza será un promedio de cantidades positivas y negativas y, por tanto, la magnitud de la covarianza será más pequeña. Por ello, la magnitud de la covarianza puede interpretarse como indicador de la intensidad de la relación lineal entre X e Y. Si no existe relación lineal, la covarianza será nula.

Ejemplo 5.2 Sea la variable estadística (X,Y), donde X :"edad en años de 30 trabajadores" e Y :"salario anual en miles de euros de los 30 trabajadores". Suponga que la distribución de frecuencias absolutas es la que se indica en las tablas siguientes.

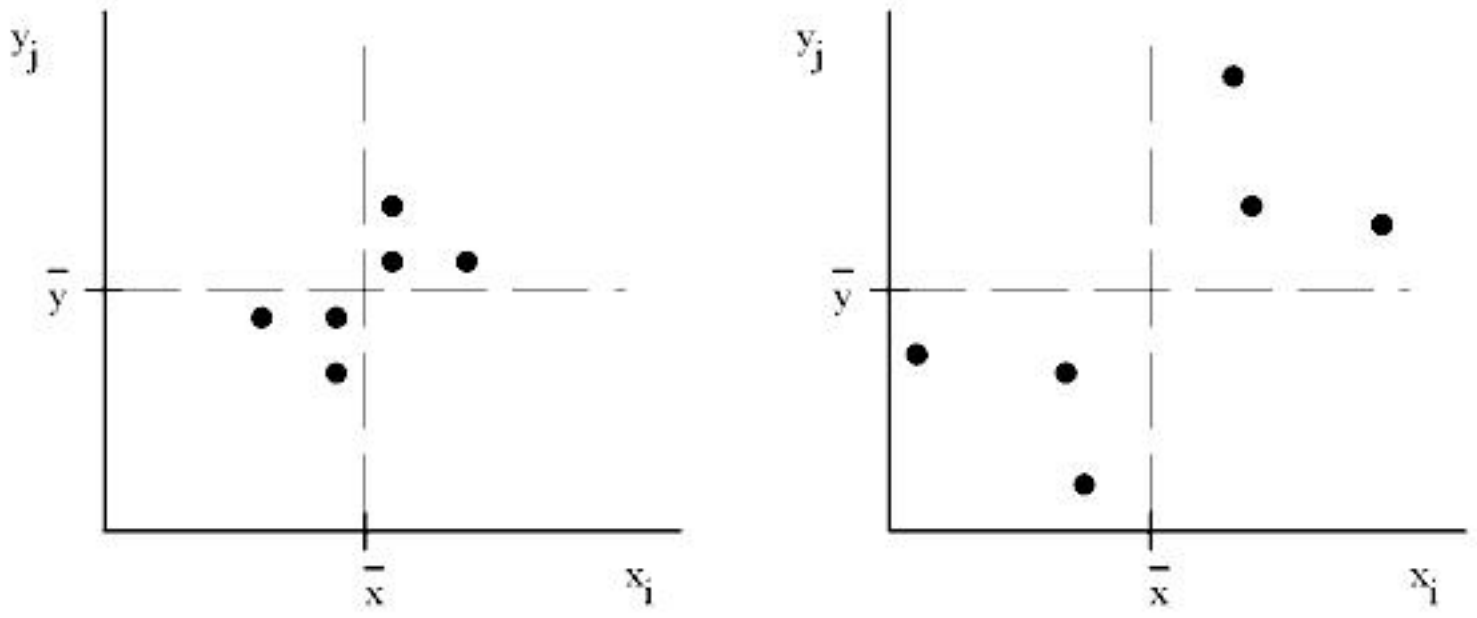
(a)

$X \backslash Y$	8	14	20	
20	5	0	0	5
26	0	20	0	20
32	0	0	5	5
	5	20	5	30

(b)

$X \backslash Y$	8	14	20	
20	4	1	0	5
26	0	18	2	20
32	1	1	3	5
	5	20	5	30

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.



En este sentido, resulta interesante buscar un coeficiente que mida el grado de intensidad de la relación lineal sin que la magnitud cambie por efecto de las varianzas. Este coeficiente es el coeficiente de correlación lineal, que se define como

$$\rho_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}.$$

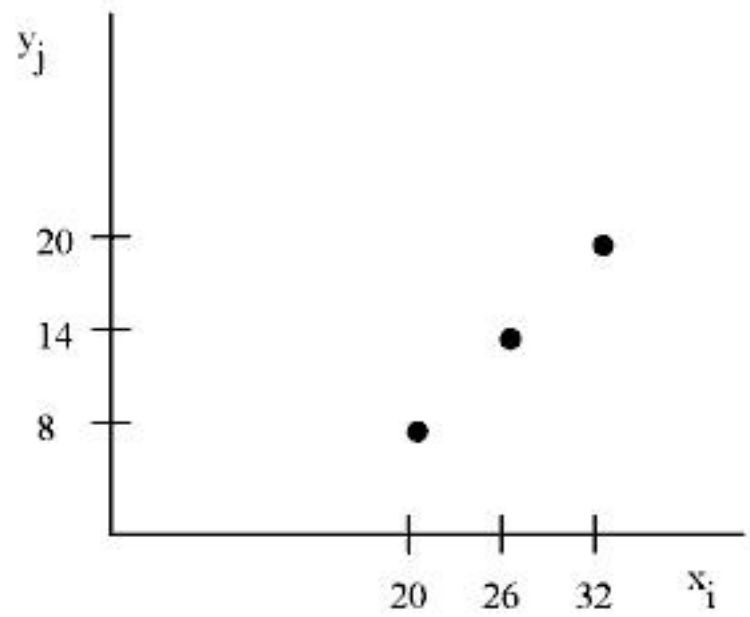
El signo de este coeficiente es el signo de la covarianza, de modo que este coeficiente indica el sentido de la relación lineal entre X e Y. Pero, además, este coeficiente es adimensional, es decir, no depende de la escala de medida de las variables. De modo que, si aumentan las varianzas de las dos variables, pero no se modifica la intensidad de la relación, la magnitud de este coeficiente será la misma. De hecho, el coeficiente de correlación entre X e Y es el mismo que entre Z^X y Z^Y, siendo Z^X y Z^Y las correspondientes variables estandarizadas. Nótese que

$$\rho_{Z^X,Z^Y} = S_{Z^X,Z^Y} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\frac{x_i-\bar{x}}{S_X}\right)\left(\frac{y_j-\bar{y}}{S_Y}\right)\frac{n_{i,j}}{N} = \frac{S_{X,Y}}{S_X S_Y} = \rho_{X,Y}.$$

Por tanto, la magnitud del coeficiente de correlación lineal es un indicador adimensional de la intensidad de la relación. En el caso de que no exista ningún tipo de relación lineal, la covarianza será nula y, por tanto, el coeficiente de correlación también será nulo. La otra situación extrema es aquélla en la que existe una relación lineal perfecta entre X e Y, es decir, las dos variables covarían siempre en sentido directo o inverso y, además, el cambio en una de las variables determina la magnitud del cambio que se produce en la otra variable. Si esto ocurre, todos los pares de valores (x_i, y_j) estarán situados sobre una recta de pendiente no nula, es decir, $n_{i,j} \neq 0$ si y sólo si el par (x_i, y_j) es tal que $y_j = a + bx_i$, con $b \neq 0$. Se tiene entonces que $\bar{y} = a + b\bar{x}$, de modo que

$$S_{X,Y} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\bar{x})b(x_i-\bar{x})\frac{n_{i,j}}{N} = b\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\bar{x})^2\frac{n_{i,j}}{N} = bS_X^2$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.



Puede advertirse que los pares observados de la variable estadística (X,Y) son tales que cuando la edad pasa de 20 a 26 años, el salario aumenta de 8 a 14 mil euros, mientras que cuando la edad pasa de 26 a 32 años, el salario aumenta de 14 a 20 mil euros. Es decir, existe una relación lineal perfecta de pendiente unitaria. Dado que existe una relación lineal perfecta y que la relación es directa, puede deducirse que $\rho_{X,Y}=1$. De hecho, ya se había obtenido que $\overline{x}=26$, $\overline{y}=14$ y $S_{XY}=12$, de modo que

$$S_X^2=\mu_{2,0}=\sum_{i=1}^{r}(x_i-26)^2\frac{n_{i\bullet}}{30}=12$$

y

$$S_Y^2=\mu_{0,2}=\sum_{j=1}^{s}(y_j-14)^2\frac{n_{\bullet j}}{30}=12\,.$$

Entonces

$$\rho_{X,Y}=\frac{S_{X,Y}}{S_X S_Y}=\frac{12}{\sqrt{12\bullet 12}}=1\,.$$

5.3 CONCEPTO ESTADÍSTICO DE REGRESIÓN

La relación entre variables estadísticas puede ser lineal o de cualquier otro tipo y, por ello, conviene diseñar instrumentos que, contando con la ayuda que proporciona la representación gráfica, permitan detectar cuál es la expresión analítica que mejor capta tal dependencia. Pues bien, ese instrumento es la regresión, es decir, la línea que une las medias condicionadas. Para que este concepto resulte operativo, sobre todo a efectos predictivos, conviene formular una hipótesis sobre el tipo de función que mejor se ajusta a la forma de la relación observada entre las variables. Tanto el concepto de regresión en términos de medias condicionadas como los ajustes funcionales se introducen a continuación.

5.3.1. Medias condicionadas

Sea una variable estadística bidimensional (X,Y) cuya distribución de frecuencias es

$$\{(x_i, y_j; n_{i,j})\}_{i=1,\dots,r;\, j=1,\dots,s}.$$

Sean, además,

$$\{(y_j, n_{i,j})\}_{j=1,\dots,s},\ i=1,\dots,r,$$

y

$$\{(x_i, n_{i,j})\}_{i=1,\dots,r},\ j=1,\dots,s,$$

las distribuciones de frecuencias de las variables estadísticas $Y/X=x_i$, $i=1,\dots,r$, y $X/Y=y_j$, $j=1,\dots,s$, respectivamente. Cada una de estas distribuciones condicionadas corresponde a una variable estadística unidimensional, que tendrá su correspondiente media. Es decir,

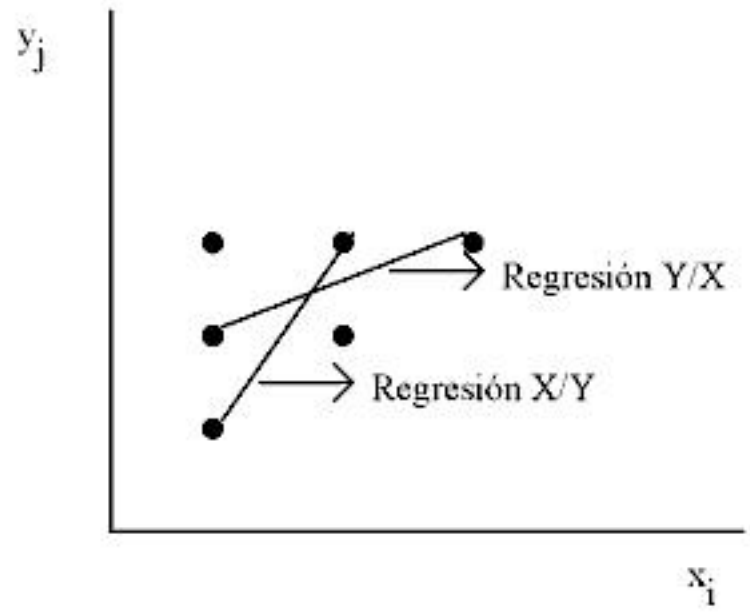
$$\bar{y}/_{X=x_i} = \sum_{j=1}^{s} y_j \frac{n_{i,j}}{n_{i,\bullet}},\quad i=1,\dots,r,$$

y

$$\bar{x}/_{Y=y_j} = \sum_{i=1}^{r} x_i \frac{n_{i,j}}{n_{\bullet,j}},\quad j=1,\dots,s.$$

Entonces, la observación de los cambios en las medias de las variables $Y/X=x_i$ cuando cambia el valor x_i muestra la forma en que los cambios en X producen cambios en Y. Del mismo modo, la observación de los cambios en las medias de las variables $X/Y=y_j$ cuando cambia el valor y_j muestra la forma en que los cambios en Y producen cambios en X. Pues bien, la línea de regresión de Y sobre X es la línea que une las medias de las variables $Y/X=x_i$, $i=1,\dots,r$, es decir, la línea que une los puntos $\left(x_i, \bar{y}/_{X=x_i}\right)$, $i=1,\dots,r$. Mientras que la línea de regresión de X sobre Y es la línea que une las medias de las variables $X/Y=y_j$, $j=1,\dots,s$, es decir, la línea que une los puntos $\left(y_j, \bar{x}/_{Y=y_j}\right)$, $j=1,\dots,s$.

En el gráfico siguiente, se muestran las líneas de regresión de Y sobre X, regresión Y/X, y de X sobre Y, regresión X/Y, para una distribución determinada.



Las líneas de regresión anteriores muestran la existencia de una relación lineal entre X e Y. Si alguna de las líneas de regresión es constante, no existe relación lineal entre las variables. Supóngase que $\overline{y}/_{X=x_i} = k$, $i = 1,\ldots,r$. Entonces

$$\overline{y} = \sum_{i=1}^{r}\sum_{j=1}^{s} y_j \frac{n_{i,j}}{N} = \sum_{i=1}^{r} \frac{n_{i,\bullet}}{N} \sum_{j=1}^{s} y_j \frac{n_{i,j}}{n_{i,\bullet}} = \sum_{i=1}^{r} \frac{n_{i,\bullet}}{N} k = k \sum_{i=1}^{r} \frac{n_{i,\bullet}}{N} = k ,$$

mientras que

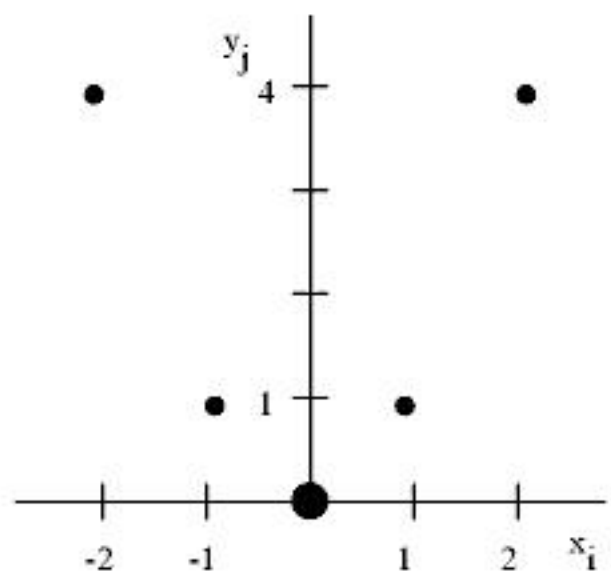
$$m_{1,1} = \sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j \frac{n_{i,j}}{N} = \sum_{i=1}^{r} x_i \frac{n_{i,\bullet}}{N} \sum_{j=1}^{s} y_j \frac{n_{i,j}}{n_{i,\bullet}} = \sum_{i=1}^{r} x_i \frac{n_{i,\bullet}}{N} k = k \sum_{i=1}^{r} x_i \frac{n_{i,\bullet}}{N} = k\overline{x} ,$$

de modo que

$$S_{XY} = m_{1,1} - m_{1,0}m_{0,1} = k\overline{x} - \overline{x}k = 0 .$$

Nótese que si la media de Y no crece ni decrece cuando cambia el valor de X, no existe tendencia de covariación lineal directa ni inversa. Aunque puede existir relación de otro tipo.

Ejemplo 5.4 Sea una variable estadística bidimensional (X,Y) con distribución de frecuencias unitaria tal que



En este caso, se tiene que $\overline{x}/_{Y=y_j}=0$, $j=1,2,3$. Por tanto, no existe relación lineal entre X e Y. Sin embargo, puede advertirse que $\overline{y}/_{X=x_i}=x_i^2$, $i=1,2,3,4,5$. De hecho, todos los pares observados (x_i,y_j) son tales que $y_j=x_i^2$, es decir, existe una relación funcional perfecta entre X e Y, pero de tipo parabólico.

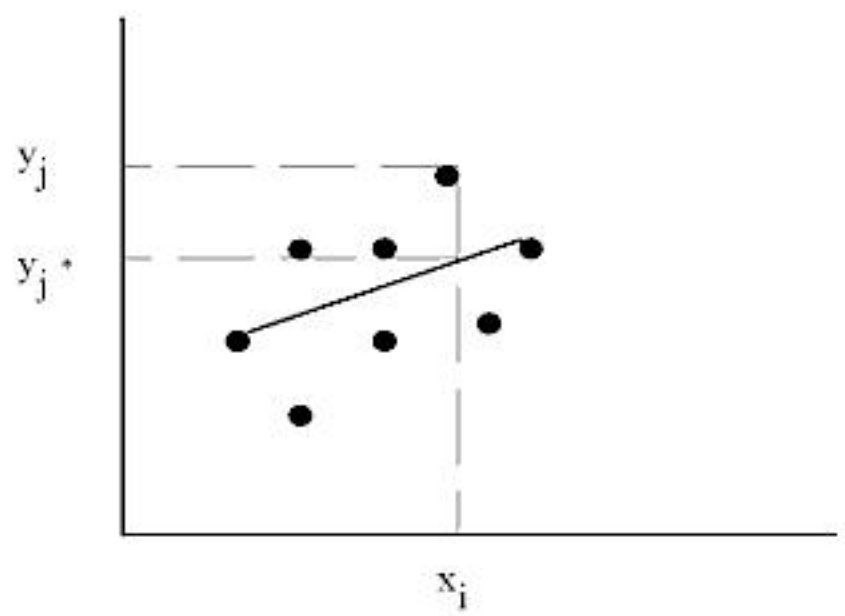
El ejemplo anterior ilustra también el hecho de que la ausencia de correlación lineal no implica independencia estadística, aunque, como se ha explicado, la independencia estadística sí implica ausencia de relación lineal.

5.3.2. Ajustes funcionales por mínimos cuadrados

El cálculo de las medias condicionadas puede ayudar a identificar la forma de la relación entre las variables X e Y. Pero, una vez elegida la forma funcional que, en principio, parece ajustarse mejor a los datos, es necesario obtener los parámetros, desconocidos a priori, que intervienen en esa formulación. En este sentido, puede acudirse al método de los mínimos cuadrados como criterio de ajuste de relaciones lineales y no lineales.

Supóngase que se asume que la variable X puede explicar los valores de la variable Y de acuerdo con la forma funcional f. Es decir, $y_j^*=f(x_i)$, donde y_j^* es el valor ajustado de la variable explicada cuando la variable explicativa toma el valor x_i. En la función f, ya sea lineal, parabólica o de cualquier otro tipo, intervendrán parámetros que habrá que determinar para obtener los valores y_j^*. El criterio de los mínimos cuadrados consiste en elegir los valores de los parámetros que minimizan la suma de cuadrados de los errores o residuos del ajuste, $e_{i,j}$, es decir, los cuadrados de las diferencias entre los valores observados y_j y los valores ajustados y_j^*.

Supóngase que f es una función lineal, entonces el gráfico siguiente expresa los errores del ajuste que se cometen cuando se efectúa el ajuste lineal.



You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

y

$$\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{i,j} = a\sum_{i=1}^{r}\sum_{j=1}^{s} x_i n_{i,j} + b\sum_{i=1}^{r}\sum_{j=1}^{s} x_i^{\,2} n_{i,j} .$$

Dividiendo por N en ambos miembros de la primera ecuación, se tiene que

$$\overline{y} = a + b\overline{x} ,$$

es decir,

$$a = \overline{y} - b\overline{x} ,$$

y sustituyendo en la segunda ecuación

$$\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{i,j} - \overline{y}\sum_{i=1}^{r}\sum_{j=1}^{s} x_i n_{i,j} = b\sum_{i=1}^{r}\sum_{j=1}^{s} x_i^{\,2} n_{i,j} - b\overline{x}\sum_{i=1}^{r}\sum_{j=1}^{s} x_i n_{i,j} .$$

Si en esta última ecuación se dividen ambos miembros por N, resulta que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j \frac{n_{i,j}}{N} - \overline{y}\sum_{i=1}^{r}\sum_{j=1}^{s} x_i \frac{n_{i,j}}{N} = b\sum_{i=1}^{r}\sum_{j=1}^{s} x_i^{\,2} \frac{n_{i,j}}{N} - b\overline{x}\sum_{i=1}^{r}\sum_{j=1}^{s} x_i \frac{n_{i,j}}{N} ,$$

es decir,

$$m_{1,1} - \overline{y}\,\overline{x} = b\left(m_2 - \overline{x}^2\right),$$

de modo que,

$$b = \frac{S_{X,Y}}{S_X^2} .$$

Los valores de los parámetros a y b que anulan las derivadas parciales corresponden al mínimo de la función si y sólo si la matriz *hessiana* de segundas derivadas parciales con respecto a los parámetros a y b, evaluadas en los valores identificados para estos parámetros, es definida positiva. Esta segunda condición se verifica siempre en los problemas de ajuste por mínimos cuadrados tratados en este texto[1].

El parámetro b se denomina coeficiente de regresión. El signo de este parámetro viene determinado por el signo de la covarianza. Nótese que si este parámetro es positivo, los valores de las variables estadísticas X e Y tienden a moverse en el mismo sentido y la covarianza será positiva. Por el contrario, si el parámetro es negativo, las dos variables tienden a moverse en sentido inverso, de modo que la covarianza será negativa. Por otra parte, el parámetro b es la pendiente de la recta $y_j^* = a + bx_i$, es decir, $\dfrac{\partial y_j^*}{\partial x_i} = b$, de modo que $dy_j^* = b dx_i$. Como muestra el gráfico siguiente, la

[1] Véase Martín y Martín (1989: 130) y Casas y Santos (1995: 148-149).

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Puede advertirse que existe una relación lineal perfecta de pendiente unitaria. De hecho, si se ajusta por mínimos cuadrados la función lineal $y_j^* = a + bx_i$, se tiene que

$$b = \frac{S_{xy}}{S_x^2} = 1,$$

mientras que

$$a = \overline{y} - b\overline{x} = 14 - 26 = -12.$$

En el ejemplo anterior, el ajuste es perfecto. Existe una relación lineal perfecta y, por tanto, no existen errores en el ajuste. Pero, en general, el ajuste por mínimos cuadrados no es tan bueno y aparecen residuos o errores del ajuste, siendo conveniente evaluar la bondad del ajuste.

5.4 MEDIDAS DE BONDAD DE AJUSTE

Una vez que la forma de la relación entre variables ha sido completamente especificada, puede evaluarse la bondad con que la regresión obtenida se ajusta a los datos observados. Desde este punto de vista, pueden utilizarse los coeficientes de determinación, que expresan la proporción de la variabilidad de una variable que resulta explicada por la variable explicativa en el ajuste obtenido.

Sea una variable estadística bidimensional (X,Y) cuya distribución de frecuencias es

$$\{(x_i, y_j; n_{i,j})\}_{i=1,\ldots,r;\, j=1,\ldots,s}.$$

Supóngase que se efectúa el ajuste por mínimos cuadrados de la relación funcional $y_j^* = f(x_i)$. El ajuste será mejor cuanto más se aproximen los valores observados y_j y los valores ajustados y_j^*. En este sentido, la bondad del ajuste estará inversamente relacionada con la suma

$$\sum_{i=1}^{r}\sum_{j=1}^{s}(y_j - y_j^*)^2 n_{i,j}.$$

La suma anterior recoge la variabilidad no explicada por el ajuste, pero parece conveniente expresar esta variabilidad como proporción de la variabilidad de la variable que se pretende explicar. La variabilidad de la variable explicada puede medirse mediante la varianza de la variable Y, S_Y^2, mientras que la variabilidad no explicada puede definirse como la varianza residual, S_r^2, es decir, como la varianza de los residuos del ajuste dada por

$$S_r^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}(y_j - y_j^*)^2 \frac{n_{i,j}}{N}.$$

Nótese que si en la función ajustada se introduce término independiente, la media de los residuos será nula.

Entonces, como coeficiente de bondad de ajuste puede emplearse el denominado coeficiente de determinación general, R^2, definido como

$$R^2 = 1 - \frac{S_r^2}{S_Y^2}.$$

Si el ajuste es perfecto, $S_r^2 = 0$, de modo que $R^2 = 1$. Y cuanto peor sea el ajuste, mayor será la varianza residual, de modo que el coeficiente R^2 será más pequeño. En general, y salvo que en la relación funcional no se incluya un término independiente, $S_r^2 \leq S_Y^2$, de modo que el coeficiente de determinación será no negativo. En el caso de que la varianza residual coincida con la varianza de la variable explicada, es decir, cuando el ajuste no consigue explicar nada de la variabilidad de la variable Y, se tiene que $R^2 = 0$.

En el caso de un ajuste lineal especificado con término independiente, $y_j^* = a + bx_i$, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación lineal. Nótese que, en este caso,

$$S_r^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - y_j^*\right)^2 \frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - (a+bx_i)\right)^2 \frac{n_{i,j}}{N} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\left(y_j - \overline{y}\right) - b\left(x_i - \overline{x}\right)\right)^2 \frac{n_{i,j}}{N} = S_Y^2 + b^2 S_X^2 - 2bS_{X,Y} =$$

$$= S_Y^2 + \left(\frac{S_{X,Y}}{S_X^2}\right)^2 S_X^2 - 2\frac{S_{X,Y}}{S_X^2} S_{X,Y} = S_Y^2 - \frac{{S_{X,Y}}^2}{S_X^2}$$

Por tanto,

$$R^2 = 1 - \frac{S_r^2}{S_Y^2} = 1 - \frac{S_Y^2 - \frac{{S_{X,Y}}^2}{S_X^2}}{S_Y^2} = \frac{{S_{X,Y}}^2}{S_X^2 S_Y^2} = {\rho_{X,Y}}^2.$$

Además, en el caso del ajuste lineal con término independiente, el coeficiente de determinación puede expresarse de otra forma. Teniendo en cuenta que

$$y_j = y_j^* + e_{i,j},$$

de modo que,

$$y_j - \overline{y} = y_j^* - \overline{y} + e_{i,j},$$

la varianza de la variable explicada Y puede expresarse como

$$S_Y^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - \overline{y}\right)^2 \frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(\left(y_j{}^* - \overline{y}\right) + e_{i,j}\right)^2 \frac{n_{i,j}}{N} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j{}^* - \overline{y}\right)^2 \frac{n_{i,j}}{N} + \sum_{i=1}^{r}\sum_{j=1}^{s}\left(e_{i,j}\right)^2 \frac{n_{i,j}}{N} + 2\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j{}^* - \overline{y}\right)e_{i,j}\frac{n_{i,j}}{N}$$

Y, dado que las condiciones que determinan los parámetros a y b implican que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} e_{i,j}\frac{n_{i,j}}{N} = 0$$

y

$$\sum_{i=1}^{r}\sum_{j=1}^{s} x_i e_{i,j}\frac{n_{i,j}}{N} = 0\,,$$

resulta que

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j{}^* - \overline{y}\right)e_{i,j}\frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(a + bx_i - \overline{y}\right)e_{i,j}\frac{n_{i,j}}{N} =$$

$$= a\sum_{i=1}^{r}\sum_{j=1}^{s} e_{i,j}\frac{n_{i,j}}{N} + b\sum_{i=1}^{r}\sum_{j=1}^{s} x_i e_{i,j}\frac{n_{i,j}}{N} - \overline{y}\sum_{i=1}^{r}\sum_{j=1}^{s} e_{i,j}\frac{n_{i,j}}{N} = 0$$

Por tanto,

$$S_Y^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j{}^* - \overline{y}\right)^2 \frac{n_{i,j}}{N} + \sum_{i=1}^{r}\sum_{j=1}^{s}\left(e_{i,j}\right)^2 \frac{n_{i,j}}{N}\,,$$

y teniendo en cuenta que

$$\overline{y}^* = \sum_{i=1}^{r}\sum_{j=1}^{s} y_j{}^* \frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - e_{i,j}\right)\frac{n_{i,j}}{N} = \sum_{i=1}^{r}\sum_{j=1}^{s} y_j \frac{n_{i,j}}{N} - \sum_{i=1}^{r}\sum_{j=1}^{s} e_{i,j}\frac{n_{i,j}}{N} = \overline{y}\,,$$

se tiene que

$$S_Y^2 = S_{Y^*}^2 + S_r^2\,.$$

Así pues, el coeficiente de determinación puede escribirse en el caso lineal como

$$R^2 = 1 - \frac{S_r^2}{S_Y^2} = 1 - \frac{S_Y^2 - S_{Y^*}^2}{S_Y^2} = \frac{S_{Y^*}^2}{S_Y^2}\,,$$

es decir, el coeficiente de determinación indica la proporción de la variabilidad de la variable Y que resulta explicada por la regresión lineal.

Finalmente, conviene señalar que los ajustes funcionales resultan muy útiles a efectos predictivos; es decir, con ellos es posible pronosticar el valor de la variable explicada a partir del conocimiento de los valores de las variables explicativas. Ahora bien, la confianza de la predicción no puede derivarse exclusivamente de la bondad obtenida en el ajuste para los datos observados, sino que dicha confianza está condicionada a una hipótesis de estabilidad en la relación. En este sentido, es importante considerar la posibilidad de que la relación obtenida sea una relación espúrea, es decir, tal que la correlación estadística no se corresponda, sin embargo, con una relación de causalidad, como ya se adelantó en el capítulo anterior.

Ejemplo 5.6 Sea otra vez la variable estadística (X,Y), donde X: "edad en años de 30 trabajadores" e Y :"salario anual en miles de euros de los 30 trabajadores". Suponga que la distribución de frecuencias absolutas es la que se indica en la tabla siguiente.

$X \setminus Y$	8	14	20	
20	5	0	0	5
26	0	20	0	20
32	0	0	5	5
	5	20	5	30

Se había comprobado que $\overline{x}=26$, $\overline{y}=14$, $S_X^2=12$, $S_Y^2=12$ y $S_{XY}=12$. Y en este caso, si se ajustaba por mínimos cuadrados la función lineal $y_j^*=a+bx_i$, se obtenía una relación lineal perfecta dada por $y_j^*=-12+x_i$. Dado que el ajuste es perfecto, se tiene que $S_r^2=0$ y, por tanto, $R^2=1$. Este mismo valor se obtiene si se expresa el coeficiente de determinación como el cuadrado del coeficiente de correlación lineal, es decir,

$$R^2=\frac{S_{XY}{}^2}{S_X^2 S_Y^2}=\frac{12^2}{12\bullet 12}=1\,.$$

A partir del ajuste $y_j^*=-12+x_i$, es posible predecir el salario anual de un trabajador una vez conocida su edad. Por ejemplo, si $x_i=30$ entonces $y_j^*=18$. Es decir, se predice que un trabajador con 30 años obtendrá un sueldo de 18 mil euros anuales.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

mínimo entre los residentes en los 30 municipios" e Y:"porcentaje de votantes de izquierda en los 30 municipios". Los resultados del análisis efectuado se muestran en la tabla siguiente, que recoge la distribución de frecuencias absolutas de la variable (X,Y).

$X \setminus Y$	40	50	60
5	5	4	1
10	3	3	2
15	1	2	3
20	1	1	4

(a) Represente los pares observados (x_i,y_j) mediante un diagrama de dispersión. A partir del gráfico, ¿diría usted que ambos porcentajes tienden a moverse en el mismo sentido o en sentido inverso?

(b) Calcule la covarianza y el coeficiente de correlación lineal.

(c) Obtenga las medias de las variables estadísticas $Y/X=5$, $Y/X=10$, $Y/X=15$, $Y/X=20$. Represente gráficamente la línea de regresión de Y sobre X.

(d) De acuerdo con el resultado del apartado anterior, si la relación entre ambas variables se recoge mediante una función lineal del tipo $y_j^* = a + bx_i$, ¿de qué signo será la pendiente de la recta ajustada?

(e) Obtenga los valores de los parámetros a y b de la función lineal anterior si se efectúa el ajuste por mínimos cuadrados.

(f) De acuerdo con los resultados del apartado anterior, ¿en cuánto puede predecirse que se incrementará el porcentaje de votantes de izquierda en un municipio cuando el porcentaje de votantes con ingresos inferiores al salario mínimo aumenta en 5 puntos porcentuales?

(g) Evalúe la bondad del ajuste obtenido. ¿Qué proporción de la variabilidad del porcentaje de voto de izquierdas no queda explicada por el modelo lineal propuesto? ¿Es fiable la predicción obtenida en el apartado anterior?

5.5. Un sociólogo argumenta que la baja tasa de natalidad en la sociedad española es fruto de decisiones individuales que dependen de las posibilidades económicas del núcleo familiar. Este sociólogo ha estudiado el número de hijos y los ingresos de un conjunto de 100 familias. Sea la variable estadística (X,Y), donde X:"ingresos familiares anuales en miles de euros de cada una de las 100 familias" e Y:"número de hijos de las 100 familias". Y sea $\{(x_i,y_i)\}_{i=1,\ldots,100}$ el conjunto de pares de valores observados de la variable (X,Y) para las 100 familias. A partir del análisis efectuado se ha encontrado que $\sum_{i=1}^{100} x_i = 1694$, $\sum_{i=1}^{100} y_i = 200$, $\sum_{i=1}^{100} x_i y_i = 3506$, $\sum_{i=1}^{100} x_i^2 = 32572$, $\sum_{i=1}^{100} y_i^2 = 810$.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

6

Series temporales y números índices

En este capítulo se concede especial relevancia al factor tiempo como referencia para observar la evolución de una magnitud. El carácter dinámico propio de la actividad social implica que en muchas ocasiones no sea posible obviar el elemento temporal que, por el contrario, constituye un aspecto clave en la toma de decisiones. Así ocurre cuando un individuo en edad de trabajar decide seguir estudiando con la esperanza de obtener mayores ingresos en el futuro. El factor tiempo también está presente en la mente de un trabajador que planifica sus vacaciones para la época estival. Supóngase, por ejemplo, que, con objeto de diseñar políticas sociales, se examina la tasa de paro de los residentes en una zona turística. En este último caso, para evaluar la situación de los desempleados es importante tener en cuenta la evolución a largo plazo de la tasa de paro; pero también debe tomarse conciencia de que en función de la afluencia de turistas en cada estación, la tasa de paro puede experimentar notables fluctuaciones dentro de cada año. Consideraciones de este tipo pueden efectuarse, desde una perspectiva descriptiva, usando los procedimientos relacionados con el denominado análisis clásico de series temporales, que se explican en el primer epígrafe del capítulo.

Supóngase ahora que una organización sindical y una asociación empresarial pertenecientes a una determinada industria —entendida ésta como el conjunto de empresas que participan en el mercado de un cierto producto—, discuten sobre el incremento salarial que debe considerarse justo para el año en curso. Previsiblemente, se atenderá poco a los incrementos reales de productividad en el sector en cuestión y los argumentos que se esgriman girarán en torno a la necesidad de mantener el poder adquisitivo en función de la evolución del índice de precios de consumo. En la discusión aparece de nuevo el factor tiempo, pero ahora como referencia sobre la que medir la variación de una magnitud. El segundo epígrafe del capítulo aporta los instrumentos necesarios para poder formular un juicio crítico sobre los argumentos defendidos por las partes en conflicto. Concretamente, se exponen los números índices como herramienta para medir el porcentaje de variación de una magnitud entre dos puntos, generalmente instantes de tiempo, dados.

6.1 CONCEPTO DE SERIE TEMPORAL Y ANÁLISIS DE SUS COMPONENTES

En este epígrafe se presenta el concepto de serie temporal y se ofrecen las definiciones de sus componentes, así como de los modelos de integración de éstos. Una serie temporal es un conjunto de observaciones referidas a una magnitud y ordenadas en el tiempo. Sea X la magnitud bajo estudio y sea x_t el valor de la magnitud X en el instante del tiempo t. Entonces el conjunto $\{x_t\}_{t=1,\ldots,T} : \{x_1,\ldots,x_T\}$ es una serie temporal.

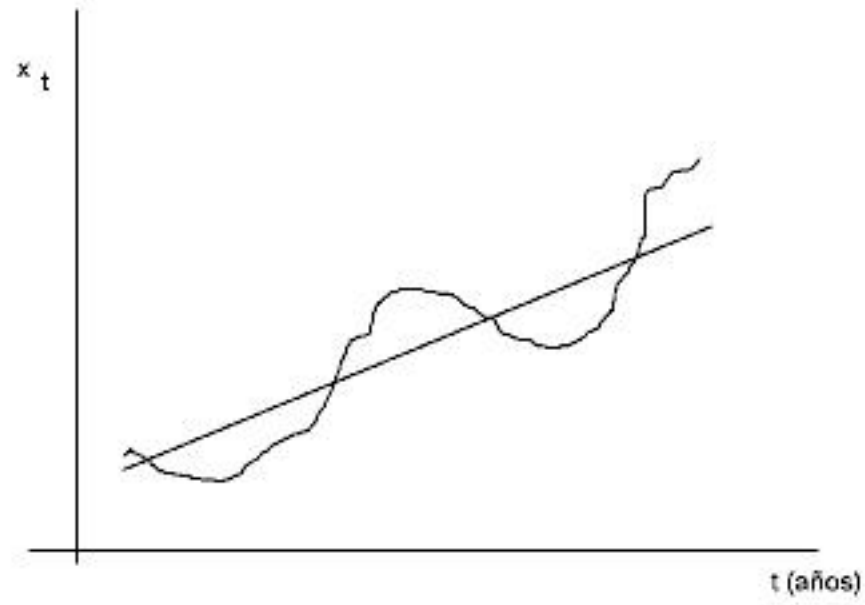
Las series temporales pueden ser analizadas con una finalidad descriptiva, si sólo se pretende describir el comportamiento registrado en el pasado, explicativa, si se intenta probar estadísticamente la existencia de relaciones dinámicas causa-efecto entre variables, o predictiva, cuando el objetivo es reducir el grado de incertidumbre sobre el futuro a partir del conocimiento del pasado. Para comprender el modo en que una serie temporal puede ser útil en alguno de estos sentidos, conviene, en primer lugar, identificar los componentes que dan lugar a los valores observados en el tiempo, y en segundo lugar, considerar diferentes modelos tanto de cada uno de dichos componentes como del modo en que éstos interactúan.

6.1.1. Componentes de una serie y esquemas de combinación

Tradicionalmente, se ha considerado que el comportamiento en el tiempo de una magnitud es el resultado de la dinámica que, conjuntamente, conduce la evolución de cuatro componentes básicos: tendencia, ciclo, variación estacional y componente irregular.

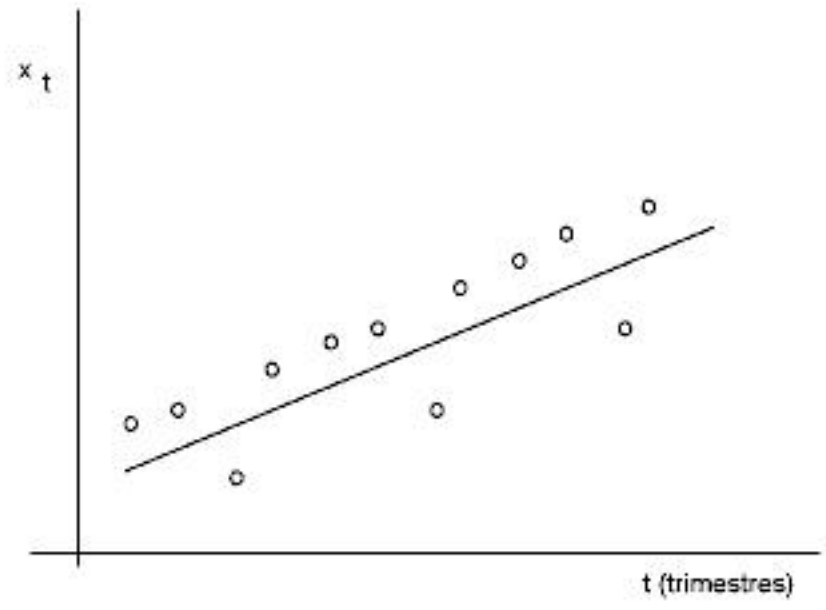
El componente tendencial, T_t, representa la conducta a largo plazo de la serie. Es decir, trata de reflejar hacia dónde tiende la serie. Las variaciones cíclicas, C_t, pueden considerarse oscilaciones más o menos regulares y periódicas en torno a la conducta de largo plazo o tendencial que se completan o compensan en un periodo largo, generalmente de varios años. En este sentido, una tendencia no lineal y una variación cíclica pueden llegar a ser indistinguibles, especialmente si la longitud temporal de la serie es reducida. De ahí que, en ocasiones, se hable de componente tendencia-ciclo.

Ejemplo 6.1 La evolución del número de coches vendidos anualmente en una determinada región, $\{x_t\}$, puede representarse como se indica en el gráfico siguiente. Tal vez, en el largo plazo esta magnitud tiende a crecer, pero en los años en los que la economía de la región se estanca se reducen las ventas de automóviles, de forma que pueden aparecer comportamientos cíclicos.



Por su parte, las fluctuaciones estacionales, S_t, son también oscilaciones en torno a la tendencia, pero se diferencian de las cíclicas en que el periodo en el que las primeras se completan o compensan es inferior o igual al año. Por último, el componente irregular o residual, ε_t, recoge las variaciones impredecibles que se producen en el corto plazo sin responder a ningún patrón sistemático.

Ejemplo 6.2 En el caso de la tasa trimestral de paro de una zona turística en decadencia que recibe casi todos los turistas en verano, las observaciones trimestrales, $\{x_t\}$, pueden revelar un cierto incremento a largo plazo. Y en torno a este movimiento tendencial existen variaciones estacionales que, como se representa en el gráfico siguiente, reflejan el aumento de la demanda de trabajo, y la correspondiente reducción en la tasa de paro, en la estación veraniega con respecto a los niveles de las otras estaciones del año.



La evolución conjunta de los cuatro componentes mencionados y su interacción atendiendo a un modelo determinado da lugar a los valores de la serie temporal, es decir, $x_t = f(T_t, C_t, S_t, \varepsilon_t)$. En general, se asume que los componentes de la serie temporal se combinan de alguna de las tres formas siguientes:

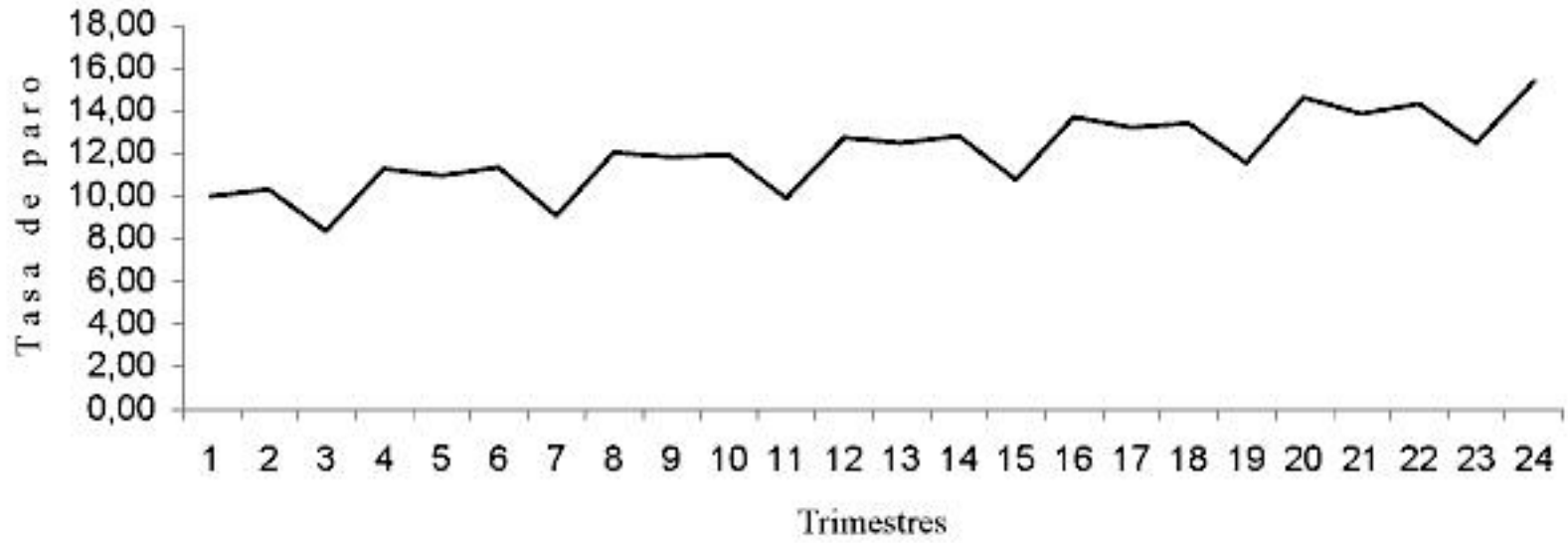
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.



El gráfico permite advertir la presencia de componente estacional. El periodo en el que se completa la variación estacional es de 4 trimestres (1 año). De modo que una media móvil de periodo 4 debe eliminar las variaciones estacionales. Si se denota la tasa de paro trimestral por x_t y la serie temporal por $\{x_t\}_{t=1,..,24}$, las medias móviles de periodo 4 pueden obtenerse como

$$x_t^{(4)} = \frac{x_t^{(4,1)} + x_t^{(4,2)}}{2}$$

siendo $x_t^{(4,1)} = \frac{x_{t-2} + x_{t-1} + x_t + x_{t+1}}{4}$ y $x_t^{(4,2)} = \frac{x_{t-1} + x_t + x_{t+1} + x_{t+2}}{4}$. Los valores de la serie de medias móviles de periodo 4, $\{x_t^{(4)}\}_{t=3,..,22}$, se recogen en la tabla siguiente.

t	x_t	$x_t^{(4,1)}$	$x_t^{(4,2)}$	$x_t^{(4)}$
1	9,99			
2	10,30		9,9750	
3	8,34	9,9750	10,2150	10,09500
4	11,27	10,2150	10,4775	10,34625
5	10,95	10,4775	10,6550	10,56625
6	11,35	10,6550	10,8475	10,75125
7	9,05	10,8475	11,0650	10,95625
8	12,04	11,0650	11,2025	11,13375
9	11,82	11,2025	11,4075	11,30500
10	11,90	11,4075	11,5800	11,49375
11	9,87	11,5800	11,7475	11,66375
12	12,73	11,7475	11,9750	11,86125
13	12,49	11,9750	12,1950	12,08500
14	12,81	12,1950	12,4375	12,31625
15	10,75	12,4375	12,6125	12,52500
16	13,70	12,6125	12,7575	12,68500
17	13,19	12,7575	12,9525	12,85500
18	13,39	12,9525	13,1800	13,06625
19	11,53	13,1800	13,3450	13,26250
20	14,61	13,3450	13,5750	13,46000
21	13,85	13,5750	13,8075	13,69125
22	14,31	13,8075	13,9975	13,90250
23	12,46	13,9975		
24	15,37			

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

En el gráfico anterior, la serie diferenciada estacionalmente es la linea representada con trazo discontinuo. Los valores de la serie diferenciada estacionalmente reflejan los incrementos anuales experimentados en cada instante del tiempo correspondiente a una determinada estación de un año con respecto a la misma estación del año anterior. Por tanto, en la nueva serie las variaciones estacionales presentes en la serie original han desaparecido, pero el incremento tendencial por trimestre también queda eliminado, atenuado o, cuando menos, transformado sustancialmente. Este efecto sobre la tendencia es el defecto fundamental de este procedimiento y, de hecho, no puede considerarse que la serie resultante de la aplicación de diferencias estacionales es equivalente a la serie original sin componente estacional.

En cualquier caso, un modelo para el componente estacional es imprescindible para obtener predicciones acertadas en series que contengan este tipo de variaciones. Por ejemplo, si se desea predecir la entrada de turistas en España en el mes de agosto, una predicción obtenida a través de la extrapolación de una tendencia global infravaloraría sistemáticamente el valor de la serie. Pues bien, uno de los procedimientos que permite evaluar la magnitud de tales fluctuaciones en cada uno de los estadios estacionales que configuran el período en que se completa la variación estacional es el método de la diferencia o razón a las medias móviles.

Sea una serie temporal $\{x_t\}$ tal que $x_t = T_t + S_t + \varepsilon_t$ o $x_t = T_t S_t + \varepsilon_t$ para la que se dispone de s observaciones por año durante n años. Cada elemento x_t puede entonces denotarse como $x_{i,k}$ si el instante del tiempo t corresponde a la estación k del año i, donde $k = 1,...,s$ y $i = 1,...,n$. Los valores de la serie temporal pueden representarse en una tabla de doble entrada como se indica.

$i \backslash k$	1	2	...	s
1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,s}$
2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,s}$
⋮	⋮	⋮	⋱	⋮
n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,s}$

En el método de las diferencias/razones a las medias móviles, la magnitud de las variaciones estacionales se establece por comparación de los datos originales con los correspondientes al componente de largo plazo; y este último se obtiene como tendencia local calculada mediante medias móviles. Los pasos del procedimiento son los siguientes.

(a) Cálculo de la serie de medias móviles centradas de s términos, que se denota por $\left\{x_{i,k}^{(s)}\right\}$. De este modo, se elimina la variación estacional y, se espera que las variaciones irregulares tiendan a compensarse. Debe tenerse en cuenta que el cálculo de esta media móvil supone la pérdida de s observaciones si s es par, o de $s-1$ observaciones si s es impar. En el primer caso, las observaciones de la serie de medias móviles son las representadas en la siguiente tabla.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$I_{2001}^{2004} = \sum_{i=1}^{2} I_{2001}^{2004}(i)\frac{1}{2} = 1.2 \bullet \frac{1}{2} + 1.3 \bullet \frac{1}{2} = 1.25.$$

Entonces puede considerarse que el precio conjunto del pan y del agua ha crecido un 25% entre 2001 y 2004.

Normalmente, las ponderaciones se fijan atendiendo a algún criterio relacionado con la naturaleza de las magnitudes que intervienen en el cálculo. Por ejemplo, en el caso de índices de precios al consumo, se asigna mayor ponderación a los bienes que tienen más peso en la cesta de la compra habitual. Por otro lado, los números índices deben cumplir una serie de propiedades deseables (Martín y Martín, 1989:213-214) que, sin embargo, no se explican aquí.

6.2.2. Índices de precios

Uno de los principales índices elaborados en España es el índice de precios de consumo (*IPC*). Se trata de un índice complejo ponderado construido a partir de los índices de precios de un conjunto de bienes de consumo (cesta de la compra). En general, se emplean como ponderaciones las participaciones de cada bien en el consumo total. Estas participaciones se evalúan teniendo en cuenta los precios del periodo base y las cantidades consumidas en el periodo base, o bien, las cantidades consumidas en el periodo en el que se evalúa el índice.

Sean $x_{i,t}$, $i = 1,\ldots,k$, los precios de los k bienes de consumo que componen la cesta de la compra en el instante del tiempo t. Si se desea evaluar el incremento general de los precios con respecto al instante del tiempo 0, puede utilizarse el número índice complejo ponderado I_0^t, definido como

$$I_0^t = \sum_{i=1}^{k} I_0^t(i)\frac{w_i}{\sum_{i=1}^{k} w_i} = \frac{\sum_{i=1}^{k} I_0^t(i) w_i}{\sum_{i=1}^{k} w_i},$$

y tomar como ponderaciones $w_i = x_{i,0}q_{i,0}$ o $w_i = x_{i,0}q_{i,t}$, donde $q_{i,0}$ y $q_{i,t}$ representan las cantidades consumidas del bien i en los instantes del tiempo 0 y t, respectivamente. Si las ponderaciones se definen como $w_i = x_{i,0}q_{i,0}$, se obtiene el denominado índice de precios de Laspeyres, que viene dado por

$$L_0^t = \frac{\sum_{i=1}^{k} \frac{x_{i,t}}{x_{i,0}} x_{i,0}q_{i,0}}{\sum_{i=1}^{k} x_{i,0}q_{i,0}} = \frac{\sum_{i=1}^{k} x_{i,t}q_{i,0}}{\sum_{i=1}^{k} x_{i,0}q_{i,0}}.$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$\sum_{i=1}^{j} n_i \geq \frac{N}{2}$, es decir, tal que la frecuencia acumulada en las modalidades x_i anteriores o iguales a ella $(i \leq j)$ sea mayor o igual que $\frac{N}{2}$.

Ejemplo 7.1 Suponga que se han registrado los niveles de estudios de los 10 trabajadores de una empresa y han resultado ser los siguientes: "sin estudios", "sin estudios", "sin estudios", "sin estudios", "estudios primarios", "estudios primarios", "estudios primarios", "estudios secundarios", "estudios secundarios", "estudios universitarios". Si se define el atributo X:"nivel de estudios de los 10 trabajadores de la empresa", y se denotan los niveles de estudio por SE (sin estudios), P (estudios primarios), S (estudios secundarios) y U (estudios universitarios), resulta que $X:\{SE,SE,SE,SE,P,P,P,S,S,U\}$.

La distribución de frecuencias absolutas de este atributo es

$$\{(x_i,n_i)\}_{i=1,\ldots,4}:\{(SE,4),(P,3),(S,2),(U,1)\}.$$

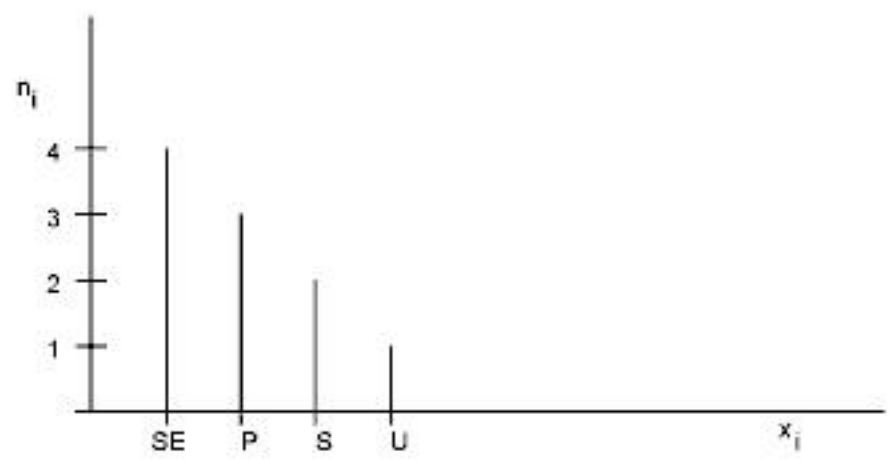
Y la distribución de frecuencias relativas es

$$\{(x_i,f_i)\}_{i=1,\ldots,4}:\left\{\left(SE,\frac{4}{10}\right),\left(P,\frac{3}{10}\right),\left(S,\frac{2}{10}2\right),\left(U,\frac{1}{10}\right)\right\}.$$

Estas distribuciones de frecuencias se expresan en la tabla siguiente.

x_i	n_i	f_i
SE	4	4/10
P	3	3/10
S	2	2/10
U	1	1/10
	10	1

Entonces, el diagrama de barras es el que se indica.



Mientras que el diagrama de sectores es el que se representa a continuación.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Ejemplo 7.2 Suponga que se han registrado los niveles de estudios, cuyas modalidades son SE (sin estudios), P (estudios primarios), S (estudios secundarios) y U (estudios universitarios), y el género, cuyas modalidades son H (hombre) y M (mujer), de los 10 trabajadores de una empresa. Se han encontrado los pares de características siguientes

$$(SE,H),(SE,H),(SE,H),(SE,M),(P,H),(P,H),(P,M),(S,H),(S,M),(U,M).$$

Si se definen X :"nivel de estudios de los 10 trabajadores de la empresa", e Y :"género de los 10 trabajadores de la empresa", entonces (X,Y) es un atributo bidimensional tal que

$$(X,Y):\left\{\left(x_i^{10},y_i^{10}\right)\right\}_{i=1,\ldots,10}:$$
$$:\{(SE,H),(SE,H),(SE,H),(SE,M),(P,H),(P,H),(P,M),(S,H),(S,M),(U,M)\}$$

La distribución bidimensional de frecuencias absolutas puede escribirse como

$$\{(x_i,y_i;n_i)\}_{i=1,\ldots,7}:\{(SE,H;3),(SE,M;1),(P,H;2),(P,M;1),(S,H;1),(S,M;1)(U,M;1)\}$$

o también como

$$\{(x_i,y_j;n_{i,j})\}_{\substack{i=1,2,3,4\\ j=1,2}}:$$
$$:\{(SE,H;3),(SE,M;1),(P,H;2),(P,M;1),(S,H;1),(S,M;1),(U,H;0),(U,M;1)\}$$

Esta última distribución de frecuencias puede expresarse en la tabla siguiente.

$X\backslash Y$	H	M
SE	3	1
P	2	1
S	1	1
U	0	1

Y la distribución de frecuencias relativas del atributo (X,Y) puede escribirse como

$$\{(x_i,y_i;f_i)\}_{i=1,\ldots,7}:$$
$$:\left\{\left(SE,H;\frac{3}{10}\right),\left(SE,M;\frac{1}{10}\right),\left(P,H;\frac{2}{10}\right),\left(P,M;\frac{1}{10}\right),\left(S,H;\frac{1}{10}\right),\left(S,M;\frac{1}{10}\right)\left(U,M;\frac{1}{10}\right)\right\}$$

o también como

$$\{(x_i,y_j;f_{i,j})\}_{\substack{i=1,2,3,4\\ j=1,2}}:$$
$$:\left\{\left(SE,H;\frac{3}{10}\right),\left(SE,M;\frac{1}{10}\right),\left(P,H;\frac{2}{10}\right),\left(P,M;\frac{1}{10}\right),\left(S,H;\frac{1}{10}\right),\left(S,M;\frac{1}{10}\right),(U,H;0),\left(U,M;\frac{1}{10}\right)\right\}$$

Y esta última distribución puede expresarse en la tabla siguiente.

$X\backslash Y$	H	M
SE	0.3	0.1
P	0.2	0.1
S	0.1	0.1
U	0	0.1

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$\gamma = \frac{N_c - N_d}{N_c + N_d},$$

donde, de nuevo, N_c es el número de concordancias y N_d es el número de discordancias. Se tiene que $-1 \leq \gamma \leq 1$. Cuando existe máxima concordancia, $N_d = 0$ y, por tanto, $\gamma = 1$. Mientras que si existe máxima discordancia, $N_c = 0$ y, por tanto, $\gamma = -1$.

Ejemplo 7.4 Sea el atributo bidimensional (X,Y), donde X :"edad de 3 trabajadores", cuyas modalidades son "baja", "mediana" y "alta" e Y :"salario anual de los 3 trabajadores", cuyas modalidades son "bajo", "medio" y "alto". Supóngase que la distribución de frecuencias absolutas es la que se presenta en la tabla de doble entrada siguiente.

$X \setminus Y$	Bajo	Medio	Alto
Baja	0	0	1
Mediana	0	1	0
Alta	1	0	0

Entonces, la relación entre los dos atributos puede evaluarse a través de los coeficientes siguientes. Dado que los dos atributos poseen tres modalidades que poseen un orden explícito, el conjunto de pares de rangos es $\left\{\left(R_i^X, R_i^Y\right)\right\}_{i=1,2,3} : \{(1,3),(2,2),(3,1)\}$, de modo que $\overline{R}^X = \sum_{i=1}^{3} \frac{R_i^X}{3} = 2$ y $\overline{R}^Y = \sum_{i=1}^{3} \frac{R_i^Y}{3} = 2$. Se tiene que

(x_i, y_i)	$\left(R_i^X, R_i^Y\right)$	$\left(R_i^X - \overline{R}^X\right)\left(R_i^Y - \overline{R}^Y\right)$	$\left(R_i^X - \overline{R}^X\right)^2$	$\left(R_i^Y - \overline{R}^Y\right)^2$
$(baja, alto)$	$(1,3)$	-1	1	1
$(mediana, medio)$	$(2,2)$	0	0	0
$(alta, bajo)$	$(3,1)$	-1	1	1
		-2	2	2

Por tanto,

$$r_S = \frac{\sum_{i=1}^{3}\left(R_i^X - \overline{R}^X\right)\left(R_i^Y - \overline{R}^Y\right)}{\sqrt{\sum_{i=1}^{3}\left(R_i^X - \overline{R}^X\right)^2 \sum_{i=1}^{3}\left(R_i^Y - \overline{R}^Y\right)^2}} = -1.$$

Este valor puede deducirse a partir de la representación gráfica de los pares $\left\{\left(R_i^X, R_i^Y\right)\right\}_{i=1,2,3} : \{(1,3),(2,2),(3,1)\}$.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Bajo el supuesto de independencia, las frecuencias registradas deberían ser

$X \setminus Y$	H	M	
SE	21	14	35
P	18	12	30
S	12	8	20
U	9	6	15
	60	40	100

Por tanto, el coeficiente de contingencia χ^2 viene dado por

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{\left(n_{i,j} - \hat{n}_{i,j}\right)^2}{\hat{n}_{i,j}} = 7.73809524 .$$

Se tiene que $0 \le \chi^2 \le 100$. Por tanto, no parece que exista una relación muy fuerte. Esta conclusión se alcanza también si se calculan los coeficientes de contingencia de Cramer, Pearson y Tschuprow, que son, respectivamente,

$$V = \sqrt{\frac{\chi^2}{100}} = 0.27817432 , \; C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = 0.28960485 , \; T = \sqrt{\frac{\chi^2}{100\sqrt{3}}} = 0.21136678 .$$

EJERCICIOS

7.1. Sea el atributo X :"nivel de ingresos de las 500 familias de un pequeño municipio", y sean los niveles de ingresos B (bajos), M (medios) y A (altos), tales que la distribución de frecuencias de este atributo es la que se expresa en la tabla siguiente.

x_i	n_i	f_i
B	100	1/5
M	300	3/5
A	100	1/5
	500	1

(a) Represente esta información mediante un diagrama de barras y un diagrama de sectores. Construya también un diagrama de Pareto.

(b) ¿Qué medidas de posición utilizaría usted para resumir la información contenida en la tabla anterior? Calcule e interprete el valor obtenido para cada una de esas medidas.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

opción de izquierda (i), centro (c) o derecha (d). Entonces, el espacio muestral asociado al experimento aleatorio puede escribirse indicando lo que sucede con cada elector, de modo que cada resultado podrá identificarse mediante un par de modalidades de los atributos que indican la opción política de cada elector. Si la primera modalidad corresponde al elector 1 y la segunda al elector 2, el conjunto de resultados posibles es

$$\Omega: \{(i,i),(i,c),(i,d),(c,i),(c,c),(c,d),(d,i),(d,c),(d,d)\}.$$

En cualquier caso, el espacio muestral asociado a un experimento puede especificarse de varias maneras, en función de lo que se quiera observar del experimento. Por ejemplo, puede que se desee obtener una predicción del voto que identifique el partido político concreto al que vota el elector.

Puesto que el objetivo es evaluar las probabilidades de ocurrencia de enunciados sobre los resultados del experimento, hay formas de expresar el espacio muestral más adecuadas que otras en el sentido de que facilitan el cálculo, aunque el resultado de la probabilidad será el mismo sea cual sea la forma en que se exprese el espacio muestral. Estos enunciados acerca del resultado del experimento aleatorio constituyen los llamados sucesos.

Un suceso es cualquier afirmación sobre los resultados del experimento, que puede ocurrir o no. Formalmente, un *suceso* es un subconjunto A del espacio muestral Ω asociado a un experimento aleatorio formado por los resultados del experimento en los que se verifica la afirmación que define al suceso. Así, un suceso A ocurre si alguno de los elementos del subconjunto A que lo define es el resultado del experimento. Los sucesos pueden ser *elementales*, constituidos por un solo elemento del espacio muestral y, por tanto, imposibles de descomponer en otros más sencillos; o *compuestos*, resultantes de la unión de varios sucesos elementales.

Los sucesos son conjuntos y, por tanto, pueden realizarse con ellos las operaciones propias de conjuntos. Algunas definiciones en este sentido se aportan a continuación.

Dos sucesos A y B son iguales, y se denotará por $A = B$, si ambos están formados por los mismos resultados del experimento. Estos dos sucesos constituyen afirmaciones equivalentes sobre el resultado de un experimento aleatorio.

Se dice que un suceso A está contenido en otro B, y se denotará por $A \subset B$, si A es un subconjunto de B; es decir, si cada uno de los resultados del experimento que pertenecen a A, pertenece también a B. De este modo, si ocurre A, puede asegurarse que ocurre B.

La unión de los sucesos A y B, que se denotará por $A \cup B$, es otro suceso que contiene todos los resultados del experimento que pertenecen a A y también todos aquéllos que pertenecen a B. Por tanto, si ocurre el suceso $A \cup B$, entonces ocurrirá A o bien ocurrirá B.

La intersección de los sucesos A y B, que se denotará por $A \cap B$, es otro suceso que contiene los resultados del experimento que, además de pertenecer a A, pertenecen también a B. Por tanto, si ocurre el suceso $A \cap B$, entonces ocurrirá A y también ocurrirá B.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

la probabilidad de ocurrencia de los sucesos que se transforman en ellos. Sea un espacio probabilístico $(\Omega, \mathcal{A}, \mathcal{P})$ y el espacio probabilizable $(\mathbb{R}^2, \mathcal{B}_2)$. Una aplicación

$$\begin{array}{rccc} (X,Y): & \Omega & \to & \mathbb{R}^2 \\ & \omega \in \Omega & \to & (X(\omega), Y(\omega)) = (x,y) \in \mathbb{R}^2 \end{array}$$

es una variable aleatoria si y sólo si $(X,Y)^{-1}(B) \in \mathcal{A}$, $\forall B \in \mathcal{B}_2$, siendo $(X,Y)^{-1}(B): \{\omega \in \Omega / (X(\omega), Y(\omega)) = (x,y) \in B\}$. Es decir, la aplicación es variable aleatoria si y sólo si cualquiera de los conjuntos de puntos del plano formados con los puntos del plano asignados a los resultados del experimento tiene como imagen inversa uno de los sucesos del conjunto de sucesos $\mathcal{A}$. Si el conjunto $\mathcal{A}$ es el conjunto de todos los sucesos que pueden definirse, cualquier aplicación será variable aleatoria.

Se define el rango de una variable aleatoria bidimensional (X,Y), que se denotará por $R_{X,Y}$, como el conjunto de puntos del plano que son imagen, mediante la aplicación (X,Y), de algún elemento de Ω. Es decir,

$$R_{X,Y}: \{(x,y) \in \mathbb{R}^2 / \exists \omega \in \Omega : (X(\omega), Y(\omega)) = (x,y)\}.$$

Ejemplo 9.2 Suponga que interesa predecir la orientación del voto de un elector A y de otro elector B. El espacio muestral puede escribirse indicando, en primer lugar, la orientación del voto del elector A y, en segundo lugar, la del elector B. Utilizando las iniciales i (*izquierda*), c (*centro*) y d (*derecha*), el espacio muestral puede expresarse como $\Omega: \{(i,i),(i,c),(i,d),(c,i),(c,c),(c,d),(d,i),(d,c),(d,d)\}$. Considerando el conjunto $\mathcal{A}$ de todos los sucesos que pueden definirse, la aplicación

$$\begin{array}{rccc} (X,Y): & \Omega & \to & \mathbb{R}^2 \\ & (i,i) & \to & (1,1) \\ & (i,c) & \to & (1,1) \\ & (i,d) & \to & (1,2) \\ & (c,i) & \to & (1,1) \\ & (c,c) & \to & (1,1) \\ & (c,d) & \to & (1,2) \\ & (d,i) & \to & (2,1) \\ & (d,c) & \to & (2,1) \\ & (d,d) & \to & (2,2) \end{array}$$

es una variable aleatoria bidimensional cuyo rango es $R_{X,Y}: \{(1,1),(1,2),(2,1),(2,2)\}$.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

9.4.1. Variable aleatoria unidimensional discreta

Se puede definir una variable aleatoria unidimensional discreta como aquélla cuyo rango es un conjunto finito o infinito numerable de puntos de $\mathcal{R}$.

Para este tipo de variables aleatorias es útil definir la función de masa. Sea el espacio probabilístico $(\Omega, \mathcal{A}, \mathcal{P})$ y la variable aleatoria X que convierte el espacio probabilístico anterior en el nuevo espacio probabilístico $(\mathcal{R}, \mathcal{B}, P_X)$. Se define la función de masa o cuantía de la variable aleatoria X, que se denotará por p_X, como la función que indica la probabilidad asignada a cada uno de los puntos de la recta real, es decir,

$$p_X(x) = P_X(\{x\}), \ \forall x \in \mathcal{R}.$$

En el caso de variables aleatorias discretas, los valores de la función de masa indican las probabilidades individuales de los puntos del rango y a la función de masa se le denomina función de probabilidad. Es decir, la función de probabilidad de la variable aleatoria discreta X se define como

$$P_X(x) = P(X = x) = p_X(x), \ \forall x \in \mathcal{R}.$$

Esta función verifica las dos condiciones siguientes.

(1) $P_X(x) \geq 0, \ \forall x \in \mathcal{R}$.

(2) $\sum_{x \in R_X} P_X(x) = 1$.

Si X es una variable aleatoria discreta, su función de distribución será discontinua, ya que

$$F_X(x) = \sum_{x_i \in R_X / x_i \leq x} P(X = x_i), \ \forall x \in \mathcal{R}.$$

La función de distribución va acumulando las probabilidades individuales de los puntos a la izquierda de aquél en el que se valora la función. De ahí que en una variable aleatoria discreta, la función de distribución sea una función en escalera. Va saltando en cada uno de los puntos del rango. La magnitud del salto es igual a la probabilidad del punto en el que éste se produce.

Ejemplo 9.7 Suponga de nuevo la variable aleatoria X del ejemplo 9.1 cuya función de distribución viene dada por

$$F_X(x) = \begin{cases} 0 & , \quad x < 1 \\ \frac{2}{3} & , \quad 1 \leq x < 2 \\ 1 & , \quad x \geq 2 \end{cases}$$

Entonces, se trata de una variable aleatoria discreta cuya función de probabilidad es

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

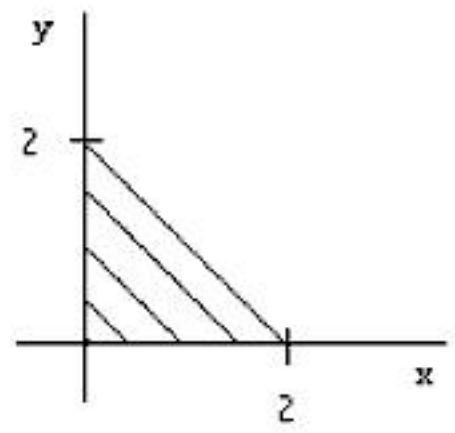
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.



(a) Obtenga la probabilidad de que el suelo comercial supere las 1000 hectáreas.
(b) Calcule la probabilidad de que el suelo agrícola supere las mil hectáreas.
(c) Si el suelo turístico es inferior a 1000 hectáreas, ¿cuál es la probabilidad de que el suelo agrícola supere las mil hectáreas?
(d) Obtenga la probabilidad de que la cantidad utilizada de suelo agrícola sea mayor que una constante a $(0 < a < 2)$.
(e) Obtenga la distribución conjunta de la variable aleatoria bidimensional (U,V), cuyos elementos recogen el desequilibrio entre suelo turístico y agrícola, así como entre suelo comercial y agrícola, es decir, $U = X - A$, $V = Y - A$. Calcule entonces la probabilidad de que se destine menos terreno a uso agrícola que a cada uno de los otros dos usos.
Nótese que, si $(u,v) = g(x,y)$ es una transformación biunívoca tal que

$$g(x,y) = \left(u = g_1(x,y), v = g_2(x,y)\right) \text{ y } g^{-1}(u,v) = h(u,v) = \left(x = h_1(u,v), y = h_2(u,v)\right),$$

entonces

$$\int_a^b \int_c^d Q(x,y)dydx = \iint_R Q(h_1(u,v),h_2(u,v))\left|J(u,v)\right|dudv,$$

donde

$$R = \left\{(u,v)/a \le h_1(u,v) \le b, c \le h_2(u,v) \le d\right\} \text{ y } J(u,v) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \neq 0, \ \forall (u,v) \in R.$$

9.6. Suponga que la distribución conjunta de los ingresos mensuales, Y, y los gastos mensuales, X, en miles de euros, de una familia se puede representar como

$$f_{X,Y}(x,y) = \begin{cases} 2, & 0 < x < y < a \\ 0, & resto \end{cases}$$

(a) Calcule el ahorro máximo mensual que puede generar la familia.
(b) Evalúe la probabilidad de que los gastos mensuales sean superiores a 1/3 si los ingresos mensuales son inferiores a 2/3.
(c) Si el ingreso mensual obtenido es y, obtenga la distribución de los gastos mensuales. ¿Cuál es la distribución de los ingresos mensuales, si el gasto mensual realizado es x?
(d) Calcule la probabilidad de que los ingresos mensuales sean superiores a los gastos mensuales realizados sabiendo que éstos equivalen a 500 euros.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

donde a y b son constantes. Nótese que, si X es discreta,

$$E[aH(X)+bG(X)]=\sum_{x\in R_X}(aH(x)+bG(x))P(X=x)=$$
$$=\sum_{x\in R_X}aH(x)P(X=x)+\sum_{x\in R_X}bG(x)P(X=x))=aE[H(X)]+bE[G(X)]$$

Mientras que, si X es continua,

$$E[aH(X)+bG(X)]=\int_{R_X}(aH(x)+bG(x))f_X(x)dx=$$
$$=\int_{R_X}aH(x)f_X(x)dx+\int_{R_X}bG(x)f_X(x)dx=aE[H(X)]+bE[G(X)]$$

En general, sea X una variable aleatoria, sean las constantes k_i, $i=1,...,n$, y sean las funciones $H_i:\Re\to\Re$, $i=1,...,n$. Entonces

$$E\left[\sum_{i=1}^{n}k_iH_i(X)\right]=\sum_{i=1}^{n}k_iE[H_i(X)].$$

Ejemplo 10.1 Suponga que en un proceso productivo se emplea un input cuyo precio en el mercado, X, puede ser 1, 2 o 3 euros con la misma probabilidad, es decir,

$$P(X=x)=\begin{cases}\frac{1}{3} & , \quad x=1,2,3\\ 0 & , \quad resto\end{cases}$$

Entonces, el precio esperado del input es

$$E[X]=\sum_{x\in R_X}xP(X=x)=\sum_{x=1}^{3}x\frac{1}{3}=1\bullet\frac{1}{3}+2\bullet\frac{1}{3}+3\bullet\frac{1}{3}=2.$$

Suponga que se decide que el precio del output, X, sea el doble del precio del input más 1. Entonces, el precio esperado del output es

$$E[Y]=E[2X+1]=\sum_{x\in R_X}(2x+1)P(X=x)=\sum_{x=1}^{3}(2x+1)\frac{1}{3}=3\bullet\frac{1}{3}+5\bullet\frac{1}{3}+7\bullet\frac{1}{3}=5.$$

Suponga ahora que la variable aleatoria X del ejemplo anterior se mueve con la misma verosimilitud en el intervalo (1,3), es decir,

$$f_X(x)=\begin{cases}\frac{1}{2} & , \quad 1<x<3\\ 0 & , \quad resto\end{cases}$$

Entonces, el precio esperado del input es

$$E[X]=\int_{x\in R_X}xf_X(x)dx=\int_1^3x\frac{1}{2}dx=\left(\frac{x^2}{4}\right)\Big]_1^3=\frac{9}{4}-\frac{1}{4}=2.$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

entonces $X(\omega_1)=x$, $X(\omega_2)=x$ y la variable aleatoria X asignará el valor x a cualquier resultado del experimento tal que se produzcan x éxitos y $n-x$ fracasos. De modo que

$$P(\{\omega_1\})=\overbrace{pp...p}^{x}\overbrace{(1-p)(1-p)...(1-p)}^{n-x}=p^x(1-p)^{n-x},$$

$$P(\{\omega_2\})=(1-p)\overbrace{pp...p}^{x}\overbrace{(1-p)(1-p)...(1-p)}^{n-x-1}=p^x(1-p)^{n-x}$$

y, en general, ésta será la probabilidad asignada a cualquiera de los resultados del experimento tales que $X(\omega)=x$. Por tanto,

$$P(X=x)=P(\{\omega\in\Omega/X(\omega)=x\})=P(\{\omega_1,\omega_2,...\})=\sum_{\omega/X(\omega)=x}P(\{\omega\})=Np^x(1-p)^{n-x},$$

siendo N el número de elementos del espacio muestral tales que ocurren x éxitos y $n-x$ fracasos. Entonces,

$$N=PR_n^{x,n-x}=\frac{n!}{x!(n-x)!}=\binom{n}{x};$$

por tanto,

$$P(X=x)=\begin{cases}\binom{n}{x}p^x(1-p)^{n-x}, x=0,1,2,...,n\\ 0 \qquad\qquad ,resto\end{cases}$$

Utilizando la función de probabilidad obtenida, resulta que

$$\mu_X=E[X]=\sum_{x=0}^{n}x\binom{n}{x}p^x(1-p)^{n-x}=\sum_{x=1}^{n}x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}=$$

$$=\sum_{x=1}^{n}\frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x}=np\sum_{x=1}^{n}\frac{(n-1)!}{(x-1)!(n-x)!}p^{x-1}(1-p)^{n-x}$$

y haciendo el cambio $y=x-1$, se tiene que

$$\mu_X=np\sum_{y=0}^{n-1}\frac{(n-1)!}{y!(n-1-y)!}p^y(1-p)^{n-1-y},$$

de modo que la expresión anterior incluye la suma de los valores de la función de probabilidad de una variable aleatoria Y que es $B(n-1,p)$ y, por lo tanto,

$$\mu_X=np.$$

Se espera que ocurran np éxitos en las n pruebas. El parámetro p es la proporción esperada de éxitos. Por otra parte, teniendo en cuenta que

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

12

Distribuciones continuas y el teorema central del límite

En este capitulo se introducen distribuciones continuas que resultan adecuadas para modelar determinados comportamientos probabilísticos. Supóngase, por ejemplo, que se desea analizar el consumo de un individuo que pertenece a un cierto estrato de renta. Puede asumirse que dicho consumo varía en un intervalo de tal modo que los consumos inferiores no sean ni más ni menos probables que los consumos altos, siempre dentro de ese intervalo. En este caso, la evaluación de la probabilidad de que el consumo del individuo en cuestión se mueva entre dos cantidades puede llevarse a cabo sin complicaciones a través de la distribución uniforme, que se presenta en el primer epígrafe del capítulo. Si, en cambio, se está examinando la intensidad del tráfico que circula por determinado punto de una red viaria y, en concreto, se estudia el tiempo que transcurre desde que comienza a observarse el tráfico hasta que pasa el primer vehículo, puede ser adecuado acudir a la distribución exponencial, de la que se ocupa el epígrafe segundo.

En el tercer epígrafe se introduce la distribución normal, que en su versión univariante permite recoger el comportamiento probabilístico de magnitudes cuyos valores más verosímiles son los valores centrales, mientras que a medida que los valores son más extremos desciende su verosimilitud. Así ocurre, por ejemplo, cuando la magnitud en cuestión es el peso o la estatura de un individuo. Por otra parte, si los técnicos de una empresa exportadora de tomates desean estudiar el calibre —longitud del diámetro expresado en mm— y el peso —expresado en gramos— de un fruto cualquiera y, en concreto, desean calcular, por ejemplo, la probabilidad de que estas magnitudes superen ciertos umbrales que garantizan mayor cotización, puede resultar adecuado utilizar una distribución normal bivariante. La distribución normal multivariante se expondrá sólo de forma introductoria, de modo que pueda advertirse que se trata de una generalización de la distribución normal univariante. Estas distribuciones normales, univariantes o multivariantes, constituyen modelos probabilísticos de extrema importancia, no sólo por los numerosos fenómenos aleatorios que parecen caer en su

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

puede ayudar a calibrar la probabilidad de que $\overline{X}$ (proporción de votantes de la muestra) difiera de p (proporción poblacional de votantes) más allá de cierta cantidad.

En definitiva, la inferencia estadística proporciona mecanismos de obtención de información sobre una población amparados en criterios que, de forma más o menos directa, tratan de manejar el carácter incierto de las conclusiones inductivas derivadas a partir de una muestra. El proceso arranca, generalmente, de la formulación de modelos teóricos sobre la distribución poblacional, de la que se pretende conocer algún parámetro (estimación) o evaluar si puede descartarse la veracidad de alguna afirmación sobre el valor del mismo (contraste de hipótesis).

Ahora bien, este objetivo puede alcanzarse desde cualquiera de las diferentes perspectivas que han sido propuestas. En el planteamiento original de la inferencia clásica formulada por Fisher, Neyman y Pearson, la muestra es la única fuente de información sobre los parámetros poblacionales; mientras que, desde el punto de vista bayesiano, también se considera la información a priori sobre ellos. Y, finalmente, en el enfoque de la teoría de la decisión, no sólo se tiene en cuenta la información muestral y la información a priori sobre parámetros como la proporción de votantes de un partido, sino que la toma de una determinada decisión dependerá también de las consecuencias que se deriven de dicha decisión. Por otro lado, es preciso considerar que, en determinadas circunstancias, es más útil recurrir a un planteamiento inferencial que no parta de la imposición de supuestos distribucionales sobre la magnitud estudiada. En este caso, se ingresa en el terreno de la inferencia no paramétrica.

Por último, es importante que se acepte la necesidad de recurrir a la inferencia para abordar situaciones que, como la anterior, pertenecen al ámbito de la investigación social. En este sentido, es oportuno recordar lo ya comentado en el bloque anterior en relación al hecho de que la incertidumbre impregna los comportamientos de los agentes sociales y, por tanto, el recurso a métodos inferenciales que superen el rígido marco del enfoque descriptivo, se muestra como una condición necesaria para que las conclusiones obtenidas sean el resultado de un aprovechamiento óptimo de la información disponible.

13.2 LA POBLACIÓN, LA MUESTRA Y LOS ESTADÍSTICOS MUESTRALES

Los datos muestrales son la información con la que se obtienen valores de estadísticos muestrales que se utilizan para construir las inferencias inductivas sobre una población. Las muestras pueden proceder de un experimento real (diseñado por el investigador, que controla las características de los individuos de la muestra antes de formarla) o de un estudio observacional (la muestra se forma con individuos no controlados previamente). Así, en el ejemplo antes citado, los miembros del partido político tendrán que arbitrar criterios para seleccionar los individuos de la población de la región considerada para obtener información sobre la proporción poblacional de votantes del partido.

Dado que, generalmente, no se conocen los factores que pueden incidir en una determinada magnitud bajo estudio, el mecanismo de obtención de la muestra adquiere una relevancia notoria, ya que los resultados inferidos pueden variar de una muestra a otra. En otras palabras, la variabilidad de la muestra implica también la de los estadísticos a

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

EJERCICIOS

13.1. Suponga una población formada por 5 alumnos que han cursado estudios durante 10, 12, 14, 16 y 18 años, respectivamente. Un individuo que desconoce esta información desea saber cuál es el número medio de años de estudio de estos alumnos.

(a) Suponga que se le permite preguntar a tres de ellos, elegidos aleatoriamente. El individuo decide entonces aproximar la media poblacional desconocida por la media obtenida para los tres alumnos a los que pregunta. ¿Cuál es la probabilidad de que la media obtenida para los tres alumnos seleccionados coincida con la media poblacional? ¿Cuál es el error máximo que se puede cometer mediante esta aproximación?

(b) ¿Qué ocurrirá con la precisión de la aproximación si sólo se permite preguntar a dos alumnos? ¿Cuál es, en ese caso, la probabilidad de que la aproximación sea acertada? ¿Y el error máximo que se puede cometer?

13.2. Suponga que se desea conocer el grado de satisfacción con sus estudios de los alumnos de una facultad. Para ello, se ha pensado en utilizar una escala de 0 a 10 y preguntar a una muestra de 20 de esos alumnos. Suponga que las respuestas de estos alumnos son las siguientes.

$$\{8,7,6,6,4,3,6,5,4,8,2,3,1,7,2,3,8,9,6,2\}$$

Obtenga los valores de la media muestral, el momento muestral de segundo orden, la cuasivarianza muestral, el mínimo muestral, el máximo muestral y la mediana muestral.

13.3. Suponga que se desea obtener una muestra aleatoria simple de tamaño 5 de una población de 1000 individuos. Explique cómo podría seleccionar los individuos de la muestra a partir de la siguiente sección de una tabla de números aleatorios.

876	614	360	354	823	807	023	892	690	724	911	600	689	171	662
304	797	156	237	177	090	573	045	399	183	618	826	987	469	957
341	156	595	836	623	372	091	519	424	878	339	452	775	454	302
438	743	161	293	394	200	405	365	191	700	043	994	560	936	994
021	155	641	215	894	392	139	275	558	903	948	988	038	371	651
479	227	343	778	356	224	753	761	450	815	296	284	884	985	908
683	126	594	445	738	791	949	470	672	501	191	944	234	197	897
676	831	400	928	309	009	457	726	292	387	538	381	688	798	944
031	062	603	454	652	446	434	257	626	351	969	325	002	162	432
077	737	802	692	369	051	709	639	382	772	236	239	736	095	085

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$\overline{X} \text{ es } N\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

Y aunque la distribución poblacional no sea normal, si el tamaño de la muestra es suficientemente grande, la distribución de la media muestral puede aproximarse a una normal en virtud del teorema central del límite. Nótese que

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{L} N\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

En cuanto a la cuasivarianza muestral, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, puede demostrarse que $E\left[S^2\right] = \sigma_X^2$. Nótese que $S^2 = \frac{n}{n-1}\left[M_2 - M_1^2\right]$, de modo que

$$E\left[S^2\right] = \frac{n}{n-1}\left(E\left[M_2\right] - E\left[M_1^2\right]\right).$$

Además, se tiene que

$$E\left[M_k^2\right] = E\left[\left(\frac{\sum_{i=1}^{n} X_i^k}{n}\right)^2\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left[X_i^{2k}\right] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} E\left[X_i^k\right]E\left[X_j^k\right]\right] =$$

$$= \frac{1}{n^2}\left(nm_{2k} + n(n-1)m_k^2\right)$$

y, por tanto, $E\left[M_1^2\right] = \frac{1}{n}\left(m_2 + (n-1)m_1^2\right)$. Entonces,

$$E\left[S^2\right] = \frac{n}{n-1}\left(m_2 - \frac{m_2 + (n-1)m_1^2}{n}\right) = \frac{n}{n-1}\left(\frac{(n-1)\left(m_2 - m_1^2\right)}{n}\right) = m_2 - m_1^2 = \sigma_X^2.$$

Ejemplo 14.1 Suponga que antes de celebrarse unas elecciones para decidir el gobierno de un país, los miembros del partido político gobernante deciden realizar una encuesta con objeto de predecir la proporción p de electores que votarán por ese partido. Suponiendo que se pregunta a 10000 encuestados si tienen intención o no de votar por el partido y que las respuestas de los encuestados son independientes, podría evaluarse la probabilidad de que la proporción de encuestados que declara que va a votar por el partido en cuestión difiera de la proporción poblacional p en menos de 0.01.

Sea X una variable aleatoria que indica la intención de voto de un elector cualquiera, es decir, X toma el valor 1 si el elector tiene intención de votar por el partido gobernante o el

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

y se obtiene una integral tipo *gamma*, de modo que

$$\int_{-\infty}^{+\infty} f_{\chi_d^2}(x)dx = \frac{1}{\Gamma\left(\frac{d}{2}\right)}\int_0^{+\infty} u^{\frac{d}{2}-1}e^{-u}du = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} = 1\,.$$

La función generatriz de momentos de esta distribución viene dada por

$$m_{\chi_d^2}(t) = \frac{1}{(1-2t)^{\frac{d}{2}}}\text{, si } t < \frac{1}{2}\,.$$

Nótese que

$$m_{\chi_d^2}(t) = E\left[e^{t\chi_d^2}\right] = \int_0^{+\infty} e^{tx}\frac{x^{\frac{d}{2}-1}}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}e^{-\frac{x}{2}}dx = \int_0^{+\infty}\frac{x^{\frac{d}{2}-1}}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}e^{-\left(\frac{1}{2}-t\right)x}dx\,.$$

Por tanto, si se efectúa el cambio de variable $\left(\frac{1}{2}-t\right)x = u$ y se asume que $t < \frac{1}{2}$, se tiene que

$$\int_0^{+\infty}\frac{x^{\frac{d}{2}-1}}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}e^{-\left(\frac{1}{2}-t\right)x}dx = \int_0^{+\infty}\frac{u^{\frac{d}{2}-1}}{\left(\frac{1-2t}{2}\right)^{\frac{d}{2}-1}2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}e^{-u}\frac{1}{\left(\frac{1-2t}{2}\right)}du$$

y se obtiene una integral tipo *gamma*, de modo que

$$m_{\chi_d^2}(t) = \int_0^{+\infty}\frac{u^{\frac{d}{2}-1}}{(1-2t)^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)}e^{-u}du = \frac{\Gamma\left(\frac{d}{2}\right)}{(1-2t)^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)} = \frac{1}{(1-2t)^{\frac{d}{2}}}\,.$$

Y a partir de la función generatriz de momentos puede obtenerse la media y la varianza de la distribución. Se tiene que

$$\frac{\partial m_{\chi_d^2}(t)}{\partial t} = -\frac{d}{2}(1-2t)^{-\frac{d}{2}-1}(-2) = d(1-2t)^{-\frac{d}{2}-1}\,,$$

mientras que

$$\frac{\partial^2 m_{\chi_d^2}(t)}{\partial^2 t} = d\left(-\frac{d}{2}-1\right)(1-2t)^{-\frac{d}{2}-2}(-2) = 2d\left(\frac{d}{2}+1\right)(1-2t)^{-\frac{d}{2}-2}\,.$$

Por tanto, $\mu_{\chi_d^2} = d$ y $\sigma^2_{\chi_d^2} = 2d\left(\frac{d}{2}+1\right) - d^2 = 2d$.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

y, dado que $\overline{X}$ y S^2 son independientes, se tiene que $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ y $\dfrac{(n-1)S^2}{\sigma^2}$ son independientes. Por tanto,

$$\frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/n-1}} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \text{ es } T_{n-1}.$$

Sean ahora dos variables aleatorias X e Y tales que X es $N(\mu_X,\sigma_X^2)$ e Y es $N(\mu_Y,\sigma_Y^2)$. Y sean $\{X_1,\ldots,X_{n_X}\}$ e $\{Y_1,\ldots,Y_{n_Y}\}$ muestras aleatorias independientes de tamaños n_X y n_Y de las variables aleatorias X e Y. Entonces, si σ_X^2 y σ_Y^2 son conocidas, se tiene que

$$\overline{X} \text{ es } N\left(\mu_X,\frac{\sigma_X^2}{n_X}\right), \qquad \overline{Y} \text{ es } N\left(\mu_Y,\frac{\sigma_Y^2}{n_Y}\right),$$

y dada la independencia de las dos muestras aleatorias, resulta que

$$\overline{X}-\overline{Y} \text{ es } N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}\right).$$

Mientras que si las varianzas son desconocidas pero iguales, es decir, $\sigma_X^2=\sigma_Y^2=\sigma^2$, se tiene que

$$\overline{X} \text{ es } N\left(\mu_X,\frac{\sigma^2}{n_X}\right), \qquad \overline{Y} \text{ es } N\left(\mu_Y,\frac{\sigma^2}{n_Y}\right),$$

y dada la independencia de las dos muestras aleatorias, resulta que

$$\overline{X}-\overline{Y} \text{ es } N\left(\mu_X-\mu_Y,\sigma^2\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)\right).$$

Por tanto,

$$\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sigma\sqrt{\dfrac{1}{n_X}+\dfrac{1}{n_Y}}} \text{ es } N(0,1).$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

15.1 EL PROBLEMA DE LA ESTIMACIÓN

Sea $\{X_1,...,X_n\}$ una muestra aleatoria de tamaño n de la variable aleatoria poblacional X cuya distribución poblacional depende de un parámetro desconocido γ, es decir, $F_X(t)=h(t,\gamma)$. De acuerdo con la naturaleza del parámetro, podrá asumirse que el valor desconocido del parámetro pertenecerá a algún conjunto determinado, denominado *conjunto paramétrico*, que se denotará por Γ. Es decir, $\gamma \in \Gamma$. Ahora bien, dentro del conjunto de valores del parámetro pertenecientes al conjunto paramétrico, pueden elaborarse criterios que orienten la selección de uno de tales valores como aproximación del valor desconocido en función de la información que la muestra contenga sobre el parámetro que se desea conocer. Si se decide resumir esa información en el estadístico $\hat{\Gamma}=g(X_1,...,X_n)$, de modo que el valor de dicho estadístico una vez obtenida una muestra $\{x_1,...,x_n\}$ se emplea como aproximación del valor del parámetro γ, entonces se dice que $\hat{\Gamma}=g(X_1,...,X_n)$ es un *estimador* del parámetro γ, mientras que $\hat{\gamma}=g(x_1,...,x_n)$ es la *estimación* puntual del parámetro γ. También es posible que el objetivo no sea obtener un valor puntual que sirva como aproximación del valor desconocido del parámetro, sino más bien, encontrar dos valores extremos de modo que pueda asumirse que el valor del parámetro se sitúa entre esos dos extremos con cierta confianza. Esos dos extremos definen un intervalo de confianza para el parámetro en cuestión, pero el problema de la estimación por intervalos será abordado en el capítulo siguiente. Los dos epígrafes siguientes se refieren, por tanto, a la estimación puntual. En primer lugar, se analizan las propiedades generales que pueden orientar la elección de un estimador. En segundo lugar, se exponen diferentes métodos para la obtención de estimadores.

15.2 PROPIEDADES DE LOS ESTIMADORES

Como se ha visto en el apartado anterior, la información muestral sobre el parámetro desconocido puede resumirse a través de diferentes estadísticos y cada uno de ellos ofrecerá una aproximación determinada del parámetro poblacional. Sin embargo, aunque no es posible *a priori* evaluar el error que se comete con cada una de estas aproximaciones, la distribución muestral de los estadísticos empleados para efectuarlas puede aconsejar la elección de alguna aproximación en detrimento de otras. Esta selección puede llevarse a cabo teniendo en cuenta las propiedades deseables que verifica cada uno de los estadísticos propuestos como estimadores del parámetro desconocido. Aunque sólo sea intuitivamente, parece conveniente que el estimador contenga la máxima información que la muestra proporciona sobre el parámetro estimado, que la esperanza del estimador sea próxima o igual al valor verdadero del parámetro que estima, que su varianza sea la menor posible o que el estimador converja en probabilidad al parámetro estimado.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

(c) Intervalo de confianza para el parámetro σ^2 (μ conocida)

Si la media poblacional es conocida, el intervalo de confianza para la varianza poblacional puede obtenerse teniendo en cuenta que

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \text{ es } \chi^2_n,$$

de modo que

$$P\left(\chi^2_{n,\frac{\alpha}{2}} \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \le \chi^2_{n,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

y, por tanto,

$$P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}\right) = 1-\alpha,$$

de donde resulta que

$$I_{\sigma^2}^{1-\alpha} = \left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,1-\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,\frac{\alpha}{2}}}\right].$$

(d) Intervalo de confianza para el parámetro σ^2 (μ desconocida)

Si la media poblacional es desconocida, el intervalo de confianza para la varianza poblacional puede obtenerse teniendo en cuenta que

$$\frac{(n-1)S^2}{\sigma^2} \text{ es } \chi^2_{n-1},$$

de modo que

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

y, por tanto,

$$P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha,$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Nótese que

$$P\left(\left|\overline{X}_n-\mu\right|<e\right)=P\left(\left|\frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}\right|<\frac{e}{\frac{\sigma}{\sqrt{n}}}\right)\cong P\left(|Z|<\frac{\sqrt{n}e}{\sigma}\right)=2N_Z\left(\frac{\sqrt{n}e}{\sigma}\right)-1.$$

Por tanto, para que $P\left(\left|\overline{X}_n-\mu\right|<e\right)\geq 1-\alpha$, debe verificarse que

$$2N_Z\left(\frac{\sqrt{n}e}{\sigma}\right)-1\geq 1-\alpha,$$

es decir, $N_Z\left(\frac{\sqrt{n}e}{\sigma}\right)\geq 1-\frac{\alpha}{2}$ y, por tanto, $\frac{\sqrt{n}e}{\sigma}\geq z_{1-\frac{\alpha}{2}}$, siendo $z_{1-\frac{\alpha}{2}}$ el cuantil correspondiente de la distribución normal estándar. La desigualdad anterior implica que

$$n\geq z_{1-\frac{\alpha}{2}}^{\ 2}\frac{\sigma^2}{e^2}.$$

Si la varianza poblacional es desconocida, la única forma de asegurar la desigualdad es tomar un tamaño muestral mayor que el máximo de la función $z_{1-\frac{\alpha}{2}}^{\ 2}\frac{\sigma^2}{e^2}$, pero esta práctica puede conducir a tamaños muestrales excesivamente grandes. Si éste es el caso, otra opción consiste en aproximar el tamaño muestral necesario sustituyendo la varianza poblacional por una estimación de ésta. Generalmente, se emplea el valor de la cuasivarianza muestral, S^2, obtenida para una muestra piloto. De hecho, si la muestra procede de una variable aleatoria poblacional con distribución normal, $\frac{\overline{X}_n-\mu}{\frac{S}{\sqrt{n}}}$ sigue una distribución T_{n-1} y, si el tamaño muestral es suficientemente grande, esta distribución es asintóticamente normal estándar.

En el caso particular en que X es $B(p)$, la media poblacional $\mu_X=p$, puede interpretarse como la proporción desconocida de individuos de la población que presentan cierta característica, mientras que la media muestral, $\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$, recoge la proporción muestral de individuos con dicha característica. Entonces,

$$\overline{X}_n\xrightarrow{\ L\ }N\left(p,\frac{p(1-p)}{n}\right),$$

de modo que, si el tamaño muestral n es suficientemente grande, podrá determinarse el valor de n tal que

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

de modo que

$$P\left(\left|\frac{\overline{X}_n - \mu_X}{\sqrt{\frac{N-n}{N-1}\frac{\sigma_X^2}{n}}}\right| < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha\,.$$

Entonces

$$I_{\mu_X}^{100(1-\alpha)\%} = \left[\overline{X}_n \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N-1}}\frac{\sigma_X}{\sqrt{n}}\right].$$

Por otro lado, si se desea que $P\left(\left|\overline{X}_n - \mu_X\right| < e\right) = 1-\alpha$, bastará con tomar un tamaño muestral n tal que

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{N-n}{N-1}}\frac{\sigma_X}{\sqrt{n}} = e\,,$$

de donde resulta que

$$n = \frac{z_{1-\frac{\alpha}{2}}^{\ 2} N\sigma_X^2}{(N-1)e^2 + z_{1-\frac{\alpha}{2}}^{\ 2}\sigma_X^2}.$$

Si la varianza poblacional es desconocida, puede optarse por tomar un tamaño muestral que garantice que no se supera el margen de error admisible con la probabilidad prefijada, es decir,

$$n = máx\left\{\frac{z_{1-\frac{\alpha}{2}}^{\ 2} N\sigma_X^2}{(N-1)e^2 + z_{1-\frac{\alpha}{2}}^{\ 2}\sigma_X^2}\right\},$$

aunque, en este caso, seguramente se está consiguiendo mayor precisión de la deseada con un coste superior al estrictamente necesario. En este sentido, otra opción consiste en sustituir σ_X^2 por el valor de la cuasivarianza muestral, S^2, obtenida para una muestra piloto.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Los procedimientos concretos de contraste de hipótesis simples se abordan en el resto del capítulo, mientras que las hipótesis compuestas se estudian en el siguiente. El epígrafe segundo de este capítulo presenta el teorema de Neyman-Pearson y, finalmente, se ilustra dicho teorema con algunas aplicaciones al contraste de hipótesis simples.

17.1 CONCEPTOS BÁSICOS: HIPÓTESIS, REGIÓN CRÍTICA Y TIPOS DE ERROR

Una hipótesis estadística es una afirmación con respecto a alguna característica poblacional y, por consiguiente, un enunciado sobre una variable aleatoria en relación a su ley de probabilidades. Admitiendo que la distribución de la variable aleatoria poblacional es de determinado tipo, puede elaborarse un enunciado sobre el parámetro del que depende la distribución de dicha variable aleatoria. Este enunciado define una hipótesis paramétrica. Pero, si no se conoce el tipo de distribución poblacional, también pueden formularse enunciados sobre características de la población o incluso sobre el tipo de distribución. Estas hipótesis, denominadas no paramétricas, se analizarán más adelante, mientras que en este capítulo y el siguiente se abordará el contraste de hipótesis paramétricas.

Sea $\{X_1,...,X_n\}$ una muestra aleatoria de una variable aleatoria poblacional X con distribución de tipo conocido tal que $F_X(t)=h(t,\gamma)$, siendo γ un parámetro desconocido perteneciente a un determinado conjunto paramétrico, es decir, $\gamma\in\Gamma$. Sobre este parámetro pueden formularse dos tipos de hipótesis: simples, de la forma $\gamma=\gamma_0$, o compuestas, tales como $\gamma\neq\gamma_0$, $\gamma>\gamma_0$ o $\gamma\leq\gamma_0$. Nótese que una hipótesis simple especifica una distribución única; mientras que, en el caso de una hipótesis compuesta, un conjunto de distribuciones pueden ser compatibles con la hipótesis.

En cualquier caso, en un contraste de hipótesis en la línea de Neyman y Pearson (1933) se formulan dos hipótesis: la hipótesis nula, H_0, cuya veracidad se pone en duda, y la alternativa, H_1, que se asumirá en caso de que la muestra aporte evidencia suficiente para rechazar la hipótesis nula. La finalidad del contraste de hipótesis es proporcionar una regla de decisión que permita rechazar o no la hipótesis nula. En definitiva, se trata de dividir el espacio muestral —es decir, el conjunto de resultados posibles de la muestra aleatoria— en dos regiones, de modo que: si la muestra observada pertenece a la llamada región crítica, se rechazará la hipótesis nula, mientras que si la realización de la muestra aleatoria no se ubica en la región crítica, se concluye que no existe evidencia estadística suficiente para rechazar la anterior hipótesis. En términos formales, dada una muestra aleatoria de tamaño n y un punto muestral $(x_1,...,x_n)\in\Re^n$, se trata de particionar el espacio muestral $\Re^n$ en dos conjuntos $C:\{(x_1,...,x_n)\in\Re^n/\,se\ rechaza\ H_0\}$ y $A:\{(x_1,...,x_n)\in\Re^n/\,no\ se\ rechaza\ H_0\}$. Es decir: si $(x_1,...,x_n)\in C$, la hipótesis nula se rechaza; y si $(x_1,...,x_n)\in A$, la hipótesis nula no se rechaza. Al conjunto C de observaciones muestrales que conducen a rechazar H_0 se le denomina región crítica o región de rechazo. Y al conjunto A de

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

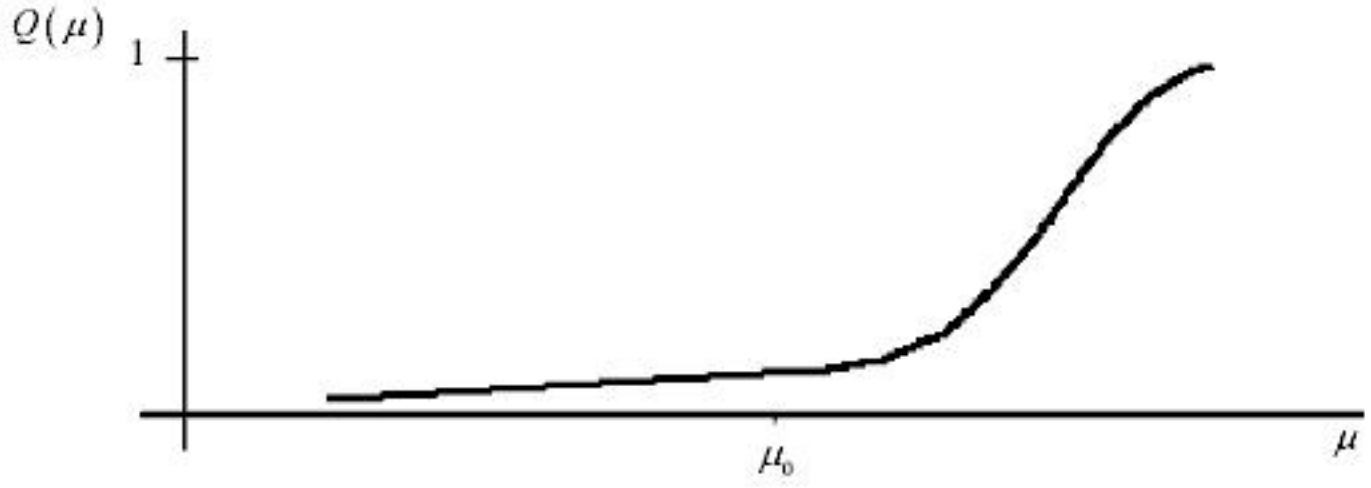
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

y, dado que, bajo la hipótesis alternativa, $\overline{X}$ es $N\left(\mu, \frac{1}{n}\right)$, con $\mu > \mu_0$,

$$Q(\mu) = 1 - \left(N_Z\left(\sqrt{n}\,(k + \mu_0 - \mu)\right)\right), \quad \mu > \mu_0 .$$

Entonces, el comportamiento del *test* es el siguiente.



Nótese, además, que, por lo general, al aumentar el tamaño de la muestra, aumenta la potencia de un buen *test*.

17.2 CONTRASTE DE HIPÓTESIS SIMPLES: TEOREMA DE NEYMAN-PEARSON

En el caso de hipótesis simples, el teorema de Neyman-Pearson permite construir la región crítica más potente para un tamaño determinado del *test*. La idea básica de este teorema es evaluar la verosimilitud de la muestra observada en las dos situaciones consideradas posibles, de modo que si la verosimilitud de la muestra bajo la hipótesis alternativa es suficientemente grande en relación con la verosimilitud de dicha muestra bajo la hipótesis nula, la decisión será rechazar esta última hipótesis.

Sea $\{X_1,\ldots,X_n\}$ una muestra aleatoria de una variable aleatoria poblacional X con distribución de tipo conocido tal que $F_X(t) = h(t,\gamma)$, siendo γ un parámetro desconocido perteneciente a un determinado conjunto paramétrico Γ tal que $\Gamma: \{\gamma_0, \gamma_1\}$. Suponga que se desea efectuar el contraste de hipótesis simples

$$H_0 : \gamma = \gamma_0$$
$$H_1 : \gamma = \gamma_1$$

El teorema de Neyman-Pearson permite deducir que, si existe la región crítica

$$C : \left\{ (x_1,\ldots,x_n) \in \Re^n \Big/ \frac{L(x_1,\ldots,x_n;\gamma_1)}{L(x_1,\ldots,x_n;\gamma_0)} \geq k \right\},$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

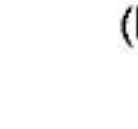
You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

la probabilidad de error de tipo I es $\alpha = 0.0565$.

Suponga ahora que 10 de las 20 personas encuestadas declaran que van a votar por el partido considerado. Entonces $\sum_{i=1}^{20} x_i = 10$, de modo que la muestra observada no pertenece a la región crítica y, en consecuencia, la decisión adoptada al 94.35% de confianza es que no puede rechazarse la hipótesis nula.



EJERCICIOS

17.1. Suponga que la edad a la que un individuo de cierta población accede a su primer empleo puede modelarse a través de una variable aleatoria X con media desconocida μ. Suponga además que, a partir de una muestra aleatoria de tamaño n de la variable aleatoria X, se desea efectuar el contraste de hipótesis

$$H_0 : \mu \geq 30$$

$$H_1 : \mu < 30$$

y se proponen las regiones críticas

$$C_1 : \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n / \bar{x} > \mathrm{k} \right\} \qquad C_2 : \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n / \bar{x} < \mathrm{k} \right\}$$

¿Cuál de las dos regiones críticas diría usted que tiene más sentido? Explique por qué.

17.2. Suponga ahora que el ministro de trabajo de un país considera que al menos la mitad de los trabajadores están satisfechos con su trabajo, mientras que los sindicatos opinan que el porcentaje de trabajadores satisfechos es inferior al 50%. Para eliminar esta discrepancia, se ha decidido preguntar a n trabajadores seleccionados aleatoriamente y cuyas respuestas son independientes y utilizar la proporción de trabajadores encuestados que se declaran satisfechos como elemento para decidir si tiene razón el ministro o los sindicatos.

(a) Si tanto el ministro como los sindicatos están de acuerdo en que sería muy grave asumir que al menos la mitad de los trabajadores están satisfechos si la realidad es otra, ¿cómo plantearía usted un contraste de hipótesis que ayudara a resolver la cuestión?

(b) Suponga que se define p como la proporción desconocida de trabajadores satisfechos en la población y se plantea el contraste de hipótesis

$$H_0 : p < 0.5$$

$$H_1 : p \geq 0.5$$

Además, se define $\bar{x}$ como la proporción de trabajadores encuestados que se declaran satisfechos y se proponen las regiones críticas

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

$$\text{(a.2)}\quad s_1^2=\frac{\sum_{j=1}^{10}\left(x_{1,j}-\bar{x}_1\right)^2}{9}=142.456\,,\quad s_2^2=\frac{\sum_{j=1}^{10}\left(x_{2,j}-\bar{x}_2\right)^2}{9}=141.956\,,$$

$$s_3^2=\frac{\sum_{j=1}^{10}\left(x_{3,j}-\bar{x}_3\right)^2}{9}=133.822\,.$$

$$\text{(a.3)}\quad \sum_{i=1}^{3}10\left(\bar{x}_i-\bar{x}\right)^2=166.867\,,\quad \sum_{i=1}^{3}\sum_{j=1}^{10}\left(x_{i,j}-\bar{x}_i\right)^2=3764.1\,.$$

(b) Con un 95% de confianza, ¿diría usted que la variabilidad en el grado de satisfacción es diferente según la licenciatura?

(c) Teniendo en cuenta el resultado anterior y utilizando también un 95% de confianza, ¿cree usted que el grado de satisfacción con los estudios es, en términos medios, el mismo para las tres licenciaturas consideradas?

19.2. Suponga que se están estudiando las diferencias salariales entre los trabajadores de cuatro empresas, A, B, C y D, dedicadas a la misma actividad. Con este fin se ha seleccionado aleatoriamente una muestra de ocho trabajadores de cada una de estas empresas. Suponga que los salarios anuales de estos trabajadores, expresados en miles de euros, son los que se recogen en la tabla siguiente y asuma que las respuestas de los trabajadores seleccionados de cada empresa constituyen valores de muestras aleatorias independientes de distribuciones normales.

A	9.6	10.8	11.6	12.2	12.4	11.8	11.2	10.2
B	10.1	10.3	11.9	12.6	12.3	12.7	11.1	11.5
C	10.6	10.9	11.2	11.5	11.7	10.2	9.5	8.6
D	11.0	13.8	13.6	11.6	12.2	13.2	13.0	12.6

(a) Compruebe que, si se define $X_{i,j}$ como la variable aleatoria que recoge el salario del individuo $j-ésimo$ de la muestra de trabajadores de la empresa $i-ésima$, siendo $i=A,B,C,D$, entonces:

$$\text{(a.1)}\quad \bar{x}=\frac{1}{32}\sum_{i=A,B,C,D}\sum_{j=1}^{8}x_{i,j}=11.484\,,\quad \bar{x}_A=\frac{1}{8}\sum_{j=1}^{8}x_{A,j}=11.225\,,$$

$$\bar{x}_B=\frac{1}{8}\sum_{j=1}^{8}x_{B,j}=11.562\,,\quad \bar{x}_C=\frac{1}{8}\sum_{j=1}^{8}x_{C,j}=10.525\,,\quad \bar{x}_D=\frac{1}{8}\sum_{j=1}^{8}x_{D,j}=12.625\,.$$

$$\text{(a.2)}\quad s_A^2=\frac{\sum_{j=1}^{8}\left(x_{A,j}-\bar{x}_A\right)^2}{7}=0.954\,,\quad s_B^2=\frac{\sum_{j=1}^{8}\left(x_{B,j}-\bar{x}_B\right)^2}{7}=0.997\,,$$

$$s_C^2=\frac{\sum_{j=1}^{8}\left(x_{C,j}-\bar{x}_C\right)^2}{7}=1.114\,,\quad s_D^2=\frac{\sum_{j=1}^{8}\left(x_{D,j}-\bar{x}_D\right)^2}{7}=0.954\,.$$

$$\text{(a.3)}\quad \sum_{i=A,B,C,D}8\left(\bar{x}_i-\bar{x}\right)^2=18.358\,,\quad \sum_{i=A,B,C,D}\sum_{j=1}^{8}\left(x_{i,j}-\bar{x}_i\right)^2=28.124\,.$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

es una región crítica apropiada con probabilidad de error de tipo I igual a α, dado que

$$P\left(C/H_0\right)=P\left(\sum_{i=1}^{k}\frac{\left(X_i-np_{i,0}\right)^2}{np_{i,0}}>\chi^2_{k-1;1-\alpha}\Big/H_0\right)=\alpha\,.$$

Nótese que, si la hipótesis nula es cierta, $E[X_i]=np_{i,0}$, $i=1,\ldots,k$, y, por tanto, se espera que el valor del estadístico U sea pequeño. De modo que si el valor del estadístico U es suficientemente grande, parece razonable rechazar la hipótesis nula.

Ejemplo 20.1 Suponga que cuatro partidos políticos, A, B, C y D concurren a unas elecciones en una determinada población y se desea contrastar, con un 95% de confianza, si los cuatro partidos políticos tienen la misma proporción de votantes en la población. Para ello, se ha decidido preguntar por su intención de voto a 1000 individuos de la población seleccionados aleatoriamente y cuyas respuestas son independientes. Suponga que las respuestas de los 1000 individuos son tales que 220 de ellos declaran que van a votar por el partido A, 260 se decantan por el partido B, 270 lo hacen por el partido C y los 250 restantes eligen el partido D.

Nótese que las respuestas de los 1000 individuos son los resultados de 1000 pruebas multinomiales de tamaño 4 del tipo $\Omega:\{A,B,C,D\}$ que pueden considerarse independientes y tales que $P(\{A\})=p_A$, $P(\{B\})=p_B$, $P(\{C\})=p_C$ y $P(\{D\})=p_D$. Si se definen las variables aleatorias X_1, X_2, X_3 y X_4, que recogen el número de individuos de la muestra que declaran que van a votar, respectivamente, por los partidos políticos A, B, C y D, entonces (X_1,X_2,X_3) es $MB(1000,p_A,p_B,p_C)$ y se desea efectuar el contraste

$$H_0: p_A=\frac{1}{4},p_B=\frac{1}{4},p_C=\frac{1}{4},p_D=\frac{1}{4}$$

$$H_1: H_0 \ \textit{falsa}$$

de modo que puede utilizarse la región crítica

$$C:\left\{(x_1,x_2,x_3,x_4)\Big/\ u=\sum_{i=1}^{4}\frac{\left(x_i-1000\frac{1}{4}\right)^2}{1000\frac{1}{4}}>\chi^2_{3;0.95}\right\}.$$

En la muestra observada se tiene que $(x_1,x_2,x_3,x_4)=(220,260,270,250)$, de manera que el valor del estadístico de contraste es $u=\sum_{i=1}^{4}\frac{(x_i-250)^2}{250}=5.6$ y, dado que $\chi^2_{3;0.95}=7.815$, el resultado del contraste es que no se rechaza la hipótesis nula. Es decir, las respuestas observadas no aportan evidencia suficiente para rechazar que los cuatro partidos políticos poseen la misma proporción de votantes en la población.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Ejemplo 20.4 Sea X una variable aleatoria que recoge el tiempo de residencia de un inmigrante en determinada población expresado en años. Suponga que, a partir de una pequeña muestra aleatoria de 10 inmigrantes, se ha observado la muestra siguiente

$$\{x_1,...,x_{10}\}:\{3.5,5.4,6.2,5.6,5.8,6.1,4.9,4.8,5.9,8.3\}.$$

Entonces, puede contrastarse, al 95% de confianza, si estas observaciones proceden de una distribución normal de media 5 y varianza unitaria empleando el *test* de Kolmogorov-Smirnov. Se desea efectuar el contraste

$$H_0: X \text{ es } \mathrm{N}\left(\mu=5,\sigma^2=1\right)$$

$$H_1: H_0 \textit{ falsa}$$

Y una región crítica apropiada para efectuar el contraste anterior con un 95% de confianza viene dada por

$$C:\left\{(x_1,...,x_{10})\Big/ D_{10}=\max_t \left\{\left|F_{10}(t)-F_0(t)\right|\right\} > D_{10,0.95}\right\}$$

donde $F_{10}(t)$ es la función de distribución empírica de la muestra y $F_0(t)=N_z\left(\frac{t-5}{1}\right)$. Los valores de ambas funciones de distribución se recogen en la siguiente tabla. Dado que interesa evaluar el máximo de la diferencia entre ambas funciones de distribución, conviene evaluar $\left|F_{10}(t)-F_0(t)\right|$ para cada uno de los valores de la muestra, pero también es interesante evaluar $\left|F_{10}(t)-F_0(t^*)\right|$ siendo t^* el valor de la muestra inmediatamente posterior a otro t. Nótese que en cada intervalo $[t,t^*)$ definido por dos valores muestrales consecutivos, la función de distribución muestral F_{10} permanece constante, mientras que la función de distribución especificada en la hipótesis nula crece desde $F_0(t)$ hasta $F_0(t^*)$. Así pues, la diferencia máxima entre ambas funciones en cada intervalo $[t,t^*)$ será $\left|F_{10}(t)-F_0(t)\right|$, o bien, $\left|F_{10}(t)-F_0(t^*)\right|$. Para valores t inferiores al mínimo muestral, la diferencia máxima es $F_0\left(x_{(1)}\right)$.

t	$F_{10}(t)$	$F_0(t)$	$\left\|F_{10}(t)-F_0(t)\right\|$	$\left\|F_{10}(t)-F_0(t^*)\right\|$
$x_{(1)}=3.5$	0.1	0.0668	0.0332	0.3207
$x_{(2)}=4.8$	0.2	0.4207	0.2207	0.2602
$x_{(3)}=4.9$	0.3	0.4602	0.1602	0.3554
$x_{(4)}=5.4$	0.4	0.6554	0.2554	0.3257
$x_{(5)}=5.6$	0.5	0.7257	0.2257	0.2881
$x_{(6)}=5.8$	0.6	0.7881	0.1881	0.2159
$x_{(7)}=5.9$	0.7	0.8159	0.1159	0.1643
$x_{(8)}=6.1$	0.8	0.8643	0.0643	0.0849
$x_{(9)}=6.2$	0.9	0.8849	0.0151	0.0995
$x_{(10)}=8.3$	1	0.9995	0.0005	–

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

20.3 OTROS CONTRASTES NO PARAMÉTRICOS

En este epígrafe podría recogerse una amplia variedad de contrastes no paramétricos. Por ejemplo, para contrastar la aleatoriedad de una muestra, el *test* de rachas de Wald-Wolfowitz —apto para características poblacionales dicotómicas o magnitudes cuyos valores puedan agruparse en dos categorías, por ejemplo, a un lado y a otro de la mediana muestral— o también el contraste del cuadrado medio de diferencias sucesivas —más apropiado en el caso de datos cuantitativos— constituyen pruebas adecuadas (Ruiz-Maya y Martín, 1995:623-630; Casas, 1996:549-567). Pero sólo se exponen algunos *tests* de localización, cuyo objetivo es contrastar una hipótesis referente al valor de alguna medida de posición, así como otros contrastes de comparación de poblaciones.

Para una distribución poblacional continua, puede formularse un contraste unilateral o bilateral sobre cualquier cuantil k y utilizar el denominado *test* de signos. A continuación se expone el procedimiento del contraste de signos para la mediana, un caso particular del contraste anterior[2]. Sea X una variable aleatoria continua con distribución desconocida y mediana Me_X también desconocida. Sea $\{X_1,...,X_n\}$ una muestra aleatoria de la variable aleatoria X. Y suponga que se desea efectuar el contraste

$$H_0 : Me_X = m_0$$
$$H_1 : Me_X \neq m_0$$

Si se define la variable aleatoria S^+ que recoge el número de observaciones muestrales mayores que m_0, se tiene que S^+ es $B\left(n,\frac{1}{2}\right)$, si la hipótesis nula es cierta. Entonces, una región crítica apropiada con probabilidad de error de tipo I menor o igual que α, viene dada por

$$C:\left\{(x_1,...,x_n) \Big/ s^+ \leq k_{\frac{\alpha}{2}} \ \vee \ s^+ \geq k_{1-\frac{\alpha}{2}}\right\},$$

donde $k_{\frac{\alpha}{2}}$ y $k_{1-\frac{\alpha}{2}}$ son valores enteros tales que

$$k_{\frac{\alpha}{2}} = máx\left\{k / P(S^+ \leq k / Me_X = m_0) \leq \frac{\alpha}{2}\right\}$$

y

$$k_{1-\frac{\alpha}{2}} = mín\left\{k / P(S^+ \geq k / Me_X = m_0) \leq \frac{\alpha}{2}\right\} .$$

---

[2] Si se supone además que la distribución poblacional es simétrica, el contraste de rangos-signos de Wilcoxon constituye una prueba no paramétrica sobre la mediana que tiene en cuenta los signos —es decir, si los valores muestrales son mayores o menores que el valor mediano especificado en la hipótesis nula—, y también la magnitud de las diferencias entre los valores observados y la mediana (Siegel, 1978:99-108; Ruiz-Maya y Martín, 1995:637-640; Casas, 1996:582-590).

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

(b) En el último año, se ha seleccionado una muestra aleatoria de 1200 clientes y se ha encontrado ahora que los clientes han acudido al comercio como se muestra a continuación.

Mañana	Tarde	Noche
650	310	240

De acuerdo con estos datos y con un 95% de confianza, ¿qué conclusión obtiene usted sobre la convicción del gerente? Explique por qué cambia la conclusión a pesar de que las proporciones muestrales de clientes en cada turno no hayan cambiado.

20.2. Suponga que una fábrica dispone de una máquina que va generando una tira continua de plástico de forma ininterrumpida. Los técnicos de la fábrica desean comprobar si la superficie de plástico generada por la máquina desde que empieza la jornada hasta que aparece el primer fallo, en metros cuadrados, sigue una distribución exponencial. Para ello, han registrado esta superficie en 100 jornadas laborales seleccionadas aleatoriamente. Los resultados obtenidos son los siguientes.

Superficies	Frecuencia
$(0,1]$	53
$(1,2]$	26
$(2,3]$	16
$(3,\infty)$	5

Sabiendo además que la media de las superficies registradas hasta que aparece el primer fallo en las 100 jornadas fue de 2 metros cuadrados, ¿qué decisión estadística adoptaría usted, al 95% de confianza, sobre la hipótesis que desean contrastar los técnicos?

20.3. Suponga que en un determinado país, los miembros de un partido político asumen que la proporción de votantes del partido, X, en una cualquiera de las provincias en las que se celebran elecciones para el gobierno del estado sigue una distribución uniforme en el intervalo $(0.3,0.4)$. Suponga que, a partir de una muestra aleatoria de 10 provincias en las que se han celebrado elecciones recientes, se han observado las proporciones de voto:

$$\{0.386,0.328,0.353,0.336,0.347,0.364,0.372,0.319,0.393,0.324\}.$$

De acuerdo con esta información, contraste la hipótesis asumida por los miembros del partido al 95% de confianza.

20.4. Suponga que en determinada negociación sindical se asume que el salario anual de un trabajador de una empresa, expresado en miles de euros, X, sigue una distribución normal. Para contrastar este supuesto, se ha seleccionado una muestra aleatoria de 10 trabajadores y se han observado los salarios:

$$\{36,28,35,26,17,25,27,31,33,32\}.$$

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Índice analítico

D

E

F

N

O

P

R

S

T

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

Esta obra ofrece al estudiante de ciencias sociales unos conocimientos suficientes para que, en el desempeño de su actividad profesional, sea capaz de identificar qué problemas exigen el recurso a los métodos estadísticos. Con este objeto, se introduce un conjunto de técnicas que le permitan afrontar con éxito la resolución de dichos problemas o, cuando menos, situarlo en disposición de abordar procedimientos estadísticos más complejos que puedan resultarle útiles. La obra está elaborada a partir de la aceptación de que la estadística debe impartirse con carácter particular para la disciplina a la que sirve de herramienta. Y asumiendo que el razonamiento estadístico debe ser prioritario frente a la demostración matemática, este libro pretende facilitar al profesor de estadística su tarea de ayudar a sus estudiantes a «entender la lógica estadística» en lugar de dirigirse hacia la aplicación «ciega» de las técnicas estadísticas.

José Juan Cáceres Hernández es Profesor Titular de Economía Aplicada del Departamento de Economía de las Instituciones, Estadística Económica y Econometría de la Universidad de La Laguna. Sus líneas de investigación preferentes cubren aspectos metodológicos relacionados con el análisis de series temporales, los algoritmos genéticos y los modelos de elección discreta, así como aplicaciones en diversos ámbitos económicos.

DELTA PUBLICACIONES

C/ Luarca, 11
28230 Las Rozas de Madrid
MADRID
Tel. 91 637 16 88

www.deltapublicaciones.com