Estadística Descriptiva

Coordinadora:
M.ª Dolores Sarrión Gavilán

Autores:
M.ª Dolores Benítez Márquez
José Luis Iranzo Acosta
Fernando Isla Castillo
M.ª Dolores Sarrión Gavilán

Mc Graw Hill

ESTADÍSTICA DESCRIPTIVA

McGraw-Hill/Interamericana de España, S.L.
Edificio Valrealty, 1.ª planta
Basauri, 17
28023 Aravaca (Madrid)

ISBN: 978-84-481-8331-8
Depósito legal: M-31107-2012

Editora: María León
Director de Educación y Desarrollo de Negocio: Álvaro García Tejeda
Diseño de cubierta: David Santás
Composición: Artesa, S.L.
Impresión: Producciones Digitales Pulmen, S.L.L.
4123567890 - 8765432019

IMPRESO EN ESPAÑA-PRINTED IN SPAIN

Prólogo

Este libro, en el que presentamos los contenidos de un curso básico de Estadística Descriptiva, es fruto de nuestra experiencia como profesores de la Universidad de Málaga en el Departamento de Estadística y Econometría (Economía Aplicada) y forma parte de un proyecto interuniversitario más ambicioso que está encaminado a facilitar tanto el aprendizaje de esta materia por parte de los alumnos, como la labor docente de los profesores universitarios que la imparten.

El libro *Estadística Descriptiva* está pensado para ser utilizado como texto de referencia en cursos introductorios de Estadística en titulaciones de grado del área de Ciencias Sociales, estando especialmente orientado a los ámbitos económico-financiero, empresarial, turístico, de las relaciones laborales y de la administración pública.

Teniendo en cuenta el tipo de alumno al que el texto va dirigido, la exposición de los contenidos teóricos se acompaña de numerosos ejemplos resueltos y ejercicios propuestos, que ayudan al alumno a asimilar los diferentes conceptos y sus aplicaciones facilitándole el estudio de la materia. Los enunciados de la mayor parte de los ejercicios y ejemplos resueltos están basados en datos reales relativos a algún aspecto de los ámbitos anteriormente mencionados.

Los contenidos expuestos en el libro de texto se estructuran en cinco capítulos que están dedicados a la presentación y organización de los datos; el análisis descriptivo de las distribuciones de una variable; los números índices; el análisis de la correlación y regresión en distribuciones de dos variables y, por último, el análisis clásico de las series temporales.

El texto se complementa con el acceso a una página web asociada, creada expresamente para facilitar la labor docente del profesor y el aprendizaje de los alumnos (www.mhe.es/estadisticadescriptiva), en la que colaboran los autores y un grupo de profesores de otras universidades, y en la que se proporciona material adicional, tanto para el alumno como para el profesor. Concretamente, en dicha página están disponibles para los alumnos los ficheros de datos de los ejemplos y ejercicios que figuran en el libro; la solución detallada de algunos de ellos con diversas aplicaciones informáticas (Excel, R o SPSS); algunos cuestionarios de autoevaluación; las soluciones numéricas de los ejercicios propuestos en el libro de texto y una colección de ejercicios propuestos. Además de lo anterior, el profesor podrá acceder a las soluciones detalladas de los ejercicios propuestos, modelos de examen, presentaciones PowerPoint en formato digital descargable, cuestionarios de evaluación en formato Moodle, etc.

Por último, deseamos expresar nuestra gratitud a las profesoras Carmen Muñoz, Núria Viladomiu y Victoria Alea, de la Universidad de Barcelona, tanto por la revisión

del material preliminar como por los ejercicios sugeridos; a las profesoras Raquel Caro y Ana María Lara, de la Universidad Pontificia de Comillas y la Universidad de Granada, respectivamente, que han proporcionado material para incorporar a la página web asociada y, por último, a la editorial McGraw-Hill, por la oportunidad que nos ha brindado de participar en este proyecto y facilitar la publicación de este libro de texto.

Málaga, agosto de 2012

Los autores

María Dolores Benítez Márquez
José Luis Iranzo Acosta
Fernando Isla Castillo
María Dolores Sarrión Gavilán

Contenido

CAPÍTULO 1

Conceptos básicos

En esta primera lección se definen los conceptos de población, elemento y variable, que son básicos en todo estudio estadístico. Se distingue entre observación exhaustiva y observación parcial y se introduce el significado de las expresiones «Estadística Descriptiva» y «Estadística Inductiva» o «Inferencia Estadística».

Inicialmente, los datos relativos a la observación de una variable aparecen en el orden y modo en que fueron recogidos y registrados, aparentemente de forma desordenada e inconexa, lo que puede hacer difícil su interpretación. El punto de partida para extraer la información contenida en los datos es su resumen y presentación en forma de tabla.

Centrándonos en datos transversales relativos a una variable, para obtener la tabla contabilizamos el número de elementos que presenta cada valor o modalidad (frecuencia absoluta) o la proporción que dichos elementos representan del total (frecuencia relativa). El conjunto formado por los distintos valores o modalidades junto con su correspondiente frecuencia se presenta en forma de tabla y constituye la denominada distribución de frecuencias.

Si el número de valores distintos que la variable presenta es grande, es aconsejable agruparlos en clases o intervalos para mayor comodidad en el tratamiento de la información.

Por otra parte, cuando las modalidades admiten una ordenación resulta útil la información relativa al número de elementos que presentan una modalidad o alguna de las anteriores, suponiendo que las modalidades están ordenadas en orden creciente (frecuencia acumulada).

En esta primera fase de resumen y exploración previa de la información, resulta bastante útil presentar la información en gráficos, de modo que la referencia visual sirva como punto de partida para el análisis estadístico. Los distintos tipos de datos (cualitativos o cuantitativos) y el que la distribución de frecuencias sea para datos agrupados o no agrupados determinarán el que los representemos gráficamente mediante: diagrama de barras (para cualitativos o cuantitativos no agrupados); tallo y hojas (para cuantitativos no agrupados); histograma y polígono de frecuencias (para datos agrupados) y diagrama de sectores, cartograma, pictograma o barra de componentes porcentuales (generalmente empleados para datos cualitativos). Por su parte, la representación

gráfica de la distribución de frecuencias acumuladas, en el caso de que tenga sentido obtenerla, será un diagrama de escalera o uno poligonal, según que los datos sean no agrupados o agrupados.

Finalizaremos la lección haciendo un breve resumen de las fuentes estadísticas más relevantes para el análisis de los ámbitos económico-financiero y turístico a nivel nacional o autonómico, indicando el tipo de información estadística que en ellas podemos encontrar.

1.1. Noción de Estadística. Estadística Descriptiva e Inferencia Estadística

En el uso común el término estadística se refiere a toda información numérica, presentada de forma ordenada y habitualmente acompañada de algún gráfico que facilita su comprensión. Esta primera conceptualización, con orígenes históricos muy remotos, cada día está más arraigada en la sociedad actual, inmersa en un mundo de cifras que llenan los medios de comunicación e impregnan nuestras referencias personales.

Pero la estadística en su concepción actual es mucho más. Es una ciencia joven que se ha consolidado a lo largo del siglo xx y que ha ido ampliando su campo de aplicación hasta alcanzar a casi todas las ramas del conocimiento y a las diferentes parcelas de la actividad humana, debido a que proporciona criterios para la toma de decisiones en situaciones de incertidumbre. Esta concepción arranca, básicamente, de la confluencia de dos corrientes conceptuales que discurrieron por caminos separados hasta finales del siglo xix: la Teoría de la Probabilidad y la Estadística en el sentido tradicional del término que la identifica con «información de los estados políticos» (etimológicamente, el vocablo estadística procede del término latino *status*).

Desde un punto de vista analítico, dentro del conjunto de conocimientos que componen la ciencia Estadística se pueden distinguir tres ramas claramente diferenciadas:

- **Estadística Descriptiva,** que se ocupa de la obtención, clasificación, representación y análisis de los datos relativos a las características de interés observados en todos los individuos de un colectivo (población o muestra) con el objeto exclusivo de describir las regularidades o el comportamiento de ese colectivo, sin pretender extraer conclusiones que trasciendan al conjunto analizado. Este tipo de estudio es de naturaleza deductiva, ya que a partir de un conjunto genérico de datos se extraen conclusiones particulares de los mismos.
- **Teoría de la Probabilidad,** que tiene por objeto el estudio de los fenómenos aleatorios y que propiamente es una rama de las Matemáticas. Se trata de una teoría de base axiomática cuyos teoremas se obtienen mediante un razonamiento lógico-deductivo. El concepto de probabilidad, sobre el que ha tenido gran influencia el desarrollo de las ciencias experimentales, ha sido objeto de continuos debates al existir diversas y encontradas opiniones sobre su significado e interpretación, e incluso sobre los contextos en que ha de utilizarse una u otra concepción del mismo.
- **Inferencia Estadística,** que se nutre de las dos partes anteriores y es a la que, usualmente, se hace referencia al hablar de la Estadística como ciencia.

Su objetivo consiste en obtener conclusiones para la población objeto de estudio, mediante el instrumental que le proporciona el cálculo de probabilidades, a partir de la observación y análisis de una muestra representativa de la población. Se realiza, por lo tanto, un proceso inductivo mediante el cual se trasciende el conjunto analizado.

Entre las distintas definiciones que recogen la finalidad y los métodos de esta disciplina incluimos la siguiente:

> «La Estadística se configura como la tecnología del método científico que proporciona instrumentos para la toma de decisiones cuando estas se adoptan en ambiente de incertidumbre, siempre que esta incertidumbre pueda ser medida en términos de probabilidad. Por ello, la Estadística se preocupa de los métodos de recogida y descripción de datos, así como de generar técnicas para el análisis de esta información»[1].

1.2. Definiciones básicas

Toda investigación estadística está referida a un colectivo o grupo que recibe el nombre de **población.** La población puede estar formada por personas, pero también por entes de la más diversa naturaleza: periodos de tiempo, actos jurídicos, áreas geográficas, hoteles, etc. El primer paso de cualquier estudio estadístico es definir con precisión la población sobre la que se realizará el estudio.

Las personas o cosas que integran la población se llaman **elementos** o **unidades estadísticas.** Estos elementos son entes observables y pueden tener existencia real (personas, coches, casas...) o pueden haber sido creados artificialmente para facilitar la investigación estadística (parcelación de terrenos agrarios para estudiar su productividad, intervalos de tiempo que se toman como unidad para analizar un fenómeno a lo largo del tiempo...).

Los elementos de la población poseen ciertas propiedades, cualidades o rasgos a los que llamamos **caracteres** o **variables estadísticas** (sexo, edad, estatura, ingresos..., para poblaciones de personas; extensión, número de estrellas, número de habitaciones, antigüedad..., para poblaciones de hoteles; número de trabajadores, salario medio mensual, número medio mensual de bajas por enfermedad..., para poblaciones formadas por empresas; etc.). La investigación estadística se centrará en el estudio de uno o varios de esos caracteres que son comunes a los elementos de la población.

A los caracteres que son susceptibles de ser medidos, es decir, a aquellos en los que el resultado de su observación se puede cuantificar de una manera objetiva, como son: la edad, el peso o la estatura de las personas, se les denomina **variables cuantitativas** o **simplemente variables,** por ser las más frecuentes en los estudios estadísticos. Los restantes forman el grupo de las **variables cualitativas** o **atributos.** A las distintas formas de presentación (manifestaciones observadas) de un atributo se les llama **modalidades** o **categorías.**

Las variables cuantitativas se clasifican, atendiendo al número de valores que pueden tomar, en **discretas** y **continuas.** Las variables discretas son aquellas que toman

[1] Martín Pliego, F. J. (1994): *Introducción a la estadística económica y empresarial,* A.C., Madrid.

valores aislados, a lo sumo en cantidad numerable (número de hijos de la pareja, número de habitaciones de un hotel, número de clientes de un establecimiento, etc.). Las variables continuas son las que pueden tomar cualquier valor en un intervalo dado; por ejemplo, la altura de las personas comprendidas entre dos tallas fijas, los salarios percibidos por los trabajadores de un determinado sector, las distancias recorridas por los turistas que llegan a un cierto destino desde distintos puntos, el ingreso familiar mensual de las familias de un determinado barrio, la tasa mensual de parados en el sector industrial, etc.[2]

Por otra parte, cuando las observaciones de una variable estadística tienen una referencia temporal (consumo mensual de electricidad de una determinada familia, temperatura diaria en el observatorio del aeropuerto de Málaga, número semanal de pernoctaciones en un determinado hotel de la costa, número mensual de parados en Málaga en el sector de la construcción, etc.), distinguiremos entre **flujo** y ***stock.*** Las variables que expresan **flujos** son aquellas para las que el valor que asignamos a un periodo de tiempo corresponde a todo él tomado como unidad, mientras que las correspondientes a variables que expresan ***stocks*** son aquellas para las que el valor que asignamos a un periodo de tiempo corresponde a una medición realizada en un instante del mismo que se toma como representante. Así por ejemplo, entre las variables citadas anteriormente, el consumo mensual de electricidad de una determinada familia sería flujo, la temperatura diaria en el observatorio del aeropuerto de Málaga sería *stock*, el número semanal de pernoctaciones en un determinado hotel de la costa se considera flujo, y el número mensual de parados en Málaga en el sector de la construcción, *stock*. (Figura 1.1.)

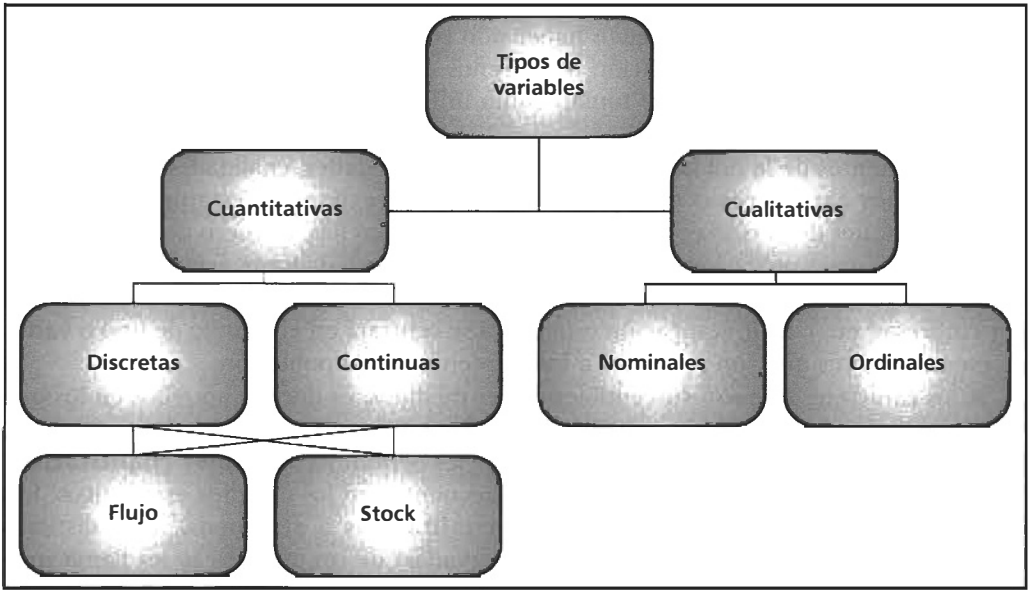


Figura 1.1. Clasificación de los caracteres y variables estadísticas

[2] Esta distinción es buena en teoría, sin embargo, en la práctica todas las variables son discretas, ya que los instrumentos de medida que utilizamos para observarlas tienen una precisión limitada.

Observemos que en las variables que expresan flujos tiene sentido la acumulación de los valores observados en distintos subperiodos de uno más amplio para obtener el valor correspondiente a este último (por ejemplo, el dato semanal como suma de los datos diarios), mientras que en las que expresan *stocks* esto no tiene sentido.

En cuanto a los atributos, se clasifican en **ordinales** y **nominales,** según que sus categorías sean susceptibles de ser ordenadas o no lo sean. Así, el color de los ojos es nominal, el tipo de alojamiento elegido por los turistas que visitan una ciudad es nominal y el nivel de estudios del cabeza de familia de los hogares de un determinado barrio de la ciudad es ordinal.

Una vez determinados los caracteres a estudiar se procede a observarlos. Esta observación puede ser **exhaustiva,** si se observa la variable de interés en todos y cada uno de los elementos de la población; **parcial,** cuando observamos la característica de interés en una parte de la población y **mixta,** si algunas características se observan exhaustivamente y otras solo parcialmente.

A su vez, la observación parcial puede hacerse en una **subpoblación** (subconjunto de elementos de la población con alguna característica en común que los distingue del resto) o en una **muestra** (los elementos que la componen no presentan ninguna característica especial que los haga diferentes del resto). La muestra puede ser **aleatoria,** si los elementos que la forman han sido seleccionados al azar, o **no aleatoria,** cuando los elementos que la forman han sido seleccionados según ciertos criterios fijados de antemano.

Cuando la observación es exhaustiva, las técnicas de la Estadística Descriptiva permiten obtener conclusiones válidas para toda la población; sin embargo, si la observación es parcial, con estas técnicas solo podremos extraer conclusiones válidas para la subpoblación o muestra que ha sido observada. Si queremos extraer conclusiones válidas para la población a partir de lo observado en una parte de ella, será necesario que esa parte sea aleatoria y recurrir, además de a la descripción de los datos, a la **Inferencia Estadística,** lo que está fuera del alcance de este curso.

Por otra parte, atendiendo a que nuestro interés se centre en conocer la evolución de una cierta característica a lo largo del tiempo o en su comportamiento en un instante del mismo, utilizaremos datos **longitudinales** (o **temporales**), **transversales** o **un panel de datos,** si nuestro interés se centra en ambos aspectos.

Llamamos **observaciones transversales** a las obtenidas de los distintos elementos en un mismo instante o intervalo de tiempo. Por ejemplo, los salarios de distintos obreros, distintas empresas, distintos sectores económicos, pero referidos siempre al mismo tiempo. Sin embargo, si fijado un obrero o una empresa, observamos el salario en distintos instantes o intervalos de tiempo, obtenemos **observaciones temporales.** Dichas observaciones expresan la evolución temporal del carácter estudiado en la unidad estadística sobre la que se realizan las distintas observaciones. Por ejemplo, el volumen salarial de los empleados de una empresa en los últimos diez años.

Cada vez se utilizan con más frecuencia los **paneles de datos,** en los que se combinan observaciones de corte transversal con observaciones temporales. Por ejemplo, volumen salarial de los empleados de 50 empresas del sector de la construcción en Málaga en los últimos 10 años.

Una característica distintiva entre las observaciones temporales y transversales es la de dependencia. Las observaciones temporales suelen ser dependientes, cada observación depende de las anteriores, lo que, en general, no suele ocurrir cuando las observaciones son transversales.

1.3. Presentación de los datos: distribuciones de frecuencias y representaciones gráficas

La realización de una investigación estadística lleva consigo la observación en cada uno de los elementos de la población, subpoblación o muestra, de los caracteres que son objeto de dicha investigación. Lo más frecuente es que se obtenga una gran masa de datos que no hace fácil su interpretación; en ese caso, se requiere un proceso previo de **condensación o reducción,** que puede conllevar la pérdida de parte de la información inicialmente recogida. El resultado de esta operación de reducción es una **tabla estadística,** o simplemente estadística, que contiene de forma ordenada y sistemática un conjunto de datos.

Una estadística se puede presentar, básicamente, de dos formas:

1. Clasificando las observaciones efectuadas sin mencionar a los elementos de la población a los que las mismas corresponden. La forma habitual de presentación en ese caso es una tabla con dos columnas o filas. En la primera aparecerán los distintos valores (o grupos de valores) observados de la variable o las distintas modalidades que el atributo haya presentado. En la segunda columna o fila, al lado o debajo de cada valor o modalidad, escribiremos las veces que el mismo se ha presentado. Se llaman **distribuciones de frecuencias.**
2. Uniendo cada observación al instante de tiempo, área geográfica, sector productivo, etc., al que la misma se refiere. Es el caso de las **estadísticas temporales, geográficas, sectoriales,** etc.

Observación. A veces, una misma estadística puede ser tratada como distribución de frecuencias o como estadística geográfica, sectorial, etc. de una variable cuantitativa, en la que los valores observados se identifican con las frecuencias. Así, si disponemos del número de oficinas bancarias en las distintas provincias andaluzas en 2007, podemos estar interesados en el número medio de oficinas por provincia, en cuyo caso la variable que estamos considerando es «n.º provincial de oficinas bancarias» con ocho observaciones (estadística geográfica). Sin embargo, los valores de esa variable, al ser números enteros, pueden interpretarse también como frecuencias, en ese caso podríamos estar interesados, por ejemplo, en conocer la distribución porcentual por provincias de las oficinas bancarias de Andalucía.

1.3.1. Distribuciones de frecuencias de una variable. Datos no agrupados

Cuando disponemos de pocas observaciones o cuando, aún siendo muchas, aparecen pocos valores distintos, trabajaremos con distribuciones de **datos no agrupados.**

Estas estadísticas se presentan en una tabla con dos filas o columnas. En la primera columna o fila suelen figurar los distintos valores de la variable, frecuentemente ordenados en orden creciente, y en la segunda el número de veces que cada valor ha sido observado, es decir, su **frecuencia absoluta de observación.**

Para un conjunto de **N** datos relativos a una variable **X,** emplearemos la siguiente notación[3]:

[3] Si lo que manejamos es un atributo, al que representamos mediante A, escribiremos $A_1,\ldots,A_k$ para representar las k modalidades distintas que A presenta en ese conjunto de N observaciones.

- Los distintos valores observados de la variable **X**, ordenados de menor a mayor, se representan mediante $x_1, x_2, \ldots, x_k$, $k \leq N$.
- La **frecuencia absoluta** de observación de cada valor x_i de la variable **X** se representa con n_i, que es el número de veces que dicho valor aparece repetido en las observaciones.
- Es claro que $n_1 + n_2 + \ldots + n_k = N$, lo que, abreviadamente, se escribirá

$$\sum_{i=1}^{k} n_i = N$$

Los distintos valores observados acompañados de sus frecuencias se suelen presentar en una tabla como la siguiente:

x_i	x_1	x_2	...	x_i	...	x_k	
n_i	n_1	n_2	...	n_i	...	n_k	**N**

O, alternativamente,

x_i	n_i
x_1	n_1
x_2	n_2
...	...
x_i	n_i
...	...
x_k	n_k
	N

Ejemplo 1.1. En una población de 60 alumnos se ha observado la variable X = «N.º de hermanos». El resultado de dicha observación es el que se muestra a continuación.

3	0	1	4	0	2	3	5	4	5
2	4	3	3	0	7	6	2	1	3
0	3	2	1	1	5	4	1	3	0
5	1	0	1	3	1	1	4	0	2
1	2	2	0	1	0	2	3	2	1
2	5	4	2	3	0	3	2	0	1

Los distintos valores observados de la variable X, ordenados de menor a mayor, son: 0, 1, 2, 3, 4, 5, 6, 7. La tabla que se obtiene después de contar el número de veces que cada uno de ellos aparece repetido es la siguiente:

x_i	n_i
0	11
1	13
2	12
3	11
4	6
5	5
6	1
7	1
	60

Diagrama o gráfico de barras

La representación gráfica más usual de la distribución de frecuencias de una variable con datos no agrupados es el **diagrama de barras.** Para obtenerlo, marcamos en el eje de abscisas los distintos valores observados y sobre cada uno de ellos se levanta una barra vertical de altura igual a su frecuencia absoluta de observación. Por ejemplo, el diagrama de barras para el Ejemplo 1.1 es el que figura en el Gráfico 1.1.

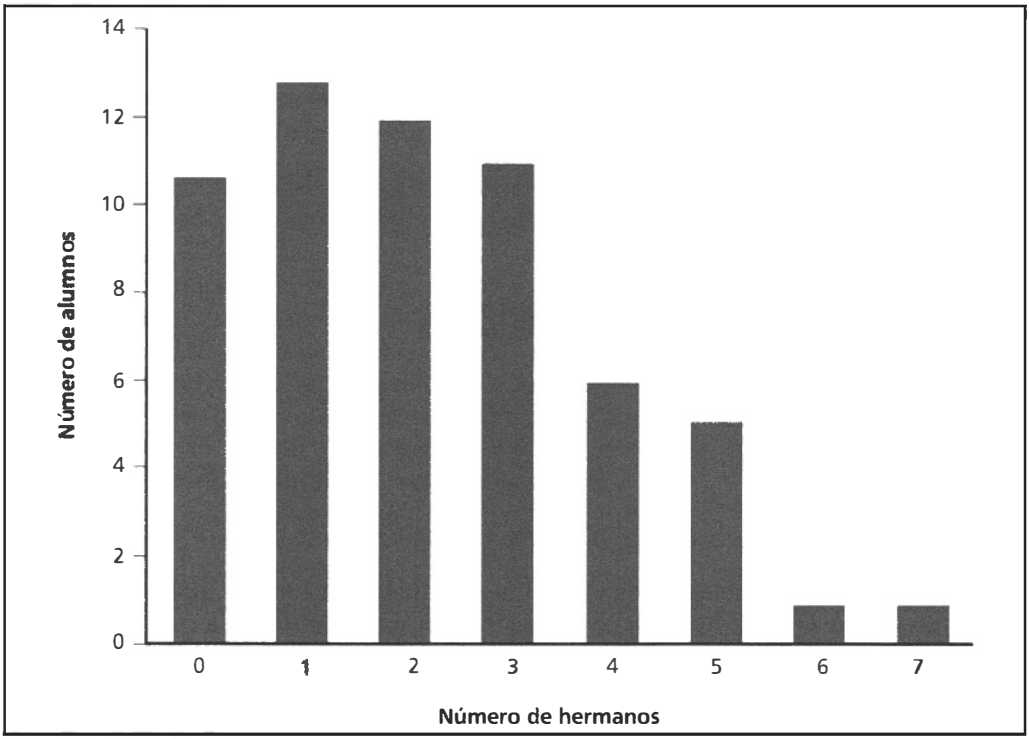


Gráfico 1.1. Diagrama de barras correspondiente a la distribución del número de hermanos

Gráfico de tallo y hojas *(stem and leaf)*[4]

El diagrama o gráfico de tallo y hojas es una técnica de organización de los datos que permite obtener simultáneamente la distribución de frecuencias y su representación gráfica.

Para construirlo se expresan los datos en la unidad necesaria para poder trabajar con dos o tres cifras significativas. Las cifras de cada observación se dividen en dos partes: la primera cifra (o dos primeras) es la parte más relevante de los valores observados y se denomina tallo; del tallo parte la hoja o última cifra significativa del valor observado.

Se dibuja una línea vertical y a su izquierda se anotan en columna los tallos desde el menor hasta el mayor. Los tallos deben ser consecutivos y abarcar todo el recorrido de la variable. A la derecha de la línea vertical se anotan las hojas correspondientes a cada tallo. Las hojas se ordenan de menor a mayor y deben ocupar el mismo espacio; habrá tantas hojas como número de observaciones.

[4] Este procedimiento de presentación y representación de los datos fue introducido por J. W. Tukey en 1977 en su libro *Exploratory Data Analysis.* Reading, MA: Addison-Wesley técnica.

Se puede completar el diagrama con las frecuencias simples o acumuladas, que se obtienen teniendo en cuenta la unidad de la hoja y el número de hojas correspondiente a cada tallo. Dichas frecuencias se suelen anotar a la izquierda de los tallos.

En el encabezado o pie del gráfico es importante indicar la unidad de la hoja (es decir, el número de observaciones que la misma representa) y del tallo (unidad de medida de la variable) para poder recuperar las observaciones en las unidades originales.

Ejemplo 1.2. Supongamos que las edades de un colectivo formado por 45 trabajadores son las siguientes: 32, 32, 32, 34, 34, 35, 35, 35 ,35, 37 ,37, 37 ,37, 38 ,39, 40, 40, 41, 42, 42, 42, 42, 42, 42, 42, 43, 43, 43, 43, 43, 45, 45, 45, 45, 45, 47, 47, 48, 49, 49, 50, 50, 51, 51, 51. El gráfico de tallo y hojas de esta muestra podría ser cualquiera de los dos que siguen (Gráficos 1.2 y 1.3).

n_i	Tallo	Hojas
15	3	222445555777789
25	4	00122222223333335555577899
5	5	00111

Unidad de las hojas: 1; 3|2 representa 32 años

Gráfico 1.2. Diagrama de tallo y hojas para la edad

Dada la poca variación que presentan los tallos es conveniente subdividirlos en partes iguales. Como las hojas toman los valores enteros de 0 a 9, únicamente se pueden subdividir los tallos en 2 o 5 partes para que sean todas ellas iguales. Si se subdividen en 2 partes, a la primera le corresponderán las hojas del 0 al 4 y a la segunda del 5 al 9.

n_i	Tallo	Hojas
5	3	22244
10	3	5555777789
15	4	001222222233333
10	4	5555577899
5	5	00111

Unidad de las hojas: 1; 3|2 representa 32 años

Gráfico 1.3. Diagrama de tallo y hojas para la edad

1.3.2. Distribuciones de frecuencias de una variable. Datos agrupados

Cuando disponemos de muchas observaciones y, además, muchas de ellas son distintas, se suele recurrir a la agrupación de las mismas en unos cuantos **intervalos** o **clases,** formando lo que se conoce como **distribución de frecuencias con datos agrupa-**

dos. Estas estadísticas se presentan en una tabla con dos filas o columnas. En la primera figuran los distintos intervalos, frecuentemente ordenados en orden creciente, y en la segunda el número de veces que han sido observados valores dentro del intervalo, es decir, su frecuencia absoluta de observación.

Ejemplo 1.3. Los establecimientos hoteleros de una gran ciudad se han agrupado por el número de plazas, obteniéndose como resultado la distribución de frecuencias que se presenta en la Tabla 1.1.

Tabla 1.1. Número de plazas

Plazas	N.º de hoteles
0-100	25
100-200	37
200-300	12
300-500	22
500-600	21
600-800	18
800-1.000	5
	140

Estas estadísticas difieren de las anteriores en que no es posible reconstruir los datos originales a partir de la información contenida en la tabla. El agrupamiento en clases conlleva la pérdida de parte de la información original, sabemos cuántas observaciones hay en cada clase, pero no cuáles son. En consecuencia, al calcular determinadas medidas numéricas (que, en general, son función de las observaciones) a partir de la tabla, cometeremos errores. Esos **errores** se llaman **de agrupamiento.**

En la obtención de una distribución de datos agrupados, a partir de los datos originales, nos encontraremos con los siguientes problemas:

- Fijar el número de clases que deben tomarse. Este debe ser el suficiente para que no se pierda excesiva información primaria y para que la estadística resulte manejable y útil para expresar las características de la variable. En este sentido, algunas de las reglas que se utilizan en la práctica son: tomar un número de clases (**k**) igual al entero más próximo a la raíz cuadrada del número total de datos ($\sqrt{\mathbf{N}}$) o al entero más próximo a la expresión $\mathbf{1+3{,}3\cdot\log N}$ (fórmula de Sturges[5]). En general, se suele recomendar un número de clases entre cinco y quince[6].
- La manera de expresar en la tabla los límites de los intervalos, de modo que no originen error. Utilizaremos intervalos solapados, es decir el extremo superior de cada intervalo coincide con el inferior del inmediato posterior, semiabiertos por la derecha y de manera tal que cubran todas las observaciones.
- La utilización de intervalos de amplitud fija o variable. En lo posible, se recomienda trabajar con intervalos de amplitud constante, ya que es más fácil el tratamiento analítico de la estadística.

[5] Sturges, M.A. (1926): «The choice of a class-interval», *Journal of the American Statistical Association*, 21.

[6] Véase, por ejemplo, el libro de Alfonso García Barbancho *Estadística elemental moderna* de la editorial Ariel.

Para las distribuciones de frecuencias con datos agrupados utilizaremos la siguiente notación:

- **N** es el número total de observaciones de la variable **X.**
- L_{i-1} - L_i, **i = 1,..., k,** son los **k** intervalos o **clases** en los que se encuentran agrupadas las observaciones. Por supuesto, **k≤N.**
- n_i, **i = 1,..., k,** es el número de observaciones en el i-ésimo intervalo o, lo que es lo mismo, la **frecuencia absoluta** de observación correspondiente al intervalo L_{i-1} - L_i.
- a_i, **i = 1,..., k,** es la **amplitud del intervalo** L_{i-1} - L_i, es decir,

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

- Si los intervalos son de distinta amplitud, pueden ser necesarias las alturas correspondientes a los mismos, que se representan con h_i, **i = 1,..., k,** y se definen como:

$$h_i = \frac{n_i}{a_i}$$

- x_i, **i = 1,..., k,** es la **marca de clase** o representante de las observaciones en el intervalo L_{i-1}-L_i, y se define como:

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

es decir, la marca de clase de cada intervalo es el punto medio del mismo.

En la siguiente tabla se muestran los valores de a_i, x_i y h_i que resultan para los intervalos del Ejemplo 1.3.

Tabla 1.2. Número de plazas (cálculos intermedios)

Plazas (L_{i-1}-L_i)	N.° de hoteles (n_i)	x_i	a_i	h_i
0 - 100	25	50	100	0,25
100 - 200	37	150	100	0,37
200 - 300	12	250	100	0,12
300 - 500	22	400	200	0,11
500 - 600	21	550	100	0,21
600 - 800	18	700	200	0,09
800 - 1.000	5	900	200	0,025
	140			

Histograma

La representación gráfica de una distribución de frecuencias de datos agrupados recibe el nombre de **histograma.** El histograma se obtiene pintando en el eje de abscisas los extremos de los intervalos y levantando sobre ellos rectángulos de área igual o proporcional a su frecuencia (véase el Gráfico 1.4).

Si todos los intervalos son de la misma amplitud, podemos pintar rectángulos de base el intervalo y de altura la frecuencia. De este modo, obtendremos rectángulos de área proporcional a la frecuencia y la constante de proporcionalidad será esa amplitud **(a)** que es común a todos los intervalos.

Si los intervalos son de distinta amplitud, se pintan rectángulos de área igual a la frecuencia, utilizando para cada intervalo una altura igual a su frecuencia dividida por su amplitud (lo que antes hemos definido como h_i).

El histograma para la distribución del número de plazas del Ejemplo 1.3, que se presenta en el Gráfico 1.4, está basado en las alturas, ya que los intervalos son de amplitud variable.

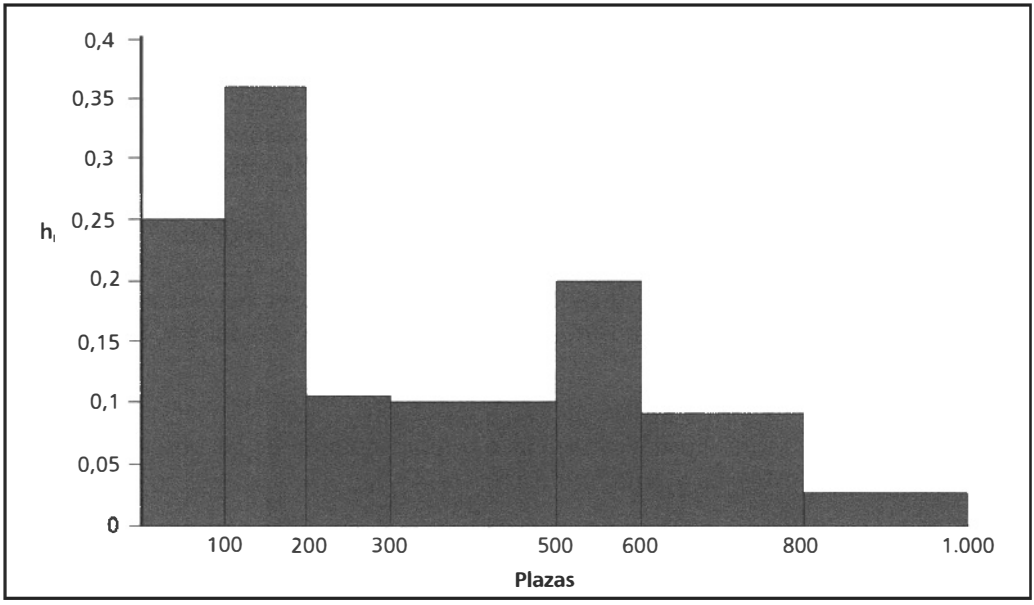


Gráfico 1.4. Histograma del número de plazas

Polígono de frecuencias

Si unimos con trazos rectos los puntos medios de los lados superiores de rectángulos contiguos, obtenemos el polígono de frecuencias. En el Gráfico 1.5 se puede observar el polígono de frecuencias correspondiente al Ejemplo 1.3.

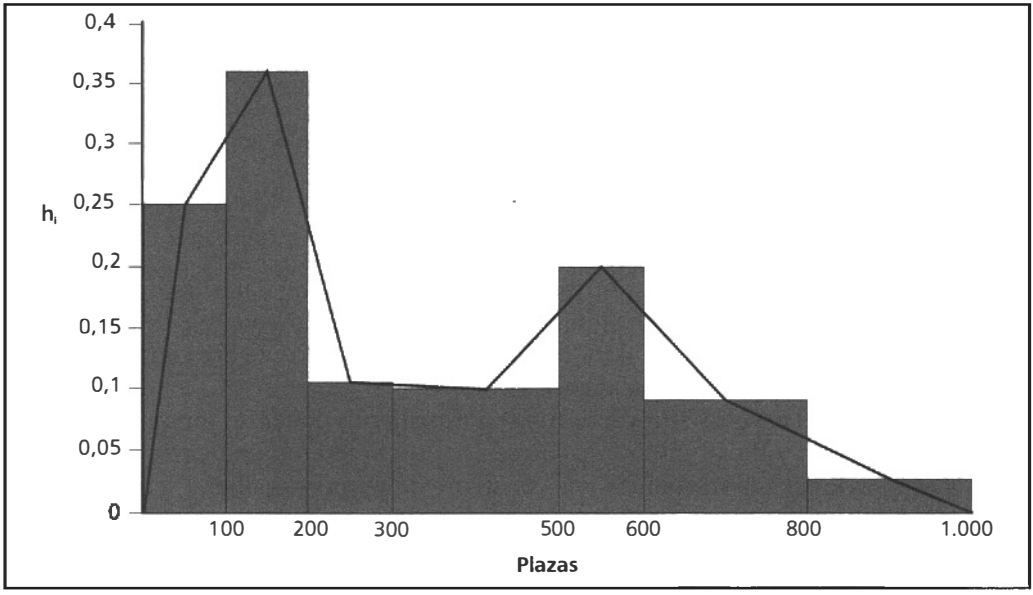


Gráfico 1.5. Polígono de frecuencias del número de plazas

1.3.3. Estadísticas derivadas: distribuciones de frecuencias acumuladas, frecuencias relativas y porcentajes

En las distribuciones de una variable, nos puede interesar conocer el número de observaciones correspondiente a un valor o intervalo y a todos los anteriores a él. Cuando las frecuencias tienen este significado se denominan **frecuencias acumuladas** o **frecuencias absolutas acumuladas.** Se representan por $\mathbf{N_i}$, i = 1,..., k, y se definen como:

$$\mathbf{N_i = \sum_{j=1}^{i} n_j}$$

Es claro que $\mathbf{N_1 = n_1}$ y que $\mathbf{N_k = N}$, si disponemos de **k valores o intervalos distintos en un conjunto de N observaciones.**

Ejemplo 1.4. En la siguiente tabla se muestran las distribuciones de frecuencias absolutas, simples y acumuladas, correspondientes a una cierta variable X.

x_i	n_i	N_i
3	1	1
4	3	4
6	5	9
7	10	19
10	15	34
	34	

En una suma de varios sumandos se llama **proporción** al cociente de dividir un sumando cualquiera por el total. A las proporciones que resultan de dividir cada frecuencia (n_i) por el número total de observaciones (la suma de todas las frecuencias, N) se les llama **frecuencias relativas** y se les representa mediante $\mathbf{f_i}$.

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

Para un conjunto de N observaciones con k valores o intervalos distintos, se cumple que:

$$\mathbf{\sum_{i=1}^{k} f_i = 1}$$

Llamamos **frecuencia relativa acumulada** del i-ésimo valor o intervalo, y la representamos mediante $\mathbf{F_i}$, al resultado de acumular hasta el mismo las correspondientes frecuencias relativas, es decir,

$$\mathbf{F_i = \sum_{j=1}^{i} f_j}$$

Es claro que,

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

y que, si los N datos están agrupados en k intervalos o valores distintos,

$$\mathbf{F_k = 1}$$

Si las frecuencias relativas se multiplican por 100 obtenemos los **porcentajes.** Para el i-ésimo valor o intervalo,

$$p_i = \frac{n_i}{N} \cdot 100 \quad \text{y} \quad \sum_{i=1}^{k} p_i = 100$$

Cuando acumulamos los porcentajes, obtenemos los **porcentajes acumulados:**

$$P_i = \frac{N_i}{N} \cdot 100 \quad \text{y} \quad P_k = 100$$

A las distribuciones de frecuencias acumuladas, frecuencias relativas o porcentajes que se obtienen a partir de la estadística primaria o básica mediante operaciones aritméticas se les llama **estadísticas derivadas.**

Siguiendo con el Ejemplo 1.3, en la Tabla 1.3 se presentan todas las frecuencias simples (absolutas, relativas y porcentuales) y las frecuencias acumuladas (absolutas, relativas y porcentuales).

Tabla 1.3. Número de plazas. Estadísticas derivadas

Plazas (L_{i-1}-L_i)	N.° de hoteles (n_i)	N_i	f_i	F_i	p_i	P_i
0 - 100	25	25	0,1786	0,1786	17,86	17,86
100 - 200	37	62	0,2643	0,4429	26,43	44,29
200 - 300	12	74	0,0857	0,5286	8,57	52,86
300 - 500	22	96	0,1571	0,6857	15,71	68,57
500 - 600	21	117	0,1500	0,8357	15,00	83,57
600 - 800	18	135	0,1286	0,9643	12,86	96,43
800 - 1000	5	140	0,0357	1,0000	3,57	100,00
	140					

1.3.4. Más representaciones gráficas

La representación gráfica de la distribución de frecuencias relativas o porcentuales con datos agrupados o sin agrupar es similar a la de frecuencias absolutas, cambiando la escala en el eje de ordenadas para trabajar con frecuencias relativas o porcentuales en lugar de con absolutas.

Diagrama de frecuencias acumuladas o escalonado

En las distribuciones de datos no agrupados, la representación gráfica de la distribución de frecuencias acumuladas (absolutas, relativas o porcentuales) es el **diagrama de frecuencias acumuladas o diagrama de escalera.** En el eje de abscisas se marcan los distintos valores, sobre cada valor se levanta una barra discontinua paralela al eje de ordenadas y de altura igual a la frecuencia acumulada (absoluta o relativa) correspondiente a dicho valor y, por último, se une, mediante trazo horizontal, cada barra con la inmediata posterior, teniendo en cuenta que antes del primer valor el trazo horizontal coincide con el eje de abscisas y después del último el trazo horizontal es de altura N, 1 o 100, según que estemos trabajando con frecuencias absolutas, relativas o porcentuales, respectivamente.

Ejemplo 1.5. El diagrama de escalera correspondiente a la distribución de frecuencias acumuladas dada en la siguiente tabla es el que se muestra a continuación.

x_i	1	2	3	5	6	
n_i	2	1	4	3	4	**14**
N_i	2	3	7	10	14	

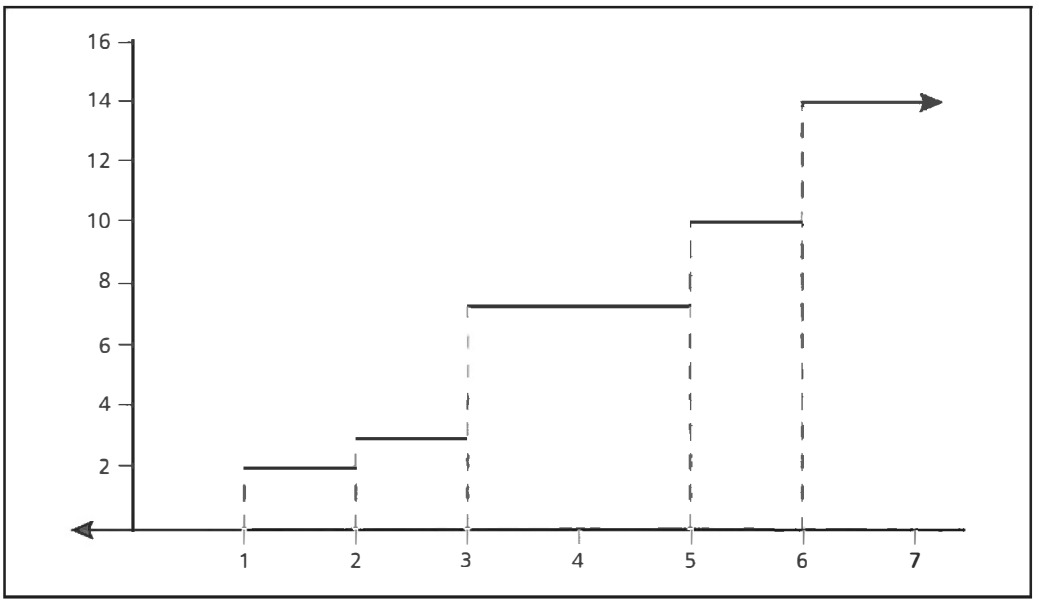


Gráfico 1.6. Diagrama de escalera para el Ejemplo 1.5

Ojiva o polígono de frecuencias acumuladas

En las distribuciones de datos agrupados, la representación gráfica de la distribución de frecuencias acumuladas (absolutas, relativas o porcentuales) es el **polígono de frecuencias acumuladas u ojiva.**

En el eje de abscisas se marcan el extremo inferior del primer intervalo y los extremos superiores de todos los intervalos. En el eje de ordenadas las frecuencias acumuladas (absolutas o relativas). Pintamos en los ejes los puntos que tienen abscisa en el extremo superior de cada intervalo y ordenada la frecuencia acumulada correspondiente a la clase que dicho extremo representa. Pintamos, también, el punto que tiene como abscisa el extremo inferior del primer intervalo y como ordenada cero. Los puntos así obtenidos se unen entre sí (cada uno con su inmediato a la derecha) mediante trazos rectos. Por último, hemos de tener en cuenta que antes del punto correspondiente al extremo inferior del primer intervalo la gráfica coincide con el eje de abscisas y que después del punto que corresponde al extremo superior del último intervalo la gráfica coincide con la recta $y = N$, $y = 1$ o $y = 100$, según que estemos trabajando con frecuencias absolutas, relativas o porcentuales.

Ejemplo 1.6. El polígono de frecuencias acumuladas que corresponde a la distribución de datos agrupados de la variable que figura en la siguiente tabla es el que se muestra en el Gráfico 1.7.

$L_{i-1} - L_i$	n_i	N_i
4-16	55	55
16-20	47	102
20-24	32	134
24-36	26	160
	160	

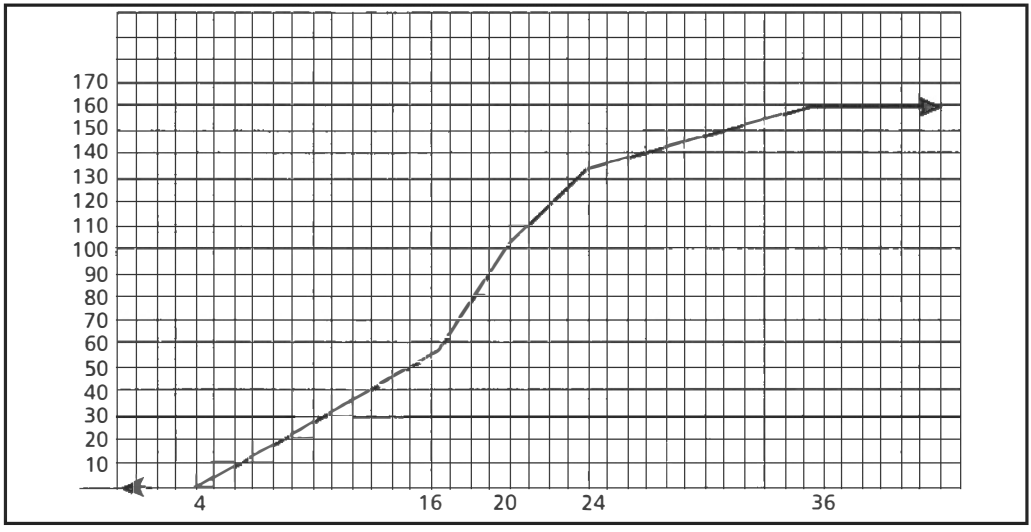


Gráfico 1.7. Polígono de frecuencias acumuladas para el Ejemplo 1.6

Observación. Al representar la distribución de frecuencias mediante un polígono como el anterior estamos aceptando que dentro de cada intervalo la distribución de las observaciones es uniforme o, lo que es lo mismo, el número de observaciones que corresponde a cada subintervalo de una clase dada es proporcional a su amplitud e independiente de su posición dentro de la misma.

Ejemplo 1.7. Los extranjeros que viven en un país de la Unión Europea quedan clasificados en la Tabla 1.4 según su país de origen. Para representar la distribución mediante un diagrama de sectores repartimos el ángulo completo (360°) entre el total de observaciones (140). La constante de proporcionalidad será entonces 360/140 (esa es la amplitud del ángulo que corresponde a una observación cualquiera, 2,571).

Tabla 1.4. Extranjeros según país de origen

	Extranjeros (10^4 personas)
Países de la U.E.	50
Otros países europeos	45
Países del norte de África	40
Otros países	5

Teniendo en cuenta el valor de la constante de proporcionalidad, ya calculado antes, las amplitudes que proporcionalmente les corresponden a las distintas modalidades son las que se presentan en la Tabla 1.5.

Tabla 1.5. Extranjeros según país de origen. Cálculo de ángulos en grados

	Extranjeros (10^4 personas)	Ángulo en grados
Países de la U.E.	50	50·2,571=128,55
Otros países europeos	45	45·2,571=115,695
Países del norte de África	40	40·2,571=102,84
Otros países	5	5·2,571=12,855
	140	359,94 (360)

El gráfico de sectores es el que se presenta a continuación.

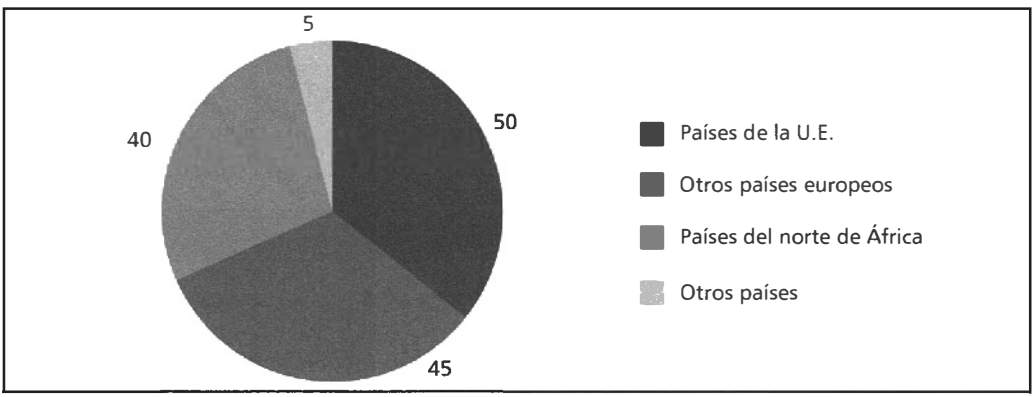


Gráfico 1.8. Extranjeros según país de origen (104 personas). Gráfico de sectores

Barra de componentes porcentuales

Es una barra vertical con escala en la que se van apilando los porcentajes correspondientes a cada una de las modalidades observadas. En el Gráfico 1.9 se ha representado la barra de componentes porcentuales correspondiente al ejemplo anterior (Ejemplo 1.7).

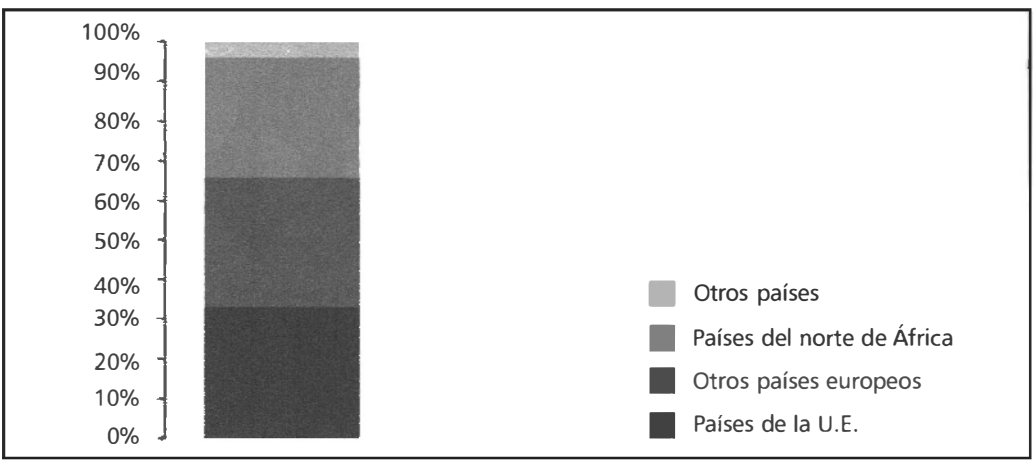
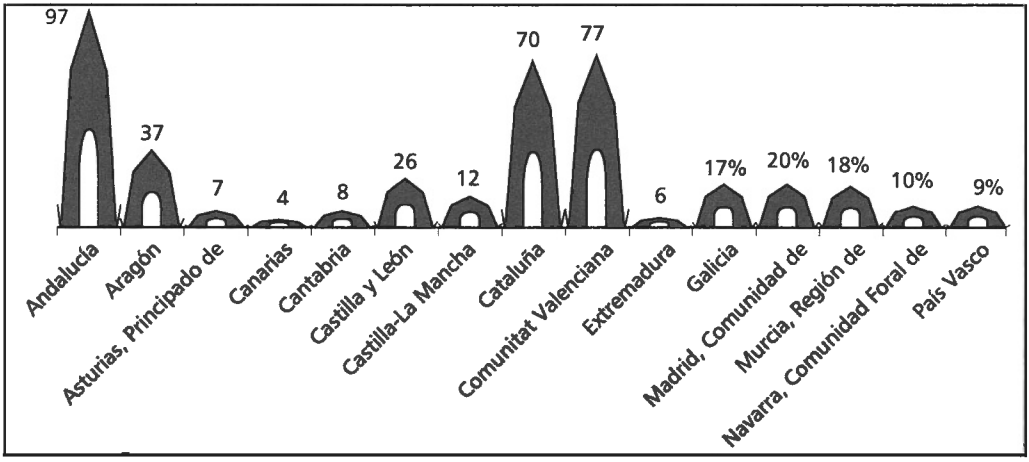


Gráfico 1.9. Barra de componentes. Extranjeros según país de origen

Pictogramas

Dibujos que representan al atributo y cuyo tamaño es, para cada modalidad, proporcional a la frecuencia con la que la misma se presenta. Por ejemplo, en el Gráfico 1.10 se muestra un pictograma relativo a la distribución por Comunidades Autónomas de los acampamentos turísticos en 2012.

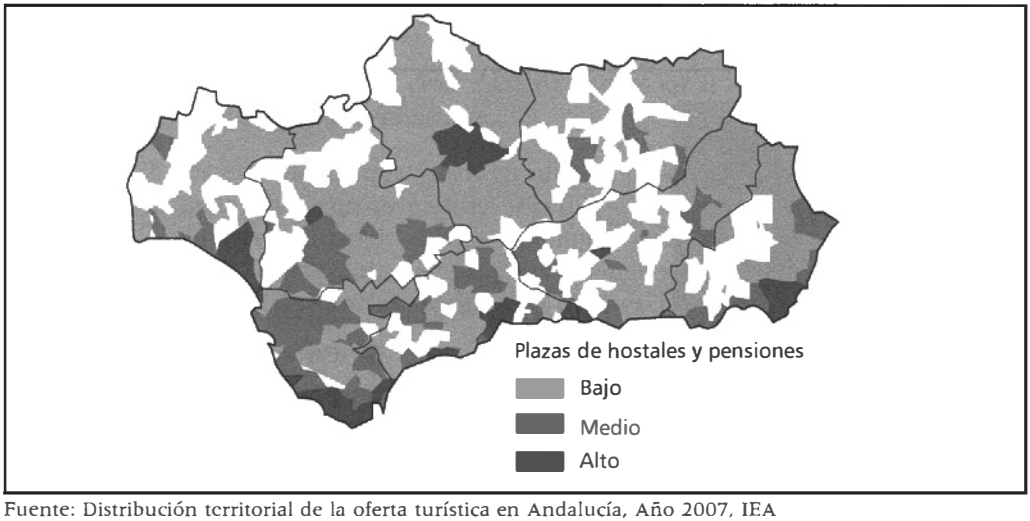


Fuente: Encuesta de Ocupación en Acampamentos Turísticos, INE

Gráfico 1.10. Número de acampamentos turísticos

Cartogramas o mapas

Se utilizan cuando las modalidades son áreas geográficas o cuando interesa visualizar una cierta variable según la zona geográfica en la que haya sido observada (estadísticas geográficas). Las distintas frecuencias (o los distintos valores o intervalos) se representan con tramas (o figuras geométricas) tanto más densas (más grandes) cuanto mayor es el valor al que representan. A modo de ejemplo, en el Gráfico 1.11 se muestra la distribución municipal del número de plazas en hostales y pensiones en los municipios de Andalucía en 2007.



Fuente: Distribución territorial de la oferta turística en Andalucía, Año 2007, IEA

Gráfico 1.11. Distribución municipal de las plazas en hostales y pensiones

1.4. Fuentes estadísticas oficiales

Las fuentes de información estadística pueden ser clasificadas en función de diversos criterios. Atendiendo al ámbito geográfico, distinguiremos entre:

- Organismos internacionales.
- Organismos nacionales.
- Organismos regionales.

Organismos internacionales

El principal organismo oficial productor de estadísticas a nivel europeo es la Oficina de Estadística de la Unión Europea (EUROSTAT, **http://epp.eurostat.ec.europa.eu/).**

A modo de ejemplo, en la Figura 1.2[7] se puede observar que, entre los distintos temas sobre los que podemos encontrar información estadística, Eurostat recoge y recopila una amplia gama de datos económicos, financieros y del turismo, con secciones dedicadas a las Cuentas nacionales (incluido el PIB); tablas *Input-Output;* tipos de cambio; tasas de interés; etc.

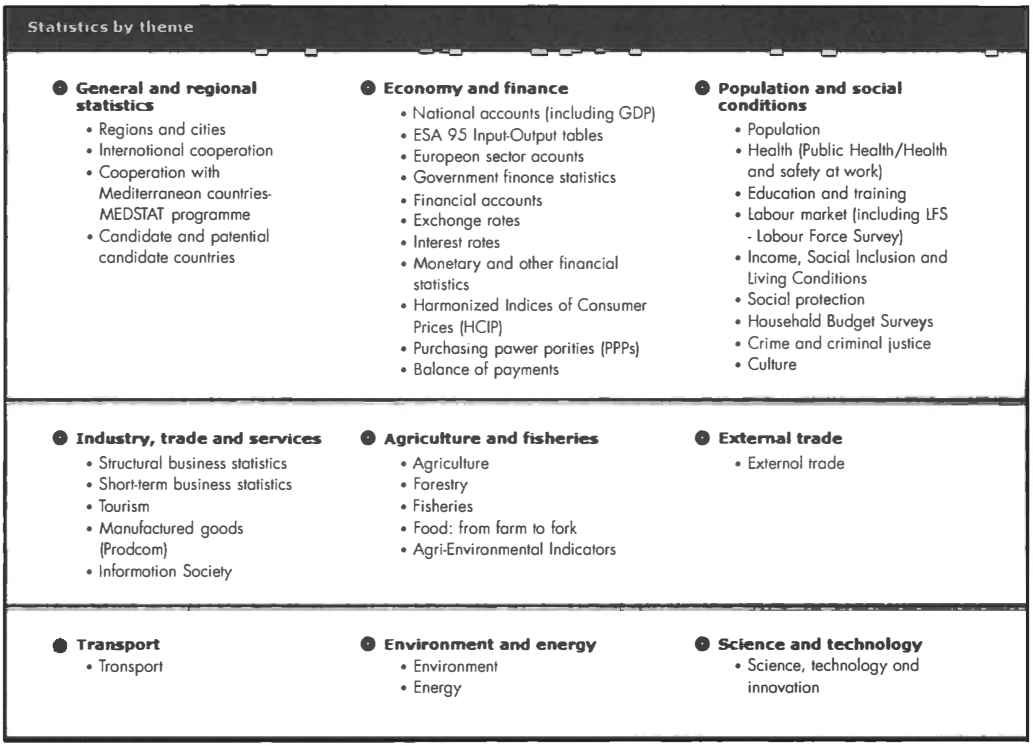


Figura 1.2. Estadísticas por temas. Página web de Eurostat

[7] Información suministrada en la web de Eurostat.

Organismos nacionales

El organismo de carácter nacional que produce un mayor número de estadísticas, tanto relativas a la Economía y las Finanzas como a cualquier otra área temática, es el Instituto Nacional de Estadística (INE; **http://www.ine.es/**). Entre las estadísticas de interés en el ámbito económico-financiero que ofrece este organismo están las que se presentan en la Figura 1.3[8].

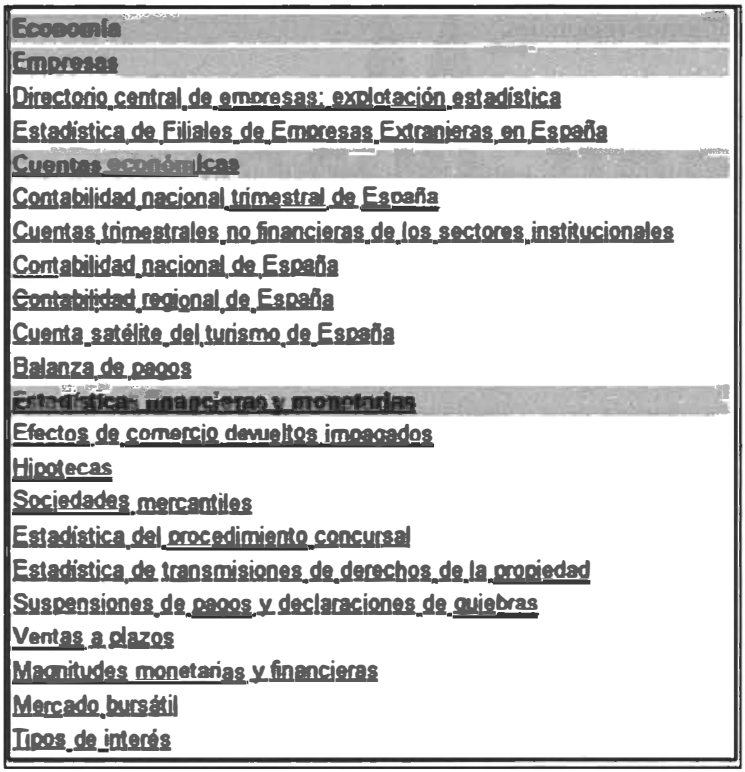


Figura 1.3. Estadísticas de Empresas, Cuentas Económicas y Estadísticas Financieras y Monetarias. Página web del INE

Otro organismo nacional con importante producción estadística en el ámbito financiero es el Banco de España (BDE, **http://www.bde.es/webbde/es/**). Como se indica en su página web oficial, el Banco de España elabora y publica las estadísticas relacionadas con sus funciones y asiste al Banco Central Europeo con las estadísticas necesarias para llevar a cabo las políticas monetaria y cambiaria únicas, referidas, entre otros aspectos, a magnitudes monetarias, bancarias y financieras, tipos de interés y balanza de pagos. En la Figura 1.4 mostramos las estadísticas que este organismo ofrece agrupadas por publicaciones.

Por último, en cuanto a los organismos de ámbito geográfico regional destacamos los Institutos de Estadística Regionales, que existen solo en algunas comunidades autónomas. En la siguiente tabla figura la denominación de las Oficinas estadísticas para las diferentes comunidades autónomas[9].

[8] Información suministrada en la web del INE.
[9] Información suministrada en la Web del INE.

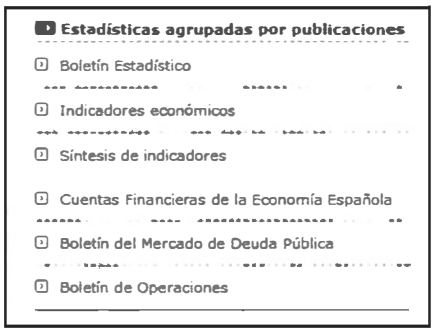


Figura 1.4. Estadísticas agrupadas por publicaciones. Página web del BDE

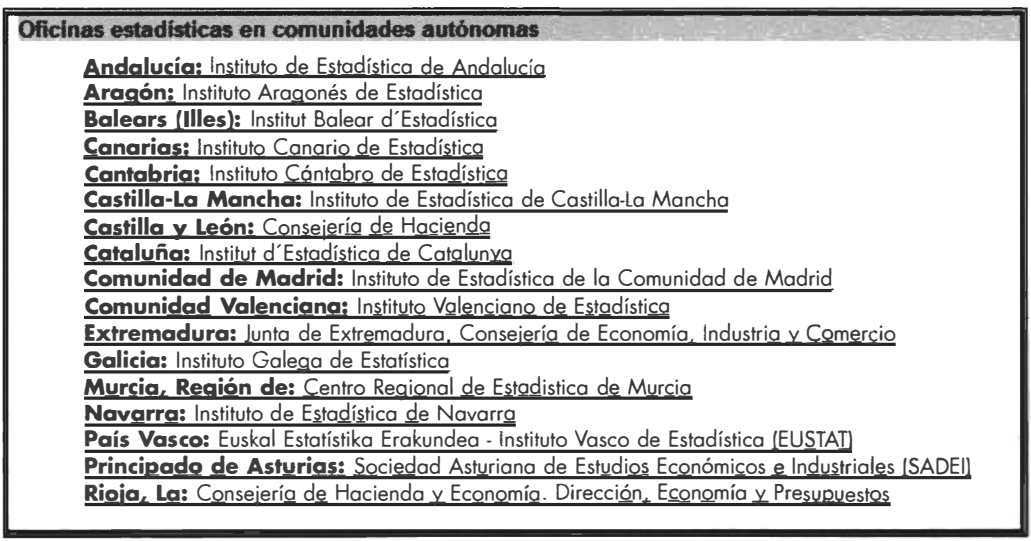


Figura 1.5. Oficinas estadísticas en comunidades autónomas. Página web del INE

1.5. Ejercicios

Ejercicio 1.1. Indique, en cada uno de los siguientes casos, si la variable mencionada es: cuantitativa, cualitativa, discreta, continua, nominal, ordinal, de flujo o *stock*.

a) Número de agencias de viajes en Andalucía. Datos anuales desde 1995 hasta 2005.
b) Número de cancelaciones de estancias en un hotel en los 5 últimos años.
c) Nivel de estudios de un conjunto de trabajadores.
d) Distribución de los ocupados atendiendo al nivel de estudios.
e) Número de ocupados en el sector industrial en Málaga en los años 2003, 2004 y 2005.
f) Salario medio mensual de los ocupados en el sector industrial en Málaga en los años 2003, 2004 y 2005.

g) Salario mensual de los trabajadores de una empresa.
h) Distribución de los visitantes según tipologías (turistas y excursionistas) en el año 2009.
i) Número de convenios en empresas turísticas (por ejemplo, de empresa, de grupo de empresas, de sector, etc.) en los últimos 10 años.

Ejercicio 1.2. En febrero de 2006 se realizó una encuesta a 50 empleados de una gran empresa. Entre otras cuestiones, se les preguntó sobre el número de horas extraordinarias que trabajan a la semana. Los resultados fueron los siguientes:

2 1 2 2 1 2 4 2 1 1
2 3 2 1 1 1 3 4 2 2
2 2 1 2 1 1 2 3 2 2
3 2 3 2 2 4 2 3 4 3
3 3 4 3 3 3 3 3 3 2

Construya una tabla estadística que resuma la información contenida en el cuadro anterior y responda, basándose en ella, a las siguientes cuestiones:

a) ¿Cuál es la población?
  a1) Número de empleados.
  a2) Número de horas extraordinarias.
  a3) El total de horas extraordinarias.
  a4) El total de empleados.
b) Atendiendo al número de elementos, ¿de qué tipo de observación diría que se trata?
  b1) Exhaustiva.
  b2) Parcial.
  b3) Muestra.
  b4) Subpoblación.
c) ¿Y atendiendo a la referencia temporal?
  c1) Transversal.
  c2) Longitudinal o temporal.
d) ¿Cuál es el número total de observaciones?
e) ¿Cuántos empleados encuestados han trabajado 3 horas extraordinarias?
f) ¿Cuántos de los empleados encuestados han trabajado 2 o menos horas extraordinarias?
g) Entre los empleados encuestados, ¿cuántos trabajaron 1 hora extraordinaria?
h) ¿Qué número de los empleados encuestados trabajaron 2 o más horas extraordinarias?
i) Entre los empleados encuestados, ¿cuántos han trabajado 3 o más horas extraordinarias?
j) Entre los empleados encuestados, ¿cuántos han trabajado 2 o 3 horas extraordinarias?

Ejercicio 1.3. El polígono de frecuencias relativas acumuladas de una variable estadística X, cuyos datos observados están agrupados en intervalos o clases es el que se presenta en el Gráfico 1.12.

Basándose en él y teniendo en cuenta que el número total de valores observados es 50, obtenga:

a) Frecuencia relativa del segundo intervalo.
b) Límite superior del segundo intervalo.
c) Marca de clase del último intervalo.
d) Porcentaje acumulado del tercer intervalo.
e) ¿Cuántas observaciones tuvieron un valor inferior a 10?
f) Porcentaje de observaciones con un valor inferior a 8.
g) ¿Cuántas observaciones tuvieron un valor superior o igual a 11? ¿Qué porcentaje del total representan?

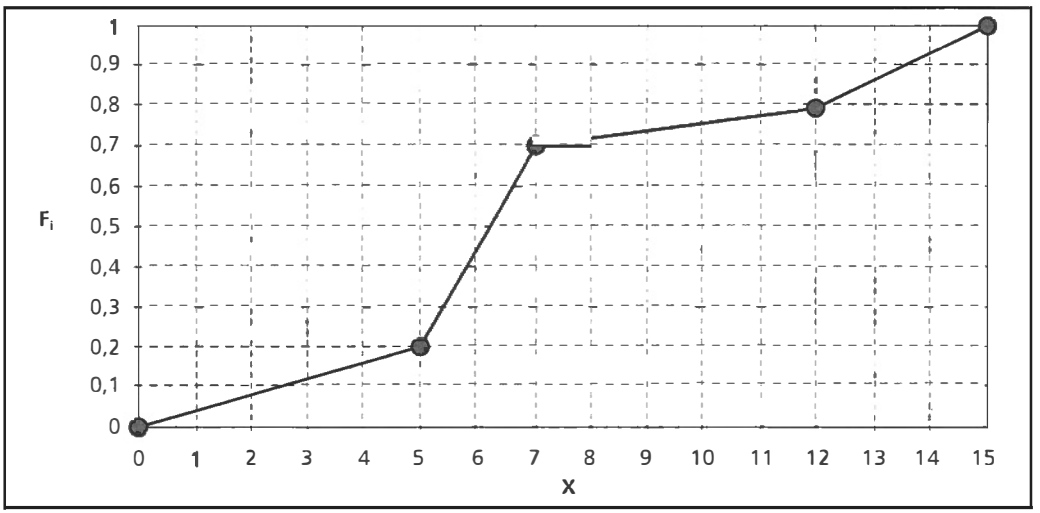


Gráfico 1.12. Polígono de frecuencias acumuladas para el Ejercicio 1.3

Ejercicio 1.4. Basándose en la información contenida en la siguiente tabla estadística, relativa al número de trabajadores en una muestra de gestorías observadas en un año determinado, responda a las cuestiones que se le plantean.

N.° de trabajadores	N.° de gestorías
1	1
2	5
3	12
4	20
5	23
6	23
7	12
8	2
9	2
Total	**100**

a) ¿Cuál es la variable a la que refiere la estadística anterior? ¿De qué tipo es? La observación, ¿es exhaustiva o parcial? ¿Y los datos?, ¿son temporales o transversales?

b) Represente gráficamente la distribución de frecuencias absolutas.

c) Determine:

c1) Número de gestorías encuestadas.

c2) Porcentaje de gestorías con 5 trabajadores.

c3) Número de gestorías con un máximo de 3 trabajadores.

c4) Número máximo y mínimo de trabajadores.

c5) Número máximo de trabajadores que tienen las 30 gestorías con menos personal.

c6) Si una empresa de software únicamente está interesada en enviar propaganda a las gestorías con más de 6 empleados, ¿a qué porcentaje de las gestorías muestreadas se dirigirá?

c7) Si la empresa de software está interesada en enviar propaganda al 25% de las gestorías con mayor empleo, ¿cuál es el número mínimo de empleados que debe tener una gestoría para estar incluida en este grupo?

c8) Si la Oficina Regional de Empleo se propone ayudar al 25% de las gestorías con menor empleo en el departamento de personal enviando un trabajador en prácticas, ¿cuántos empleados como máximo deberán tener para poderse beneficiar de dicha ayuda?

Ejercicio 1.5. Indique, en cada uno de los siguientes casos, cuáles hacen referencia a variables cualitativas, y cuáles a variables cuantitativas y, en este último supuesto, cuáles son discretas y cuáles son continuas:

a) Altura de una persona.

b) Temperatura de un mostrador frigorífico.

c) Número de anuncios emitidos en un intermedio publicitario por cierta cadena de televisión.

d) Tiempo necesario para la fabricación de una pieza.

e) Peso neto de una botella estándar de aceite.

f) Pulsaciones por minuto de una mecanógrafa.

g) Categoría de un hotel.

h) Opinión sobre su bienestar en el trabajo. Carácter de la población, al que se le asignan los valores 0, 1, 2, 3 o 4, según que la opinión del individuo sea mala, regular, normal, buena o excelente.

i) Tamaño de una empresa medido por el número de trabajadores.

j) Edad (en años) de los empleados del departamento de personal.

k) Número de accidentes mortales en una empresa.

Ejercicio 1.6. Complete los datos que faltan en la tabla y el gráfico que se presentan a continuación e indique cuál es la población, quiénes son sus elementos, cuál es la variable analizada y de qué tipo es.

Empleados	n_i	N_i
Menos de 2		
2 a 5		
5 a 20	5.914	16.602
20 a 100		
100 o más		19.027

Fuente: Elaboración propia

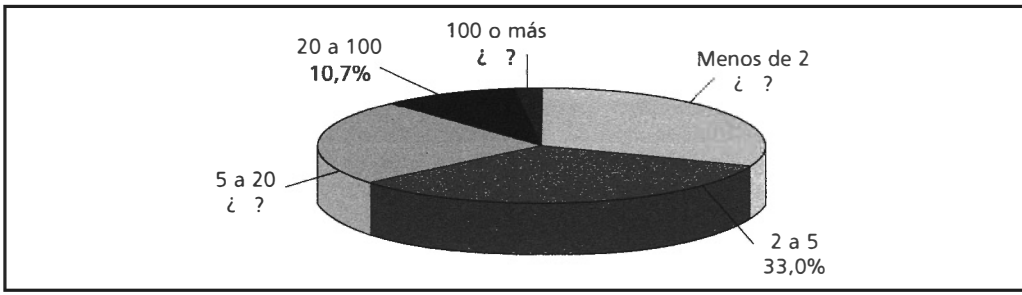


Gráfico 1.13. Distribución de las empresas atendiendo al n.º de empleados

Ejercicio 1.7. Utilizando los datos de la Encuesta Anual de Servicios del Instituto Nacional de Estadística correspondientes a un determinado año, se han construido los Gráficos 1.14 y 1.15 relativos al número de empresas en diferentes sectores. Complete los datos que faltan en ambos gráficos e indique cuál es la población, quiénes son sus elementos, cuál es la variable analizada y de qué tipo es.

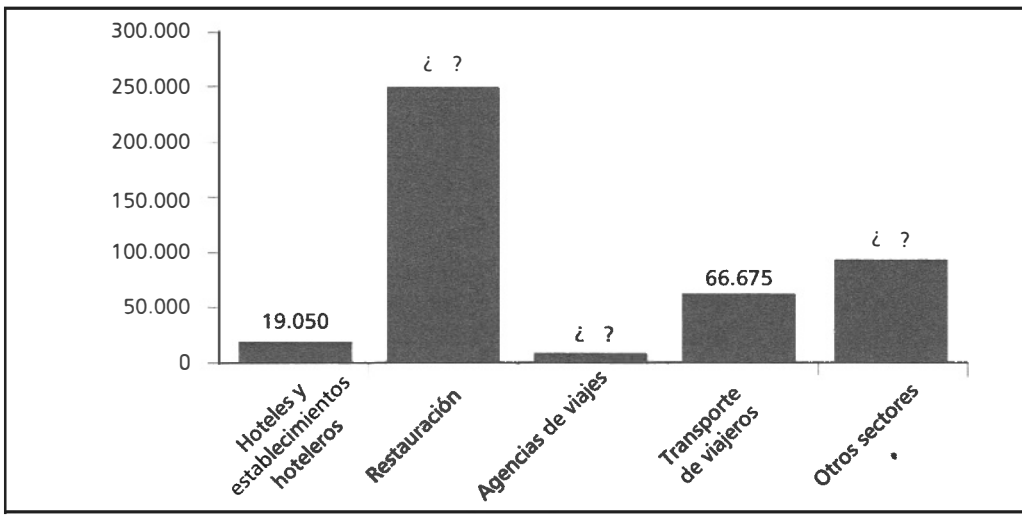


Gráfico 1.14. Número de empresas por sectores

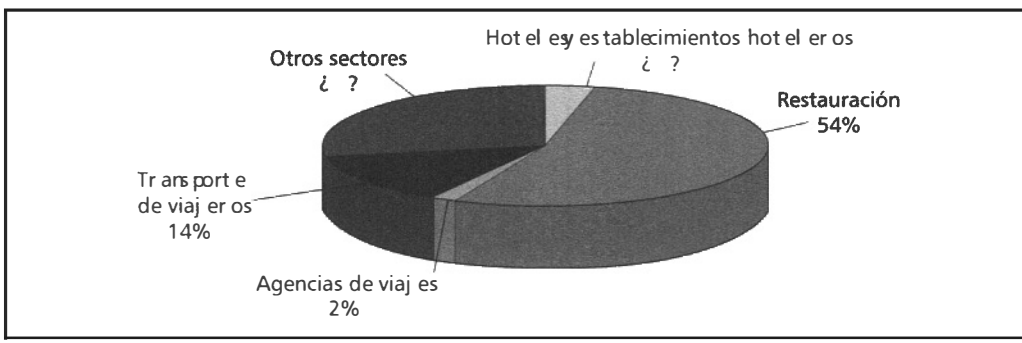


Gráfico 1.15. Porcentaje de empresas por sectores

Ejercicio 1.8. En la siguiente tabla figuran los salarios, en euros, de los 900 trabajadores de una determinada empresa.

Salario (€)	N.° de trabajadores
400-600	50
600-800	200
800-900	300
900-1.000	100
1.000-1.600	250

a) Determine la población, los elementos, la característica de interés y el tipo de la misma.
b) Represente gráficamente la distribución de frecuencias absolutas.
c) Obtenga la distribución de frecuencias absolutas acumuladas y represéntela gráficamente.
d) Determine el porcentaje de trabajadores con salario superior a 820 euros.

Ejercicio 1.9. El siguiente diagrama de tallo y hojas (*stem and leaf*) (Gráfico 1.16) corresponde a la observación de la variable X=«Peso en gramos» de un cierto número de paquetes gestionados por un servicio de mensajería.

```
1 1 2: represents 12
leaf unit: 1
            n: 100
   2     110 | 58
         111 |
   4     112 | 59
   6     113 | 68
  12     114 | 24679
  16     115 | 0234
  20     116 | 8999
  28     117 | 01235699
  40     118 | 00122356689
  49     119 | 011446688
 (10)    120 | 2233556699
  41     121 | 11346667999
  30     122 | 1233456789
  20     123 | 035899
  14     124 | 01789
   9     125 | 6678
   5     126 | 12
         127 |
   3     128 | 27
   1     129 | 2
```

Gráfico 1.16. Diagrama de tallo y hojas

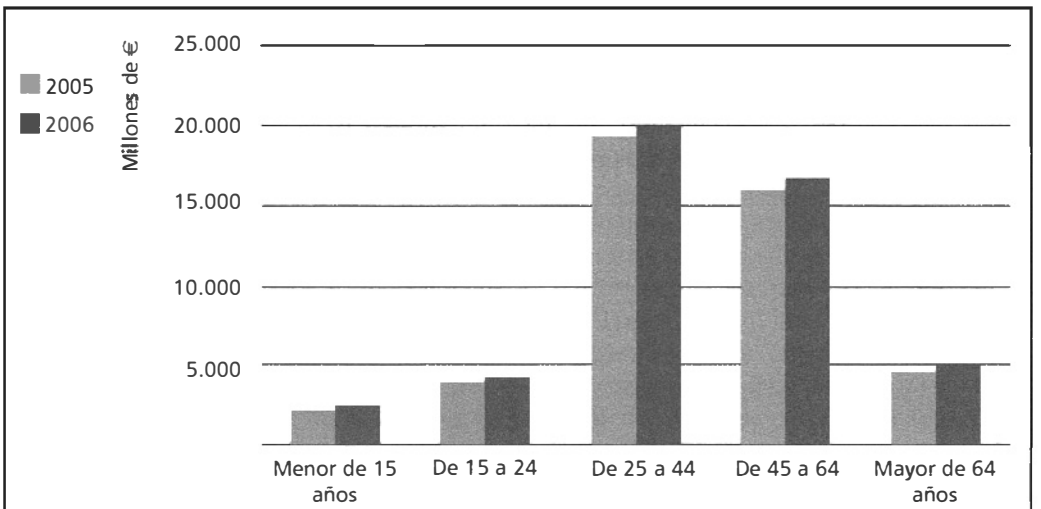
Basándose en la información contenida en el gráfico, resuelva las siguientes cuestiones:

a) Peso mínimo que denominaremos $X_{MIN.}$
b) Peso máximo que denominaremos X_{MAX}.
c) Número de observaciones que denominamos N.
d) Número total de paquetes que pesan exactamente 1.256 g.
e) Número de paquetes que pesan como máximo 1.149 g.
f) Número de paquetes en la muestra que pesan más de 1.250 g.
g) Proporción de paquetes en la muestra que pesan como máximo 1.150 g.
h) Proporción de paquetes en la muestra que pesan como mínimo 1.216 g.

Ejercicio 1.10. Según la Encuesta de Estructura Salarial para el año 2002 elaborada por el INE, la ganancia media anual por trabajador en el sector de la hostelería fue de 13.174,63 euros. En dicha encuesta se informa también de que el 25% de los trabajadores de dicho sector tuvieron una ganancia igual o inferior a 9.525,75 euros; el 50% de los trabajadores tuvieron una ganancia anual inferior a 12.369,36 euros y solo un 10% de los trabajadores de dicho sector tuvo una ganancia anual superior 20.561,15 euros. Se pide:

a) Construya una distribución de frecuencias para la ganancia anual por trabajador en el sector de la hostelería en el año 2002 que recoja la información anterior.
b) Indique cuál es la población a la que la distribución se refiere, cuál es la variable observada y de qué tipo es. Indique también, en caso de que tenga sentido, si dicha variable puede considerarse flujo o *stock*.
c) Represente gráficamente la distribución porcentual y la distribución porcentual acumulada.

Ejercicio 1.11. Basándose en la tabla y el gráfico que se muestran a continuación, responda a las siguientes cuestiones:



Fuente: EGATUR, IET

Tabla 1.6. Gasto total de los turistas según edad. Millones de euros y variación interanual

Millones de euros

		2005				2006				Total 2005	Total 2006
		Trim. I	Trim. II	Trim. III	Trim. IV	Trim. I	Trim. II	Trim. III	Trim. IV		
TOTAL		8.009,3	10.980,2	17.049,0	9.967,1	7.868,4	15.522,4	17.707,5	10.128,2	46.005,6	48.226,5
Menor de 15 años	Hombres	108,8	200,5	523,0	150,7	113,8	252,7	602,6	179,1	983,0	1.148,1
	Mujeres	131,2	220,5	537,9	196,7	145,8	278,8	611,7	192,3	1.086,3	1.228,6
De 15 a 24	Hombres	295,8	423,1	866,7	312,8	318,6	463,2	901,3	377,4	1.898,3	2.060,4
	Mujeres	287,9	452,1	933,0	355,3	273,9	486,4	937,0	441,0	2.028,3	2.138,3
De 25 a 44	Hombres	2.057,7	2.763,5	4.046,5	2.364,4	1.865,4	2.878,4	4.278,8	2.518,2	11.232,1	11.540,8
	Mujeres	1.347,9	1.887,2	3.236,6	1.638,7	1.181,8	2.096,7	3.442,4	1.629,8	8.110,4	8.350,6
De 45 a 64	Hombres	1.614,4	2.083,0	2.972,3	2.033,0	1.676,5	2.472,1	2.941,0	1.930,1	8.702,7	9.019,7
	Mujeres	1.270,3	1.718,1	2.540,7	1.721,4	1.342,7	2.107,8	2.504,5	1.631,1	7.250,5	7.586,0
Mayor de 64 años	Hombres	445,5	648,0	713,5	621,8	481,8	761,3	776,6	640,4	2.428,8	2.660,1
	Mujeres	449,9	584,2	679,0	572,3	468,2	725,2	711,5	589,0	2.285,3	2.493,9

Gráfico 1.17. Gasto total de los turistas según edad. Millones de euros y variación interanual

a) ¿Cuál fue el tramo de edad más importante con respecto al gasto turístico en 2006?

b) ¿Qué porcentaje del gasto total realizado en 2006 correspondió a los turistas con edad entre 25 y 64 años?

c) ¿Qué porcentaje del gasto total realizado en 2006 correspondió a los turistas hombres?

d) ¿En qué grupos de edad se han producido los mayores incrementos en gasto turístico, en términos relativos, con respecto a 2005?

CAPÍTULO 2

Análisis descriptivo de una variable

En el tema anterior hemos descrito varias formas de presentar los datos. Las tablas y las representaciones gráficas suponen un primer resumen de las observaciones de una variable; ahora bien, la dimensión de las primeras y la interpretación subjetiva de las segundas obliga a ir más allá, buscando medidas objetivas y numéricas que resuman los aspectos más relevantes de una distribución de frecuencias.

En este tema vamos a estudiar medidas numéricas que nos sirvan para describir las principales características de la distribución de frecuencias de una variable.

Comenzaremos definiendo y analizando distintas medidas de tendencia central o promedios; es decir, medidas que se utilizan para resumir mediante un único valor los distintos valores observados.

En primer lugar estudiaremos la media aritmética, su definición, sus propiedades, el problema de su cálculo cuando se trabaja con datos agrupados y el problema de los intervalos no determinados, como «menor que» o «mayor que».

En el cálculo de la media entran todas las observaciones; esto, que en muchas ocasiones puede ser ventajoso, representa un inconveniente cuando entre las observaciones hay valores atípicos o extremos, ya que la media se ve influida por ellos, resultando ser una medida poco robusta. En aquellas situaciones en las que se considere inadecuado dar el mismo peso a las observaciones extremas que al resto, será preferible utilizar la mediana en lugar de la media.

La mediana se define como el valor de la variable que, ordenadas las observaciones en orden creciente o decreciente, deja a ambos lados el mismo número de elementos. Es, como la media aritmética, un promedio muy utilizado y no se ve afectada por los valores atípicos, ya que al ordenar las observaciones estos quedarán al principio o al final, pero sin importar su diferencia con el resto. En este sentido diremos que la mediana es robusta.

La moda, que es el valor más frecuente en el conjunto de datos, presenta una importante ventaja frente a los anteriores y es que puede utilizarse con datos cualitativos. Puede no ser única, caso en el que pierde representatividad. La presencia de dos o más modas se debe, generalmente, a una mezcla de dos o más grupos heterogéneos y, siempre que sea posible, convendrá analizarlos por separado.

Completaremos la información proporcionada por los promedios con la correspondiente a los cuantiles, medidas de posición que dividen a la distribución en partes que, aproximadamente, contienen el mismo número de observaciones complementarias.

La correcta utilización de un promedio, como valor que sintetiza la información contenida en la distribución, depende de su representatividad, entendida como proximidad o alejamiento de las observaciones al correspondiente promedio. Por esta razón se plantea la necesidad de definir medidas objetivas que cuantifiquen la separación de las observaciones entre sí y con respecto al promedio que las representa. A estas medidas las llamaremos medidas de dispersión.

Una forma simple de medir la dispersión en una distribución es calcular el recorrido, es decir, la diferencia entre los valores máximo y mínimo. Al utilizar solo los dos extremos, esta medida se ve muy afectada por observaciones anómalas y su valor puede inducir a una idea equivocada. Medidas menos sensibles a datos atípicos se consiguen utilizando las diferencias entre cuantiles extremos: Recorridos intercuantílicos.

Otras medidas de dispersión son las que reflejan la diferencia o separación entre los valores observados y el promedio. En este sentido, introduciremos los conceptos de varianza y desviación típica, como medidas de dispersión con respecto a la media, y el coeficiente de variación, medida relativa de dispersión que permite comparar la dispersión con respecto a la media en distintas distribuciones.

La forma de la distribución será analizada numéricamente, para complementar la información gráfica, mediante las medidas de asimetría y de apuntamiento o curtosis.

Abordaremos el análisis de la concentración o desigualdad en el reparto del total, tanto gráficamente, mediante la curva de Lorenz, como numéricamente, mediante el índice de Gini.

Por último, completaremos la descripción de una variable con el análisis del gráfico de caja y bigotes que, como veremos, recoge algunas de las principales características de la distribución, mediana y cuartiles, y que, además de indicarnos cuáles son los valores atípicos, nos permite obtener una idea de la dispersión de los datos y de su asimetría.

2.1. Promedios y medidas de posición

Los promedios o medidas de tendencia central se utilizan con la finalidad de resumir en un único valor toda la información disponible en la distribución de frecuencias. La información que proporciona el promedio se puede completar con la dada por las medidas de posición no centrales, valores que dejan a su izquierda (o por debajo) un porcentaje fijado del total de las observaciones ordenadas en orden creciente.

2.1.1. Media aritmética simple

La media aritmética simple de un conjunto de datos numéricos se define como la suma de todos los datos (distintos o no) dividida por el número de ellos. La media aritmética es un valor de la variable, no necesariamente observable, que viene dado en la misma unidad de medida que la variable.

Cálculo de la media aritmética. La media aritmética de un conjunto de observaciones relativas a una variable $\mathbf{X}$ se representa con $\overline{\mathbf{x}}$ y, atendiendo al tipo de datos con el que estamos trabajando, se calculará como se indica a continuación.

Caso 1. Pocas observaciones. Si $x_1, \ldots, x_N$ son las **N** observaciones de una variable **x**, su media $\bar{x}$ se calcula mediante la siguiente fórmula

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Ejemplo 2.1. El gasto semanal en productos perecederos (en euros) de cuatro familias fue de 104,51, 105, 120 y 200, respectivamente. ¿Cuál fue el gasto medio?

$$\bar{x} = \frac{104{,}51 + 105 + 120 + 200}{4} = \frac{529{,}51}{4} = 132{,}38 \text{ euros}$$

Caso 2. Muchas observaciones, datos no agrupados. Si $x_1, \ldots, x_k$ son los distintos valores observados de una variable **X** en un total de **N** observaciones y $n_1, \ldots, n_k$ son sus correspondientes frecuencias, su media $\bar{x}$ se calcula mediante la fórmula

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

Ejemplo 2.2. La distribución de frecuencias del gasto semanal (en euros) de un grupo de 120 familias es la que figura en la siguiente tabla. ¿Cuál fue el gasto medio?

Tabla 2.1. Distribución del gasto

Gasto (euros)	Familias
104	40
105	30
120	10
200	40
	120

Para calcular la media, $\bar{x}$, construimos una columna auxiliar en la que figuren los productos $x_i \cdot n_i$, que representan los totales correspondientes a cada valor de la variable. Por ejemplo, las 40 familias con un gasto de 104 euros, tienen un gasto total de $104 \cdot 40 = 4.160$ euros. La media es la suma de los elementos de esa columna dividida por el total de observaciones.

Tabla 2.2. Distribución del gasto (ampliación)

Gasto (euros)	Familias	$x_i \cdot n_i$
104	40	4.160
105	30	3.150
120	10	1.200
200	40	8.000
Totales	**120**	**16.510**

Luego, $\bar{x} = \dfrac{16.510}{120} = 137{,}583$ **euros.**

Caso 3. Muchas observaciones, datos agrupados. En este caso, como no podemos calcular el valor exacto de la media, ya que no conocemos los valores observados, damos como media aritmética la de la distribución de datos no agrupados que resulta al sustituir cada intervalo por su marca de clase. El valor resultante será una aproximación del valor medio real de la distribución. Concretamente, si las N observaciones de una variable X están agrupadas en k intervalos a los que corresponden las frecuencias $n_1, \ldots, n_k$, su media $\bar{x}$ se calcula mediante la fórmula

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}$$

siendo $x_1, \ldots, x_k$ las marcas de clase de los k intervalos.

Ejemplo 2.3. La distribución de frecuencias del gasto (en euros) de un grupo de 1.200 familias es la que figura en la siguiente tabla. ¿Cuál fue el gasto medio?

Tabla 2.3. Distribución del gasto

Gasto (euros)	Familias
100-150	400
150-175	300
175-225	400
225-300	100
	1.200

Para calcular la media, $\bar{x}$, construimos una columna auxiliar en la que figuren las marcas de clase (x_i) y otra con los productos $x_i \cdot n_i$ que, a diferencia del caso 2, representan una estimación del gasto total de las familias que figura en cada intervalo. Esta estimación conlleva un error de agrupamiento. Daremos como media una aproximación de la media real, la suma de los elementos de esa columna dividida por el total de observaciones.

Tabla 2.4. Distribución del gasto (ampliación)

Gasto (euros)	Familias (n_i)	x_i	$x_i n_i$
100-150	400	125	**50.000**
150-175	300	162,5	**48.750**
175-225	400	200	**80.000**
225-300	100	262,5	**26.250**
Totales	**1.200**		**205.000**

Luego, $\bar{x} = \dfrac{205.000}{1.200} = 170{,}83$ **euros.**

2.1.2. Media aritmética ponderada

A veces, la obtención de la media de una variable requiere tener en cuenta la importancia relativa de los elementos de la población porque esta es distinta.

Ejemplo 2.4. En una asignatura se han realizado tres exámenes durante el curso. El segundo vale doble que el primero y el tercero vale triple que el primero. La nota final es la media de las tres calificaciones. Si las calificaciones de un alumno en los correspondientes exámenes son 7, 6 y 4, su calificación final será:

$$\bar{c} = \frac{7 + 6\cdot 2 + 4\cdot 3}{6} = \frac{31}{6} = 5{,}17 \text{ puntos}$$

La media así calculada se denomina media ponderada.

En general, se llama media aritmética ponderada de las N observaciones de una variable X, $x_1,\ldots, x_N$, respecto de los pesos o ponderaciones $w_1,\ldots, w_N$, al número definido como

$$\bar{x}_p = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i}$$

siendo los pesos números no negativos y al menos uno positivo.

2.1.3. Propiedades de la media aritmética

a) Si x_m y x_M representan al menor y al mayor valor observado, respectivamente, $x_m \leq \bar{x} \leq x_M$.

b) La media es el centro de gravedad de la distribución, es decir, la suma de las desviaciones entre las observaciones y su media es nula. Formalmente lo escribimos

$$\sum_{i=1}^{k} (x_i - \bar{x}) \cdot n_i = 0$$

c) La suma de los cuadrados de las desviaciones entre las observaciones y una constante cualquiera se hace mínima cuando dicha constante es la media. Formalmente lo escribimos

$$\min_c \sum_{i=1}^{k} (x_i - c)^2 \cdot n_i = \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i$$

d) Si a todos los valores de la variable les sumamos una constante, su media queda aumentada en dicha constante; es decir, a la media le afecta un cambio de origen en la variable del mismo modo que a esta. Formalmente, si de la distribución (x_i, n_i), $i = 1,\ldots, k$, pasamos a la distribución (y_i, n_i), $i = 1,\ldots, k$, siendo $y_i = x_i + a$, se cumple que $\bar{y} = \bar{x} + a$.

e) Si todos los valores de la variable se multiplican por una constante, su media queda multiplicada por dicha constante; es decir, a la media le afecta un cambio de «escala» en la variable del mismo modo que a esta. Formalmente, si de la distribución (x_i, n_i), $i = 1,\ldots, k$, pasamos a la distribución (y_i, n_i), $i = 1,\ldots, k$, siendo $y_i = b \cdot x_i$, se cumple que $\overline{y} = b \cdot \overline{x}$.

f) Si las observaciones de una variable están divididas en varios grupos disjuntos (los grupos no tienen elementos en común y cada elemento pertenece a un grupo), la media del grupo completo se obtiene como media ponderada de las medias de los subgrupos, siendo las ponderaciones los tamaños de los subgrupos a los que corresponden las medias.

Grupo	1	2	...	r
Media	x_1	x_2	...	x_r
Tamaño	N_1	N_2	...	N_r

$$\overline{x} = \frac{\sum_{i=1}^{r} N_i \cdot \overline{x}_i}{N} = \frac{N_1 \cdot \overline{x}_1 + N_2 \cdot \overline{x}_2 + \ldots + N_r \cdot \overline{x}_r}{N}$$

Ventajas e inconvenientes de la utilización de la media como promedio

- Es una medida de posición central sencilla de calcular.
- La media se define de forma objetiva y es única para cada distribución.
- En su cálculo intervienen todas las observaciones.
- Utilizar toda la información representa un inconveniente cuando entre las observaciones hay valores atípicos o extremos, ya que la media se va a ver influida por ellos y puede resultar poco representativa del conjunto.
- En aquellas situaciones en las que se considere inadecuado dar mucho peso a las observaciones extremas, se preferirá utilizar como promedio la mediana en lugar de la media aritmética.

2.1.4. La mediana

La mediana se define como el valor de la variable que, ordenadas las observaciones en orden creciente o decreciente, ocupa el lugar central; es decir, aquel valor que deja a ambos lados el mismo número de observaciones.

Cálculo de la mediana

Caso 1. Pocas observaciones.

a) Si el número de observaciones **(N)** es **impar** y están ordenadas en orden creciente, la mediana es la observación que ocupa el lugar central; es decir,

$$Me_x = x_{\frac{N+1}{2}}$$

Ejemplo 2.5. Los gastos semanales (en miles de euros) de siete empresas son:

30, 50, 60, 70, 110, 120, 140

La mediana es 70.000 euros, la observación que ocupa el cuarto lugar.

b) Si el número de observaciones **(N)** es **par** y las observaciones están ordenadas, la mediana es la media de las dos observaciones que ocupan los lugares centrales. Formalmente,

$$\mathbf{Me_x} = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$$

Ejemplo 2.6. Los gastos semanales (en miles de euros) de ocho empresas son:

15, 25, 45, 55, 65, 85, 105, 115

La mediana es $\frac{55+65}{2}$ **= 60 miles de euros, es decir, 60.000 euros.**

Caso 2. Muchas observaciones y datos no agrupados.

Exactamente lo mismo que en el caso anterior.

Si **N** es **impar,** $\mathbf{Me_x} = x_{\frac{N+1}{2}}$ y si **N** es **par,** $\mathbf{Me_x} = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}$

Ejemplo 2.7. Las facturaciones semanales (en millones de euros) de tres multinacionales A, B y C con 39, 60 y 50 empresas, respectivamente, son las que figuran en las siguientes tablas. Calcule la facturación semanal que se corresponde con la mediana en cada una de las multinacionales.

a)

Tabla 2.5. Distribución de la facturación (Multinacional A)

x_i	n_i	N_i	
10	10	10	
20	12	22	←
30	7	29	
40	7	36	
50	3	39	
	39		

N = 39 es impar. $\mathbf{Me_x} = x_{\frac{N+1}{2}} = x_{20} =$ **20 millones de euros**

b)

Tabla 2.6. Distribución de la facturación (Multinacional B)

x_i	n_i	N_i
20	12	12
40	10	22
60	8	30
80	7	37
100	5	42
120	8	50
140	10	60
	60	

N = 60 es par.

$$Me_x = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} = \frac{x_{30} + x_{31}}{2} = \frac{60 + 80}{2} = 70 \text{ millones de euros}$$

c)

Tabla 2.7. Distribución de la facturación (Multinacional C)

x_i	n_i	N_i
100	18	18
125	13	31
150	10	41
200	9	50
	50	

N = 50 es par.

$$Me_x = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} = \frac{x_{25} + x_{26}}{2} = \frac{125 + 125}{2} = 125 \text{ millones de euros}$$

Caso 3. Muchas observaciones y datos agrupados.

Como no disponemos de toda la información relativa a los valores observados, nos tenemos que conformar con calcular un valor aproximado de la mediana.

El valor aproximado de la mediana es el que acumula frecuencia **N/2,** bajo la hipótesis de reparto uniforme dentro de cada intervalo. Concretamente, si alguno de los intervalos acumula frecuencia **N/2,** entonces la mediana es su extremo superior. En otro caso, la mediana se encuentra en el primer intervalo cuya frecuencia acumulada es mayor que **N/2 (intervalo mediano).** Si el intervalo mediano es L_{i-1} - L_i, con un reparto uniforme dentro del mismo, la situación es la que se presenta en la siguiente figura.

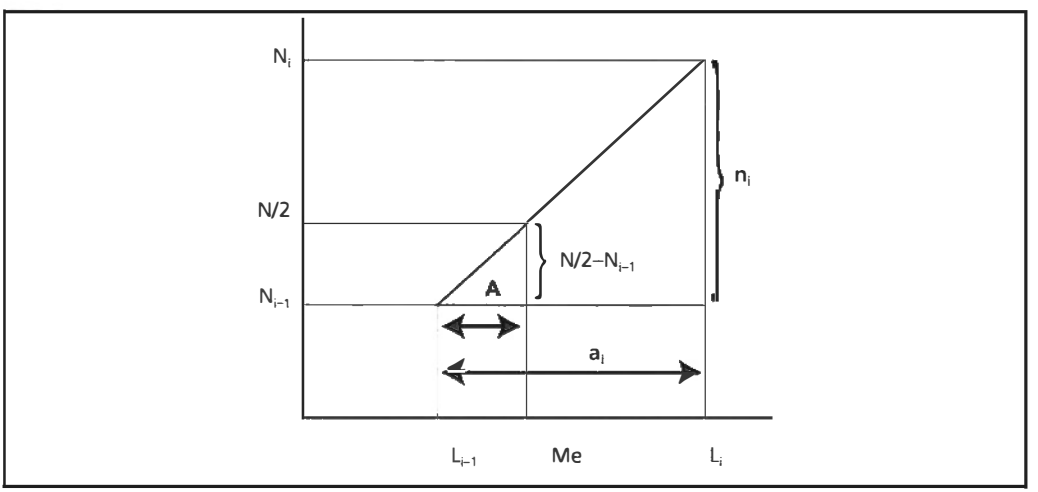


Figura 2.1. Determinación gráfica de la mediana

Donde

$$\frac{A}{a_i} = \frac{\frac{N}{2} - N_{i-1}}{n_i}$$

Por lo tanto,

$$A = \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i$$

Y, como $A = Me - L_{i-1}$, se cumple que

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

Ejemplo 2.8. En un grupo de 50 personas, la cantidad (en euros) de la que podrían disponer para gastarse en cierta actividad es la que se refleja en la siguiente tabla. Calcule la mediana de dicha distribución.

Tabla 2.8. Distribución del gasto

Cantidad	N.° de personas	N_i
40-100	10	10
100-200	20	30
200-500	15	45
500-1.000	5	50
	50	

N/2=25. El primer intervalo cuya frecuencia acumulada es mayor que 25 es 100-200 y, por lo tanto, ese es el intervalo mediano. Aplicando la fórmula anterior se obtiene

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 100 + \frac{25 - 10}{20} \cdot 100 = 100 + \frac{1.500}{20} = 175 \text{ euros}$$

Ventajas e inconvenientes de la utilización de la mediana como promedio

- Es una medida de posición central sencilla de calcular.
- Tiene una fácil interpretación.
- En la determinación de la mediana solo intervienen los valores centrales de la distribución y es, por lo tanto, insensible a los valores extremos.
- Puede calcularse en distribuciones en las que los valores extremos sean desconocidos, siempre que tengamos información sobre las frecuencias con las que estos se presentan (por ejemplo, si los intervalos inicial y final son abiertos).

2.1.5. La moda

La moda **(Mo)** es el valor o valores de la variable que más veces se repite(n); es decir, el valor o valores de la variable al (los) que corresponde la mayor frecuencia.

La moda no tiene por qué ser única, las distribuciones reciben el nombre de unimodales, bimodales, etc., según tengan una, dos, etc., modas, respectivamente. Al ser un valor de la variable, viene expresado en la misma unidad de medida que esta.

No existe una fórmula general para expresar este promedio. Para calcularlo consideramos los tres casos distintos con los que nos podemos encontrar.

Caso 1. Pocas observaciones.

Su cálculo es inmediato, basta con localizar el valor que más veces se repite.

Ejémplo 2.9. Los gastos diarios (en euros) de un grupo de 5 personas fueron 100, 100, 200, 50, 70. **La moda es 100 euros.**

Caso 2. Muchas observaciones, datos no agrupados.

Para determinarla, se observa la columna de las frecuencias absolutas y la moda es el valor de la variable que se corresponde con la mayor frecuencia absoluta.

Ejemplo 2.10. Los precios de cotización de las acciones (en euros) de un grupo de empresas en un día determinado quedan reflejados en la siguiente tabla. Calcule el precio de cotización más frecuente en ese día.

Tabla 2.9. Distribución de las cotizaciones

Precio cotización (euros)	Empresas
50	2
70	4
75	4
90	**6**
100	3
115	5

El precio de cotización más frecuente es 90 euros.

Ejemplo 2.11. Los gastos diarios en servicios de transporte (en miles euros) de un grupo de empresas fueron los que se presentan en la siguiente tabla. Obtenga el gasto más frecuente.

Tabla 2.10. Distribución del gasto en servicios de transporte

Gasto (euros)	Empresas
40	1
43	2
54	**6**
60	4
85	**6**

Esta distribución tiene dos modas: **Mo^1=54.000 euros** y **Mo^2=85.000 euros**.

Caso 3. Muchas observaciones y datos agrupados.

Como para los otros promedios ya estudiados, en este caso no podemos calcular el valor exacto de la moda. Se define el intervalo modal como aquel que tiene mayor frecuencia por unidad de amplitud. Puede ser único o no.

Si la amplitud de todos los intervalos es la misma, podemos decir que el intervalo modal es el que presenta mayor frecuencia. Si los intervalos son de amplitud variable, el intervalo modal es el que lleva asociada la mayor altura.

La determinación del valor que daremos como moda dentro del intervalo modal se hace según el siguiente criterio: «la distancia de la moda a los extremos del intervalo modal debe ser inversamente proporcional a la frecuencia (altura, si la amplitud es variable) del intervalo contiguo a dicho extremo». La aplicación de este criterio proporciona las fórmulas para el cálculo de la moda cuando trabajamos con datos agrupados. Estas son

$$\mathbf{Mo} = \mathbf{L}_{i-1} + \frac{\mathbf{n}_{i+1}}{\mathbf{n}_{i-1} + \mathbf{n}_{i+1}} \cdot \mathbf{a}_i \quad \text{(amplitud constante)}$$

$$\mathbf{Mo} = \mathbf{L}_{i-1} + \frac{\mathbf{h}_{i+1}}{\mathbf{h}_{i-1} + \mathbf{h}_{i+1}} \cdot \mathbf{a}_i \quad \text{(amplitud variable)}$$

Ejemplo 2.12.

2.12.1. La cantidad de dinero (en euros) que los integrantes de un grupo de 280 jóvenes llevan en sus bolsillos se recoge en la siguiente tabla.

Tabla 2.11. Distribución del gasto

Dinero (euros)	N.° de jóvenes
0-5	4
5-10	21
10-15	**109**
15-20	78
20-25	44
25-30	24
	280

Los intervalos tienen todos amplitud 5. El de mayor frecuencia es el intervalo [10,15) y ese es, por lo tanto, el modal.

La moda es

$$\mathbf{Mo} = \mathbf{L}_{i-1} + \frac{\mathbf{n}_{i+1}}{\mathbf{n}_{i-1} + \mathbf{n}_{i+1}} \cdot \mathbf{a}_i = 10 + \frac{78}{21 + 78} \cdot 5 = \mathbf{13{,}94 \text{ euros}}$$

2.12.2. Observados los precios de venta de 390 modelos de televisor en decenas de euros, se ha obtenido la siguiente distribución.

Tabla 2.12. Distribución del precio de venta

Precio de venta (10 €)	N.° de modelos
0-25	20
25-50	140
50-100	180
100-150	40
150-200	10
	390

Obtener la moda de la distribución del precio de venta para ese conjunto de modelos.

Como los intervalos son de amplitud variable, el intervalo modal es el de mayor altura. Añadimos a la tabla anterior una columna en la que figuren las alturas.

Tabla 2.13. Distribución del precio de venta (ampliación)

Precio de venta (10 €)	N.° de modelos	$h_i = n_i/a_i$
0-25	20	0,8
25-50	140	**5,6**
50-100	180	3,6
100-150	40	0,8
150-200	10	0,2
	390	

A la vista de los elementos de esa columna se concluye que el intervalo modal es [25,50). La moda es, por lo tanto

$$\mathbf{Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i = 25 + \frac{3,6}{3,6 + 0,8} \cdot 25 = 45,455 \text{ decenas de euros}}$$

Ventajas e inconvenientes de la moda

- La moda es el único de los promedios estudiados que está definido para caracteres de tipo cualitativo, además de para variables.
- Es de cálculo sencillo, de fácil interpretación y en distribuciones de datos no agrupados siempre coincide con un valor observado.
- En su determinación no intervienen todos los valores de la distribución.
- No es única.

2.1.6. Otras medidas de posición (cuantiles)

Los cuantiles son valores de la variable que dividen a la distribución de los valores ordenados en orden creciente en partes iguales; es decir, en intervalos que comprenden el mismo número de observaciones. Los cuantiles más utilizados son: los cuartiles, los deciles y los percentiles (o centiles).

Los cuartiles (Q_j) son los tres valores de la variable que dividen a la distribución en cuatro partes iguales; es decir, en cuatro intervalos tales que dentro de cada uno de ellos se encuentra, aproximadamente, el 25% de los datos.

Para estadísticas de datos agrupados

$$\mathbf{Q_j = L_{i-1} + \frac{\frac{jN}{4} - N_{i-1}}{n_i} \cdot a_i, \quad j = 1,2,3}$$

donde $\mathbf{L_{i-1}}$-$\mathbf{L_i}$ es el primer intervalo cuya frecuencia acumulada es mayor que $\frac{jN}{4}$, es decir, $\mathbf{N_{i-1} \leq \frac{jN}{4} < N_i}$

Los deciles (D_j) son los valores (nueve) de la variable que dividen a la distribución en diez partes iguales, de modo que cada una de ellas contiene el 10% de los datos.

Para estadísticas de datos agrupados

$$D_j = L_{i-1} + \frac{\frac{jN}{10} - N_{i-1}}{n_i} \cdot a_i, \quad j = 1,\ldots,9$$

donde L_{i-1}-L_i es el primer intervalo cuya frecuencia acumulada es mayor que $\frac{jN}{10}$, es decir, $N_{i-1} \leq \frac{jN}{10} < N_i$.

Los percentiles (Pj) son los 99 valores de la variable que dividen a la distribución en 100 partes iguales, de modo que en cada una de ellas se encuentra el 1% de los datos.

Para estadísticas de datos agrupados

$$P_j = L_{i-1} + \frac{\frac{jN}{100} - N_{i-1}}{n_i} \cdot a_i, \quad j = 1,\ldots,99$$

donde L_{i-1}-L_i es el primer intervalo cuya frecuencia acumulada es mayor que $\frac{jN}{100}$ es decir, $N_{i-1} \leq \frac{jN}{100} < N_i$.

Cuando se trata de pocas observaciones o muchas observaciones pero con datos no agrupados, los cuartiles, deciles y percentiles se calculan de modo análogo a como se calcula la mediana.

Ejemplo 2.13. El número de días al mes que utilizan el cajero automático de una sucursal bancaria 120 clientes de la misma queda recogido en la siguiente tabla.

Tabla 2.14. Distribución del uso de un cajero (días)

Días	Clientes
5	20
7	30
8	20
9	40
11	7
13	3
	120

Determinar los tres cuartiles y el séptimo decil.

Solución. En primer lugar, obtenemos la distribución de frecuencias acumuladas.

Tabla 2.15. Distribución del uso de un cajero (ampliación)

x_i	n_i	N_i
5	20	20
7	30	50
8	20	70
9	40	110
11	7	117
13	3	120
	120	

$\frac{N}{4} = \frac{120}{4} = 30$. Por lo tanto, $Q_1 = 7$ **días**.

$\frac{2N}{4} = \frac{240}{4} = 60$, luego $Q_2 = Me = 8$ **días**.

Por último, $\frac{3N}{4} = \frac{360}{4} = 90$ y, en consecuencia, $Q_3 = 9$ **días**.

Para calcular el decil 7, calculamos $\frac{7N}{10} = \frac{7 \cdot 120}{10} = 84$ y, teniendo en cuenta la distribución de frecuencias acumuladas, se obtiene que $D_7 = 9$ **días**.

Ejemplo 2.14. Observados los precios de venta de 390 modelos de televisor, se ha obtenido la siguiente distribución

Tabla 2.16. Distribución del precio de venta

Precio de venta (10 €)	N.° de modelos
0-25	20
25-50	140
50-100	180
100-150	40
150-200	10
	390

Determine los deciles 2 y 8 y los percentiles 75 y 95.

Solución. La distribución de frecuencias acumuladas es la que se muestra en la siguiente tabla.

Tabla 2.17. Distribución del precio de venta (ampliación)

Precio de venta (10 €)	N.° de modelos	N_i
0-25	20	20
25-50	140	160
50-100	180	340
100-150	40	380
150-200	10	390
	390	

2N/10=78 → $D_2 = 353{,}6$ €
8N/10=312 → $D_8 = 922{,}2$ €
75N/100=292,5 → $P_{75} = 868{,}1$ €
95N/100=370,5 → $P_{95} = 1.381{,}3$ €

2.2. Medidas de dispersión

En el apartado anterior se han definido algunos promedios o medidas de posición central, con ellas se pretende resumir en un único valor la información sobre la tendencia de las observaciones en la zona central de la distribución. Sin embargo, su correcta utilización como valores que sintetizan dicha información depende de su representatividad, entendida como proximidad o alejamiento de las observaciones al correspon-

diente promedio. Por esta razón se plantea la necesidad de definir medidas objetivas que cuantifiquen la separación de las observaciones entre sí y con respecto al promedio que se supone las representa, a estas medidas se les llama **medidas de dispersión**.

2.2.1. Medidas de dispersión absolutas

Recorridos

El recorrido (R) de un conjunto de observaciones de una variable **X** se define como la diferencia entre el mayor y el menor valor de los observados.

$$R = x_M - x_m$$

Es una medida que viene expresada en la misma unidad de medida que la variable. Cuanto mayor es el recorrido, mayor es el campo de variación de la variable y también su dispersión. Es fácil de calcular pero, como solo depende de los valores extremos, basta con que uno de los extremos esté anormalmente alejado (dato atípico) para obtener una idea falsa sobre la dispersión en el conjunto de observaciones.

Para obtener una medida menos sensible a datos atípicos, se definen los **recorridos intercuantílicos**, como la diferencia entre el mayor y el menor cuantil. Surge así, el **recorrido intercuartílico, R_Q**, que se define como

$$R_Q = Q_3 - Q_1$$

Esta medida cuantifica la longitud del intervalo en el que se encuentran las observaciones centrales que suponen el 50% del total. Cuanto menor es su valor, medido en la misma unidad que la variable, menor es la dispersión en la distribución de los valores centrales de la misma.

De manera análoga, el **recorrido interdecílico** (R_D) y el recorrido intercentílico (R_c) se definen como

$$R_D = D_9 - D_1 \quad \text{y} \quad R_c = P_{99} - P_1$$

e indican las longitudes de los intervalos en los que se encuentran el 80% y el 98% de las observaciones centrales, respectivamente.

Ejemplo 2.15. El número de días al mes que utilizan el cajero automático de una sucursal bancaria 120 clientes de la misma queda recogido en la siguiente tabla.

Tabla 2.18. Distribución del uso de un cajero (días)

Días	Clientes
5	20
7	30
8	20
9	40
11	7
13	3
	120

Calcule el recorrido y los recorridos intercuartílico e interdecílico para dicha distribución.

Solución.

El recorrido es 8 días. Por otra parte, teniendo en cuenta la información que contiene la siguiente tabla

Tabla 2.19. Distribución del uso de un cajero (ampliación)

Días	Clientes	N_i
5	20	20
7	30	50
8	20	70
9	40	110
11	7	117
13	3	120
	120	

y que

$$N/4=30$$

$$3N/4=90$$

resulta que Q_1 = **7 días**, Q_3 = **9 días** y, por lo tanto, R_Q = **2 días.**

De modo análogo, para obtener R_D, nos basamos en que **N/10=12 y 9N/10=108,** lo que, junto con la información contenida en la Tabla 2.19, nos permite concluir que D_1 = **5 días,** D_9 = **9 días** y que, por lo tanto, R_D = **4 días**.

Todas estas medidas, fáciles de calcular, tienen el inconveniente de no reflejar la «separación» entre los datos y el «centro» de la distribución, ya que en su definición no interviene ningún promedio. En consecuencia, no nos dan información sobre la representatividad o no del promedio que se haya calculado. De entre las medidas de dispersión que involucran en su definición a algún promedio y que, por lo tanto, nos pueden proporcionar información sobre su representatividad, nos centraremos en la varianza y la desviación típica (ambas basadas en la media aritmética).

Varianza

Para un conjunto de **N** observaciones de una cierta variable **X**, $x_1,\ldots, x_N$, se define su **varianza,** a la que representamos mediante S_x^2, como

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

donde $\bar{x}$ es la media aritmética de ese conjunto de observaciones.

La varianza, al ser una media de los cuadrados de las desviaciones entre las observaciones y su media, $(x_i - \bar{x})$, viene dada en el cuadrado de la unidad en que venga expresada la variable.

Cuanto más concentrados estén los valores alrededor de la media aritmética, más próximas a cero serán las desviaciones y, en consecuencia, menor la varianza. Si, por el contrario, las observaciones están muy dispersas con respecto a la media, los cuadrados de las desviaciones serán grandes y, por tanto, la varianza será grande.

Fórmulas para el cálculo de la varianza

Es fácil ver que

$$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N}x_i^2}{N} - (\bar{x})^2$$

Por otra parte, la expresión de la fórmula anterior en términos de los distintos valores observados de la variable, $x_1, \ldots, x_k$, y sus correspondientes frecuencias absolutas es

$$S_x^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - (\bar{x})^2$$

Cuando trabajamos con datos agrupados, obtenemos un valor aproximado. Dicho valor se consigue aplicando la fórmula anterior a la distribución de datos no agrupados que resulta al sustituir cada intervalo por su marca de clase.

Propiedades de la varianza

- $S^2 \geq 0$.
- $S^2 = 0 \Leftrightarrow$ Todos los datos son iguales.
- La varianza es invariante frente a cambio de origen; es decir, si de (x_i, n_i) pasamos a (y_i, n_i) con $y_i = x_i + a$, entonces $S_y^2 = S_x^2$.
- Si de (x_i, n_i) pasamos a (y_i, n_i) con $y_i = b \cdot x_i$, entonces $S_y^2 = b^2 \cdot S_x^2$; es decir, a la varianza le afectan los cambios de escala en la variable.

Ejemplo 2.16. La distribución de frecuencias del gasto semanal en alimentación (expresado en euros) de un grupo de 1.200 familias es la que figura en la siguiente tabla. Determine la varianza del gasto semanal en ese grupo de familias.

Tabla 2.20. Distribución del gasto

Gasto (euros)	Familias
100-150	400
150-175	300
175-225	400
225-300	100
	1.200

Para calcular la varianza utilizamos la fórmula

$$S_x^2 = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - (\bar{x})^2$$

donde x_i representa a la marca de clase del intervalo L_{i-1} - L_i. En la **Tabla 2.21** figuran los cálculos intermedios que son necesarios.

Tabla 2.21. Distribución del gasto (ampliación)

Gasto (euros)	Familias	x_i	$x_i\, n_i$	$x_i^2\, n_i$
100-150	400	125	50.000	6.250.000
150-175	300	162,5	48.750	7.921.875
175-225	400	200	80.000	16.000.000
225-300	100	262,5	26.250	6.890.625
Totales	**1.200**		205.000	37.062.500

A partir de los cálculos anteriores concluimos que el gasto semanal medio es

$$\bar{x} = \frac{205.000}{1.200} = 170{,}83 \text{ euros}$$

y que la varianza del gasto semanal es

$$S_x^2 = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - (\bar{x})^2 = \frac{37.062.500}{1.200} - 170{,}83^2 = 1.702{,}53 \text{ euros}^2$$

Desviación típica

La desviación típica es la raíz cuadrada positiva de la varianza, se representa con S_x. Su interpretación es similar a la de la varianza, con la diferencia de que la desviación típica viene medida en la misma unidad que la variable. Cumple las mismas propiedades que la varianza, teniendo en cuenta que cuando se efectúa el cambio $y_i = b \cdot x_i$, se cumple que $S_y = |b| \cdot S_x$.

La desviación típica que corresponde a la distribución del **Ejemplo 2.16** es

$$S_x = +\sqrt{1.702{,}53} = 41{,}26 \text{ euros}$$

Variable tipificada

Si una variable **X** tiene desviación típica no nula, podemos transformarla en otra (que se llama tipificada de **X**) y que tiene media nula y desviación típica uno. Concretamente, dada la variable **X** con medida $\bar{x}$ y desviación típica $S_x > 0$, llamamos variable tipificada de **X** a la variable **Z** definida por

$$Z = \frac{X - \bar{x}}{S_x}$$

Para cada valor x_i de **X** se obtiene el valor z_i de **Z** mediante $z_i = \frac{x_i - \bar{x}}{S_x}$.

Las variables tipificadas se suelen utilizar para comparar las posiciones relativas de los valores de variables que tienen medias distintas o que vienen expresadas en unidades que no son comparables.

Ejemplo 2.17. Un alumno obtuvo la calificación de 8,4 en Matemáticas y de 9 en Estadística, ¿en cuál de las dos asignaturas obtuvo mejor calificación con respecto a su clase si $\bar{x}_M = 7{,}6$; $S_M = 1$; $\bar{x}_E$ y $S_E = 1{,}6$?

Para poder comparar las dos notas tipificamos (situamos el origen de ambas calificaciones en cero y las medimos en unidades de desviación típica)

$$z_M = \frac{x_M - \bar{x}_M}{S_M} = \frac{8{,}4 - 7{,}6}{1} = 0{,}8 \quad \text{y} \quad z_E = \frac{x_E - \bar{x}_E}{S_E} = \frac{9 - 8{,}2}{1{,}6} = 0{,}5$$

Por lo tanto, con respecto a la clase obtuvo mejor nota en Matemáticas.

Propiedades de la variable tipificada

- $\bar{z} = 0$
- $S_z = 1$

2.2.2. Medidas de dispersión relativas

La comparación de la dispersión de distintas variables puede hacerse a través de la desviación estándar si las variables que se comparan tienen medias próximas y vienen expresadas en la misma unidad. La dispersión será menor en la distribución con menor varianza. Cuando esto no ocurre, para comparar la dispersión existente en distintas distribuciones, se emplean medidas de dispersión adimensionales o relativas.

$$\textbf{Dispersión relativa} = \frac{\textbf{Dispersión absoluta}}{|\textbf{Promedio}|}$$

De ellas, la más utilizada es el coeficiente de variación de Pearson.

Coeficiente de variación.

El coeficiente de variación de la distribución de una variable **X**, CV_x, se define como $CV_x = \frac{S_x}{|\bar{x}|}$ y, generalmente, se expresa en tanto por ciento, es decir,

$$CV_x = \frac{S_x}{|\bar{x}|} \cdot 100$$

Para que esté bien definido, la media debe ser distinta de cero.

Es claro que no se ve afectado por la unidad en que venga la variable, ya que numerador y denominador están expresados en la misma unidad y, en consecuencia, el cociente es adimensional.

Cuanto mayor es el coeficiente de variación, mayor es la dispersión y menos representativa la media.

Ejemplo 2.18. El salario mensual de dos grupos de trabajadores de empresas similares ubicadas en países distintos es el que figura en las siguientes tablas.

Grupo A

Tabla 2.22. Distribución del salario mensual en el país A

Salario (10^3 €)	N.° de trabajadores
10	14
12	16
14	20
16	15
18	18
20	17
	100

Grupo B

Tabla 2.23. Distribución del salario mensual en el país B

Salario (10^3 $)	N.° de trabajadores
8	5
10	10
11	15
13	30
15	20
18	16
20	4
	100

Se pide:

a) Salario medio entre los trabajadores de la empresa ubicada en el país A.
b) Salario medio entre los trabajadores de la empresa ubicada en el país B.
c) ¿En qué grupo es más representativo el salario medio?
d) ¿Qué grupo es más homogéneo en cuanto al salario?

Solución

Para el grupo A

Tabla 2.24. Distribución del salario mensual en el país A (ampliación)

Salario (10^3 €)	N.° de trabajadores	$x_i\, n_i$	$x_i^2\, n_i$
10	14	140	1.400
12	16	192	2.304
14	20	280	3.920
16	15	240	3.840
18	18	324	5.382
20	17	340	6.800
Totales	**100**	**1.516**	**24.096**

Teniendo en cuenta la información contenida en la tabla y aplicando las correspondientes fórmulas de cálculo, obtenemos que la media en el grupo del país **A** es $15{,}16 \cdot 10^3$ € y que la desviación típica es $3{,}337 \cdot 10^3$ €.

Para el grupo B resulta

Tabla 2.25. Distribución del salario mensual en el país B (ampliación)

Salario (10^3 $)	N.° de trabajadores	$x_i\, n_i$	$x_i^2\, n_i$
8	5	40	320
10	10	100	1.000
11	15	165	1.815
13	30	390	5.070
15	20	300	4.500
18	16	288	5.184
20	4	80	1.600
	100	**1.363**	**19.489**

Por lo tanto, la media en el grupo del país **B** es $13{,}63 \cdot 10^3$ $ y la desviación típica resulta $3{,}019 \cdot 10^3$ $.

Para comparar la representatividad de ambas medias, calculamos los coeficientes de variación, resultando que $\mathbf{CV_A = 22{,}01\%}$ y que $\mathbf{CV_B = 22{,}15\%}$, por lo que, el valor medio más representativo es el correspondiente a **A** y es, además, en el grupo de trabajadores de ese país en donde el salario presenta menos dispersión o es más homogéneo.

2.3. Medidas de forma

Las medidas de forma complementan la información que, sobre la distribución de frecuencias, proporcionan los promedios y las medidas de posición y dispersión que hemos abordado en los epígrafes previos.

El punto de partida en el estudio de la forma de una distribución de frecuencias es su representación gráfica mediante el diagrama de barras (datos no agrupados) o el histograma

(datos agrupados). De la simple inspección visual de dicha gráfica, podremos concluir si la distribución tiene una única moda o no; si presenta mucha variabilidad o no; si tiene una forma regular (las más frecuentes son: campana, L, J y U) o no.

En este apartado estamos interesados en proporcionar medidas que permitan cuantificar el grado de asimetría y apuntamiento existente en la distribución.

2.3.1. Medidas de asimetría

Diremos que una distribución es simétrica cuando valores equidistantes de uno central (promedio) se presentan con la misma frecuencia (Gráfico 2.1).

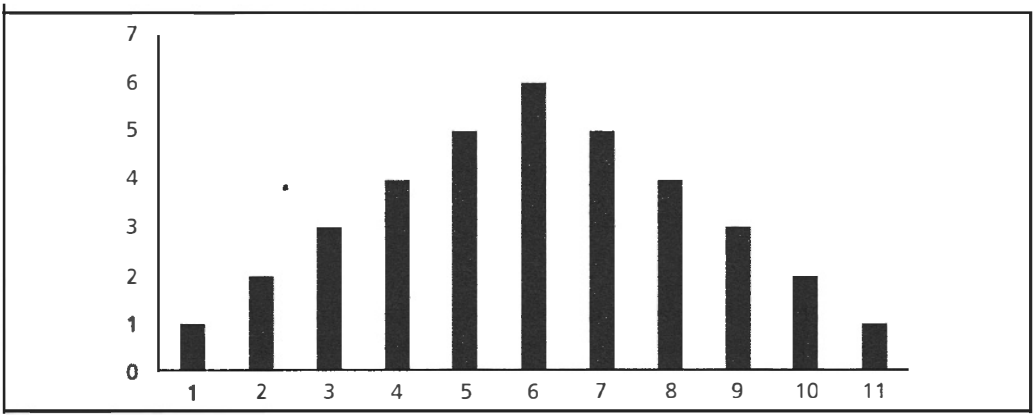


Gráfico 2.1. Ejemplo de una distribución simétrica

Cuando, además de ser simétrica, la distribución (diagrama de barras o histograma) tiene forma de campana, como en el dibujo anterior, los tres promedios estudiados, media, mediana y moda, coinciden.

Si la distribución no es simétrica, decimos que es asimétrica, distinguiendo entre: asimétricas positivas o a la derecha (Gráfico 2.2) y asimétricas negativas o a la izquierda (Gráfico 2.3).

Las distribuciones asimétricas positivas se caracterizan porque la gráfica presenta «cola a la derecha», es decir, las frecuencias descienden más lentamente por esa zona; en las asimétricas por la izquierda, las frecuencias descienden más lentamente por «la cola de la izquierda».

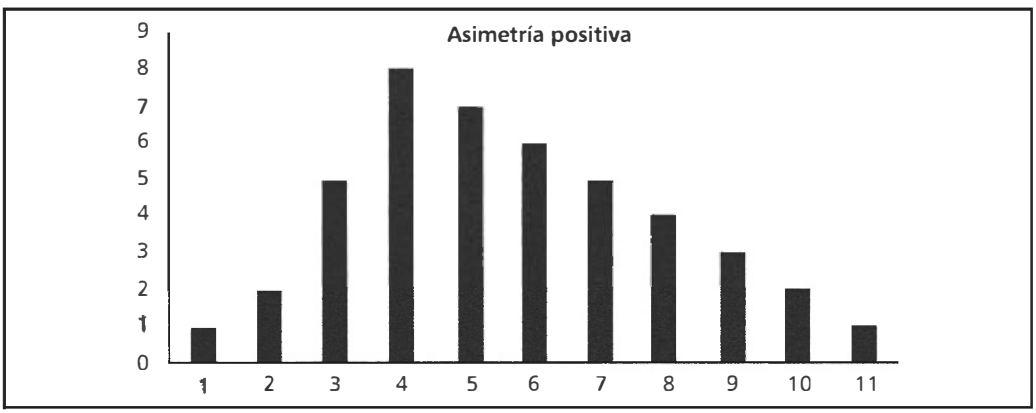


Gráfico 2.2. Ejemplo de una distribución asimétrica a la derecha

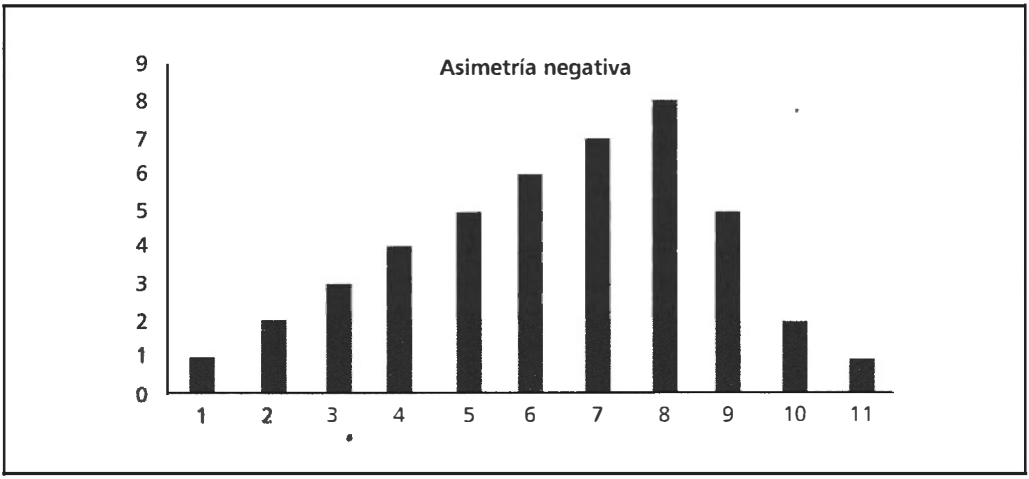


Gráfico 2.3. Ejemplo de una distribución asimétrica a la izquierda

Para cuantificar la asimetría en **distribuciones campaniformes y unimodales,** como las de los gráficos anteriores, utilizaremos **el coeficiente de asimetría de Pearson, A_p,** que se define como

$$A_p = \frac{\bar{x} - M_o}{S_x}$$

Cuando la distribución es campaniforme y simétrica este coeficiente es cero, ya que, media y moda coinciden. Por lo tanto, si es distinto de cero, la distribución no es simétrica; indicando el coeficiente que la distribución es asimétrica positiva o a la derecha cuando es positivo y asimétrica negativa o a la izquierda en caso contrario.

Ejemplo 2.19. Represente gráficamente la distribución salarial del grupo de trabajadores de la empresa ubicada en el país B y calcule su coeficiente de asimetría de Pearson.

Tabla 2.26. Distribución del salario mensual en el país B

Salario (10^3 $)	N.° de trabajadores
8	5
10	10
11	15
13	30
15	20
18	16
20	4
	100

Solución

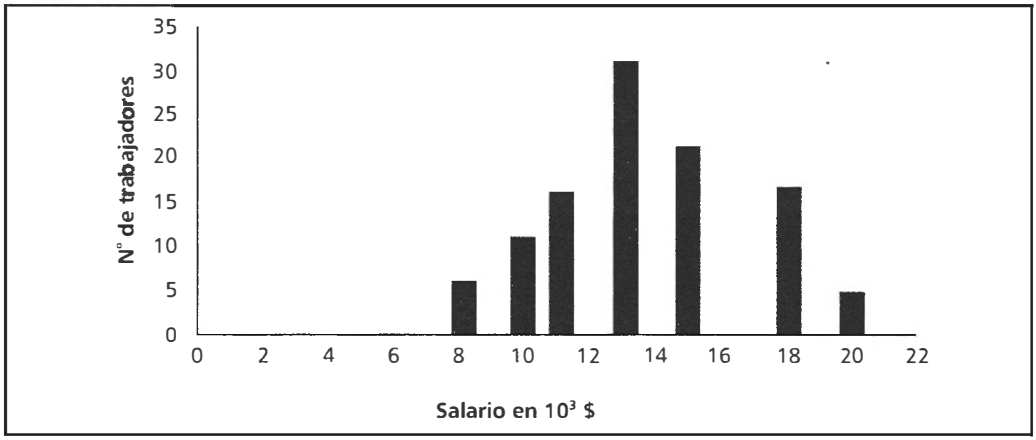


Gráfico 2.4. Diagrama de barras de la distribución salarial

A la vista del gráfico anterior, podemos afirmar que la distribución es campaniforme y parece presentar asimetría positiva.

El coeficiente de asimetría de Pearson resulta

$$A_p = \frac{\bar{x} - M_o}{S_x} = \frac{13{,}63\text{-}13}{3{,}019} = 0{,}21$$

Lo que confirma, teniendo en cuenta la forma, que la distribución es asimétrica positiva o a la derecha.

Para cuantificar la asimetría en una distribución cualquiera, no necesariamente campaniforme, se suele utilizar el **primer coeficiente de Fisher** (γ_1). Este coeficiente está basado en el concepto de momento central o con respecto a la media aritmética, m_x^r, que, para una distribución de frecuencias (x_i, n_i), $i = 1,\ldots,k$, se define como

$$M_{x_y}^r = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^r \cdot n_i}{N}$$

Si la distribución es simétrica, todos los momentos centrales de orden impar son nulos. Como el momento central de orden uno es siempre nulo, el primer momento central de orden impar que puede proporcionar información sobre la no simetría de la distribución es el de orden tres. Si dicho momento es distinto de cero, la distribución no es simétrica. Basándose en esta idea, Fisher define el coeficiente de asimetría (g_1) del siguiente modo

$$g_1 = \frac{\dfrac{\sum_{i=1}^{k} (x_i - \bar{x})^3 \cdot n_i}{N}}{S_x^3} = \frac{m_x^3}{S_x^3}$$

La interpretación de este coeficiente es similar a la del ya estudiado de Pearson. Si el coeficiente es positivo, la distribución es asimétrica a la derecha y si es negativo, asi-

métrica a la izquierda. Si es cero, en general, no podemos afirmar que la distribución sea simétrica.

Ejemplo 2.20. A partir de la información que figura en la Tabla 2.27, analice, gráfica y numéricamente, la asimetría de la distribución del volumen de negocio de las empresas concursadas en España en el primer trimestre de 2011.

Tabla 2.27. Empresas concursadas en España. Primer trimestre de 2011

Volumen de negocio (10^6 €)	N.° de empresas concursadas
Menos de 2	987
2 a 5	241
5 a 10	113
	N = 1.342

Fuente: Elaboración propia a partir de la *Estadística del Procedimiento Concursal*, INE (2012)

Solución. Al tratarse de una distribución de datos agrupados en intervalos, la representación de frecuencias más adecuada es el histograma, teniendo en cuenta que, al ser intervalos de amplitud variable, cada rectángulo debe tener como altura el cociente entre la frecuencia y la amplitud que corresponden al intervalo que figura en su base. En la Tabla 2.28 se presentan las alturas necesarias para representar el histograma y en el Gráfico 2.5 el histograma correspondiente a la distribución del volumen de negocio de las empresas concursadas.

Tabla 2.28. Empresas concursadas en España. Primer trimestre de 2011

Volumen de negocio (10^6 €)	N.° de empresas concursadas	Amplitudes (a_i)	Alturas (h_i)
Menos de 2	987	2	493,500
2 a 5	241	3	80,667
5 a 10	113	5	22,600
TOTAL	**N = 1.342**		

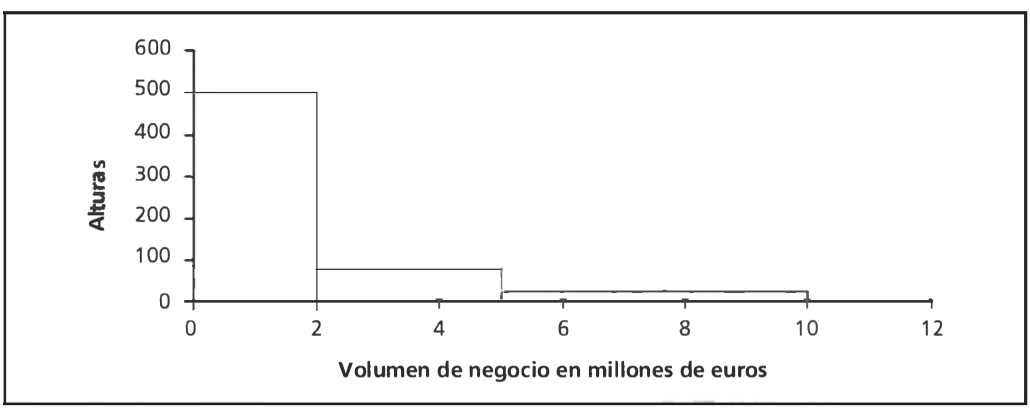


Gráfico 2.5. Volumen de negocios de las empresas concursadas

A la vista del gráfico podemos concluir que, la distribución es unimodal y fuertemente asimétrica a la derecha. Podría verse como campaniforme (si imaginamos el polígono de frecuencias) o con forma de L (histograma). Para el análisis numérico de la asimetría podríamos utilizar cualquiera de los coeficientes estudiados. Calcularemos el primer coeficiente de Fisher **(g_1).** En la Tabla 2.29 se muestran los cálculos necesarios para obtener esta medida.

Tabla 2.29. Empresas concursadas en España. Primer trimestre de 2011

Volumen de negocio (10^6 €)	n_i	x_i	$n_i \cdot x_i$	$n_i^2 \cdot x_i$	$(x_i - \bar{x})^3 \cdot n_i$
Menos de 2	987	1	987	987	–981,494
2 a 5	241	3,5	847	2.964,5	819,797
5 a 10	113	7,5	847,5	6.356,25	18.819,485
TOTAL	**N = 1.342**		**2.681,5**	**10.307,75**	**18.657,788**

A partir de la información proporcionada en la Tabla 2.29, obtenemos que

$$g_1 = \frac{m_x^3}{S_x^3} = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^3 \cdot n_i}{N}}{S_x^3} = \frac{\dfrac{18.657,788}{1.342}}{\left(\dfrac{10.307,75}{1345} - 1,998^2\right)^{\frac{3}{2}}} = 1,9623$$

El valor de dicho coeficiente indica que, como ya se había observado antes, la distribución es asimétrica a la derecha.

2.3.2. Medidas de apuntamiento

Al grado de apuntamiento, o deformación en sentido vertical, de la distribución de frecuencias con respecto a una distribución tipo se le denomina **curtosis**. La distribución que se toma como referencia para cuantificar la curtosis es la distribución normal, que tiene forma de campana y es simétrica. La comparación de la distribución de frecuencias observadas con la distribución normal, permite su clasificación en uno de los siguientes tipos:

- **Distribución platicúrtica:** menos apuntada que la distribución normal.
- **Distribución mesocúrtica:** igual de apuntada que la distribución normal.
- **Distribución leptocúrtica:** más apuntada que la distribución normal.

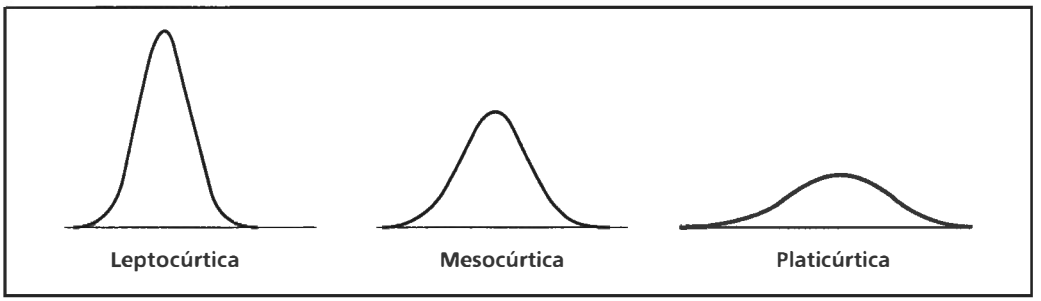


Gráfico 2.6. Distribuciones con distinto grado de apuntamiento

Uno de los coeficientes más utilizados para cuantificar la curtosis de una distribución es el segundo coeficiente de Fisher (g_2), que se define como:

$$\gamma_2 = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^4 \cdot n_i}{N}}{S_x^4} - 3 = \frac{m_x^4}{S_x^4} - 3$$

Teniendo en cuenta que el valor de este coeficiente para la distribución normal es cero, si el valor del mismo para una distribución de frecuencias campaniforme es positivo, diremos que la distribución es leptocúrtica; si es cero, mesocúrtica y si es negativo, platicúrtica.

Ejemplo 2.21. En la Tabla 2.30 se presenta la distribución de las Sociedades Anónimas Cotizadas en mercados de valores bursátiles españoles en el año 2010, atendiendo al número de consejeros con que cuentan. Calcule el coeficiente de curtosis de Fisher y comente el resultado obtenido.

Tabla 2.30. Número de consejeros en Sociedades Anónimas Cotizadas

Número de consejeros	N.° de Sociedades Anónimas Cotizadas
0 – 5	6
5 – 9	42
9 – 13	62
13 – 16	27
16 - 19	9
Total	**146**

Fuente: Elaboración propia a partir de las Estadísticas de Gobierno Corporativo de las Sociedades Cotizadas. Comisión Nacional del Mercado de Valores (2012)

Solución. Tal y como se puede observar en el Gráfico 2.7, la distribución es campaniforme y moderadamente asimétrica a la derecha, por lo que tiene sentido utilizar el coeficiente pedido para comparar su apuntamiento con el que presenta la distribución normal.

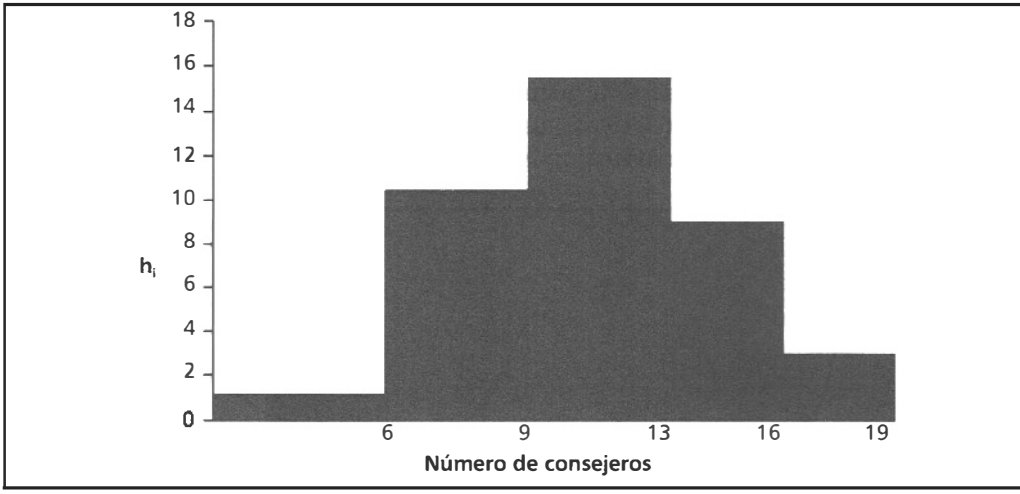


Gráfico 2.7. Histograma de la distribución del número de consejeros

Los cálculos necesarios para obtener el valor del coeficiente de curtosis se presentan en la siguiente tabla.

Tabla 2.31. Sociedades Anónimas Cotizadas (ampliación)

N.º Consejeros	N.º Sociedades	x_i	$x_i \cdot n_i$	x_i^2	$x_i^2 \cdot n_i$	$(x_i - x)^4 \cdot n_i$
0-5	6	2,5	15	6,25	37,5	25.170,468
5-9	42	7	294	49	2.058	6.655,136
9-13	62	11	682	121	7.502	2,589
13-16	27	14,5	391,5	210,25	5.676,75	6.586,514
16-19	9	17,5	157,5	306,25	2.756,25	21.023,027
TOTALES	**146**		**1.540**		**18.030,5**	**59.437,734**

A partir de los resultados presentados en la tabla anterior, se obtienen la media y la varianza del número de consejeros para el conjunto de sociedades considerado

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{1.540}{146} = 10{,}548 \text{ consejeros}$$

$$S_x^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2 = \frac{18.030{,}5}{146} - 10{,}548^2 = 12{,}237 \text{ consejeros}^2$$

Ahora, teniendo en cuenta la fórmula del coeficiente de curtosis y la última columna de la tabla anterior, obtenemos que

$$g_2 = \frac{\dfrac{\sum_{i=1}^{k} (x_i - \bar{x})^4 \cdot n_i}{N}}{S_x^4} - 3 = \frac{\dfrac{59.437{,}734}{146}}{12{,}237^2} - 3 = -0{,}281$$

lo que indica que la distribución es menos apuntada que la normal (platicúrtica).

2.4. Medidas de concentración

Las medidas de concentración tratan de poner de relieve el mayor o menor grado de igualdad en el reparto del total de los valores de la variable entre los elementos de una población; es decir, se utilizan como indicadores del grado de equidistribución en la variable.

Por ejemplo, si la variable es el ingreso de un grupo de familias, diremos que hay poca concentración en la distribución del ingreso si el total del ingreso está repartido de modo que todas las familias ingresan más o menos lo mismo. El caso extremo en este sentido sería que cada familia hubiese ingresado la misma cantidad, esto es

$$\frac{\text{Total ingreso}}{\text{N.º de familias}}$$

Este es el caso extremo denominado **equidistribución.**

Si, por el contrario, hay un grupo de esas familias (pocas en relación al resto) que contribuyen con una alta participación en el total del ingreso, mientras que las restantes, que son la mayoría, contribuyen con una pequeña parte, diremos que hay mucha concentración o que la distribución del reparto del ingreso es muy desigual. El caso extremo ahora sería que una única familia hubiese ingresado el total y que las restantes no hubiesen ingresado nada. Este es el caso de concentración máxima.

Entre las dos situaciones extremas anteriormente descritas es evidente que hay una amplia gama de posibilidades. El objetivo de las medidas de concentración es medir la proximidad o lejanía de una distribución dada a una de esas situaciones extremas.

La **medida de concentración o desigualdad** más utilizada es el **índice de Gini,** que está basado en el análisis gráfico de la desigualdad mediante la **curva de Lorenz**.

Curva de Lorenz

Dada una distribución de frecuencias de una variable (x_i,n_i), $i=1,\ldots,k$, para datos no agrupados o $(L_{i-1}\text{-}L_i, n_i)$, $i=1,\ldots,k$, para datos agrupados, llamamos curva de Lorenz a la poligonal que resulta de unir con trazos rectos los puntos (P_i,Q_i), $i=0,\ldots,k$, definidos en la siguiente expresión

$$P_0 = Q_0 = 0$$

$$P_i = \frac{N_i}{N} \cdot 100 \text{ y } Q_i = \frac{\sum_{j=1}^{i} x_j n_j}{\sum_{j=1}^{k} x_j n_j} \cdot 100, \text{ para } i = 1,\ldots,k$$

Donde x_i representa a un valor de la variable (datos no agrupados) o a la marca de clase del correspondiente intervalo (datos agrupados).

Ejemplo 2.22. El gasto mensual en prensa escrita entre los adultos de una cierta localidad costera es el que figura en la siguiente tabla. Represente gráficamente la curva de Lorenz de dicha distribución.

Tabla 2.32. Distribución del gasto mensual en prensa

Gasto (€)	N.° de adultos
0-9	723
9-18	2.037
18-27	1.445
27-36	1.857
36-60	2.224
60-101	1.714
	10.000

Para calcular las coordenadas de los puntos que determinan la curva, procedemos del siguiente modo:

- Obtenemos las marcas de clase (x_i), que son necesarias para estimar el gasto total (V).
- Calculamos el total que corresponde a cada intervalo, que es $(x_i \cdot n_i)$.
- Acumulando los totales parciales (a los que denominaremos V_i) y dividiendo por el total de la variable **(V)**, se obtiene el valor de Q_i

$$Q_i = \frac{\sum_{j=1}^{i} x_j \cdot n_j}{\sum_{j=1}^{k} x_j \cdot n_j} \cdot 100 = \frac{V_i}{V} \cdot 100$$

- Por último, calculamos las frecuencias acumuladas (N_i) y, a partir de ellas, los porcentajes acumulados de población que corresponden a cada intervalo

$$P_i = \frac{N_i}{N} \cdot 100$$

El resultado de esos cálculos para la distribución que se presenta en la Tabla 2.32 figura en la siguiente tabla.

Tabla 2.33. Distribución del gasto mensual en prensa (ampliación)

Gasto (€)	N.º de adultos	x_i	$x_i \cdot n_i$	$\sum_{j=1}^{i} x_j n_j = V_i$	$\frac{V_i}{V} \cdot 100 = Q_i$	N_i	$\frac{N_i}{N} \cdot 100 = P_i$
0-9	723	4,5	3.253,50	3.253,50	0,89	723	7,23
9-18	2.037	13,5	27.499,50	30.753,00	8,39	2.760	27,6
18-27	1.445	22,5	32.512,50	63.265,50	17,26	4.205	42,05
27-36	1.857	31,5	58.495,50	121.761,00	33,22	6.062	60,62
36-60	2.224	48	106.752,00	228.513,00	62,35	8.286	82,86
60-101	1.714	80,5	137.977,00	366.490,00	100,00	10.000	100
TOTAL	**10.000**		**366.490,00**				

La curva de Lorenz que resulta es la que se muestra en el siguiente gráfico.

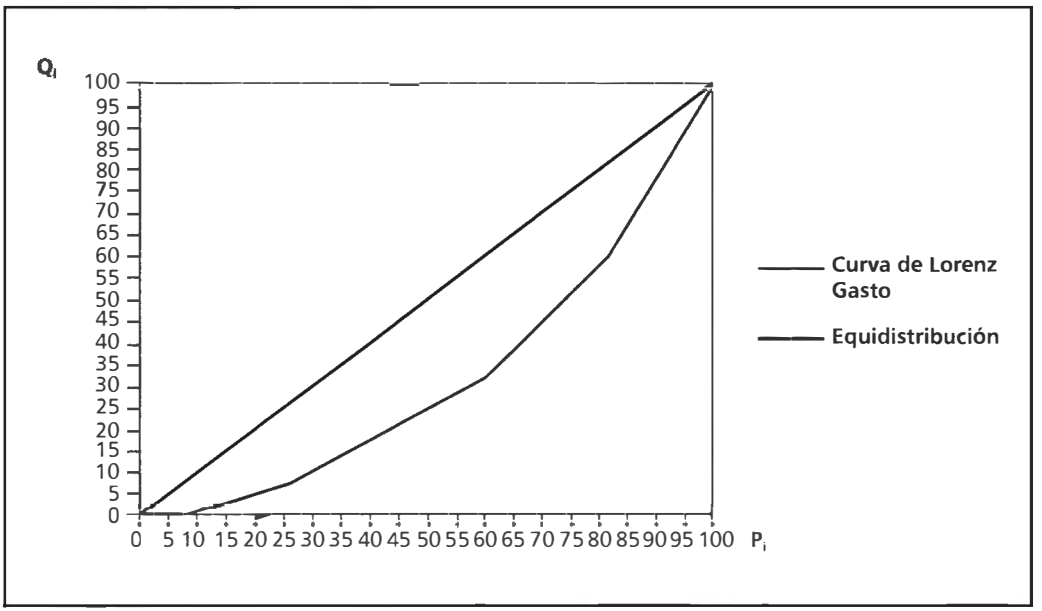


Gráfico 2.8. Curva de Lorenz del gasto mensual

La curva es siempre creciente y está por debajo de la diagonal del cuadrado de lado 100, a la que se denomina **línea de equidistribución** y es la curva de Lorenz de una distribución igualitaria.

La máxima concentración en el reparto de ese gasto total (**366.490 €**) entre los 10.000 adultos se obtendría si la distribución de la variable gasto fuese la que se presenta en la siguiente tabla.

Tabla 2.34. Distribución del gasto mensual en prensa (máxima concentración)

Gasto (€)	N.° de adultos
0	9.999
366.490	1
	10.000

En ese caso, la tabla con los valores de $\mathbf{P_i}$ y $\mathbf{Q_i}$ que permiten determinar la curva de Lorenz sería:

Tabla 2.35. Distribución del gasto mensual en prensa concentración máxima (ampliación)

Gasto (€)	N.° de adultos	N_i	P_i	$x_i \cdot n_i$	v_i	Q_i
0	9.999	9.999	99,99	0	0	0
366.490	1	10.000	100	336.490	100	100
	10.000					

Y la curva o poligonal de Lorenz sería la que figura en el gráfico siguiente.

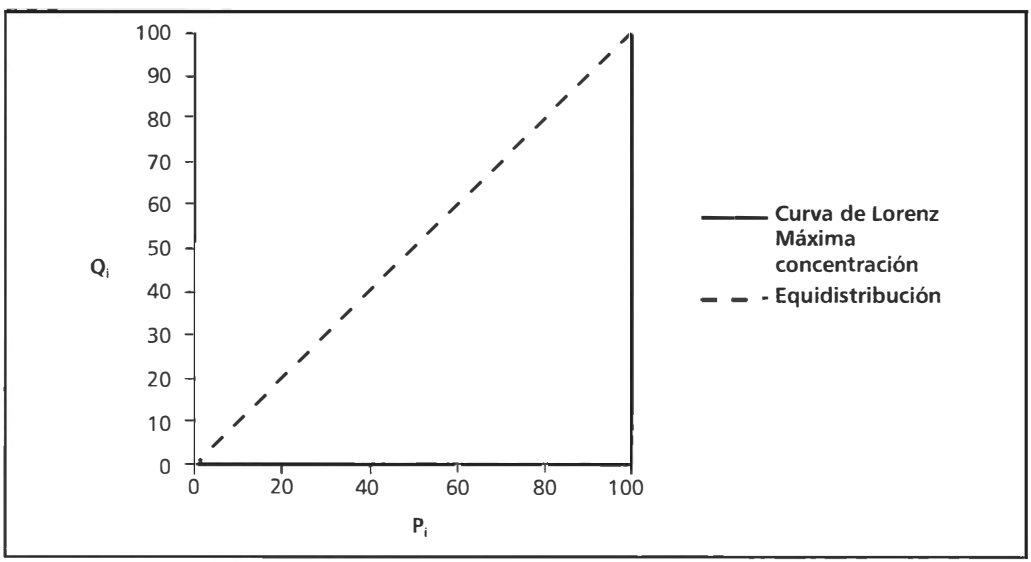


Gráfico 2.9. Curva de Lorenz para la Tabla 2.35 (máxima concentración)

El caso extremo, lados inferior y derecho del cuadrado, es el caso límite correspondiente a **concentración máxima** en población infinita.

El índice de Gini

El área de la región limitada por la diagonal del cuadrado y la curva (poligonal) de Lorenz recibe el nombre de **área de Lorenz.** La proporción que la misma representa del área del triángulo delimitado por las líneas de equidistribución y de concentración máxima es una medida analítica de la concentración (o de desigualdad en el reparto del total) que recibe el nombre de **índice de Gini.**

Teniendo en cuenta que el valor mínimo del área de Lorenz es cero (caso de equidistribución) y que el mayor valor es precisamente el área del triangulo, se concluye que **el índice de Gini (G) toma valores entre cero y uno,** indicando su proximidad a esos extremos menor o mayor nivel de desigualdad en el reparto del total.

$$\begin{cases} G = 0 & \textbf{Equidistribución} \\ G = 1 & \textbf{Máxima concentración} \end{cases}$$

La fórmula que utilizamos para el cálculo del índice de Gini es:

$$G = \frac{\sum_{i=1}^{k-1}\left(P_i \cdot Q_{i+1} - Q_i \cdot P_{i+1}\right)}{10.000}$$

Ejemplo 2.23. Obtener el índice de Gini de la distribución del gasto en prensa escrita de los adultos de la localidad a la que nos hemos referido en el ejemplo anterior.

Teniendo en cuenta la fórmula anterior, la forma de disponer los cálculos sería la que se muestra en la siguiente tabla y, en consecuencia, el valor del índice de Gini para esta distribución resulta G=0,36.

Tabla 2.36. Distribución del gasto mensual en prensa (ampliación)

$\frac{V_i}{V} \cdot 100 = Q_i$	$\frac{N_i}{N} \cdot 100 = P_i$	$P_i\ Q_{i+1}$	$Q_i\ P_{i+1}$
0,89	7,23	60,66	24,56
8,39	27,60	476,38	352,80
17,26	42,05	1.396,90	1.046,30
33,22	60,62	3.779,66	2.752,61
62,35	82,86	8.286,00	6.235,00
100,00	100,00		
		$T_1 = 13.999,59$	$T_2 = 10.411,27$

2.5. Transformaciones lineales de las variables

Sea **X** una variable estadística con distribución de frecuencias (x_i,n_i), $i=1,\ldots,k$. Diremos que otra variable **Y** es una transformación lineal de **X** si la distribución de frecuencias de **Y** es (y_i,n_i), $i=1,\ldots,k$, con $y_i = a + b \cdot x_i$ para algún par de números **a** y **b.**

Los promedios y las medidas de posición, que hemos estudiado en el punto primero de este tema, son valores de la variable y, por lo tanto, una transformación lineal en la variable les afecta en la misma medida que a esta, ya que las frecuencias que corresponden a cada valor y a su transformado son las mismas. El efecto de una de estas transformaciones sobre la media lo hemos visto desglosado en el comportamiento de esta frente a cambios de origen y cambios de escala, propiedades 4 y 5. En la siguiente tabla queda recogido lo anterior.

Tabla 2.37. Efecto de una transformación lineal sobre los promedios y medidas de posición

	Media	Mediana	Moda	Cuantiles (b>0)
$X, (x_i, n_i)$	$\bar{x}$	Me_x	Mo_x	Q_{jx}, D_{jx} o P_{jx}
$Y, (y_i, n_i)$ $y_i = a + bx_i$	$\bar{y} = a + b\bar{x}$	$Me_y = a + bMe_x$	$Mo_y = a + bMo_x$	$Q_{jy} = a + bQ_{jx}$ $D_{jy} = a + bD_{jx}$ $P_{jy} = a + bP_{jx}$

La varianza y la desviación típica se basan en las desviaciones entre las observaciones y su media, $x_i - \bar{x}$. Por lo tanto, la suma de constantes (o cambio de origen) no les afecta, ya que el de la observación con el del promedio se compensan, $x_i + a - (\bar{x} + a) = x_i - \bar{x}$. En cuanto a la multiplicación por constantes, como $b \cdot x_i - b \cdot \bar{x} = b \cdot (x_i - \bar{x})$, la varianza y la desviación típica sí se ven afectadas por este tipo de cambios.

A los recorridos, que hemos definido como diferencia entre dos valores de la variable, les pasa lo mismo. El cambio de origen no les afecta, ya que en la diferencia se compensa, pero el cambio de escala les afecta.

El coeficiente de variación, que hemos definido a partir de la desviación típica y la media, no se ve afectado por los cambios de escala y sí por los de origen.

El coeficiente de variación, cuando se ve afectado, se recalcula teniendo en cuenta las variaciones que experimentan la media y la desviación con el cambio. Eso mismo ocurre con la medida de concentración estudiada, el índice de Gini. Este índice no se ve alterado por un cambio de escala, ya que en su definición la variable interviene a través de los porcentajes del total; sin embargo, el cambio de origen sí le afecta, pero en ese caso hay que recalcularlo.

Las relaciones entre las medidas de dispersión absolutas de una variable y de su transformada son las que figuran en la siguiente tabla.

Tabla 2.38. Efecto de una transformación lineal sobre las medidas de dispersión

	Desviación	Varianza	Recorridos (b > 0)
$X, (x_i, n_i)$	S_x	S_x^2	R_x, $R_{Q,x}$, $R_{D,x}$ o $R_{C,x}$
$Y, (y_i, n_i)$ $y_i = a + bx_i$	$S_y = \lvert b \rvert S_x$	$S_y^2 = b^2 S_x$	$R_y = b\,R_x$ $R_{Q,y} = bR_{Q,x}$ $R_{D,y} = bR_{D,x}$ $R_{C,y} = bR_{C,x}$

Ejemplo 2.24. Suponga que en el año 2009 el salario medio mensual de los trabajadores de una empresa fue de 1.350 euros con una desviación típica de 250 euros. Para ese mismo año, el salario mediano de dichos trabajadores fue de 1.100 euros y el índice de Gini correspondiente a la distribución del salario fue de 0,35. Indique, razonadamente, cómo afectaría a dichas medidas un incremento salarial mensual del 4%. ¿Qué puede decir en relación a la variación que experimentarían las mismas si, en lugar del incremento anterior, se decide aplicar una subida lineal de 50 euros a todos los empleados? Justifique sus respuestas.

2.6. Valores atípicos y diagrama de caja

Los valores atípicos son los que se alejan significativamente del resto de las observaciones.

En una distribución de frecuencias de una variable consideraremos **atípicos** a aquellos valores inferiores al primer cuartil menos 1,5 veces el recorrido intercuartílico ($x_i < Q_1 - 1{,}5 \cdot R_Q$) o superiores al tercer cuartil más 1,5 veces dicho recorrido ($x_i > Q_3 + 1{,}5 \cdot R_Q$). A los valores $Q_1 - 1{,}5 \cdot R_Q$ y $Q_3 + 1{,}5 \cdot R_Q$ se les llama límites admisibles y se les representa con **LI** y **LS,** respectivamente. El intervalo **[LI, LS]** es el **intervalo de valores admisibles** y mide 4 veces el recorrido intercuartílico. Los valores que quedan fuera del mismo son los que consideraremos atípicos. Uno de los gráficos más utilizados para la detección de valores atípicos es el **diagrama de caja,** que, además, resulta bastante útil para tener una primera impresión sobre la forma de la distribución.

Un **diagrama de caja** (o de caja y bigotes) es una representación gráfica de los cuartiles y de los límites admisibles de la distribución. Concretamente, el diagrama consiste en una caja cuyos lados verticales se sitúan sobre los cuartiles primero y tercero (Q_1 y Q_3). Dentro de la caja se traza un segmento vertical que la atraviesa y que está situado sobre la mediana (Q_2). Además, se trazan dos segmentos horizontales (bigotes) que parten de los extremos de la caja y llegan hasta el menor y el mayor valor observado que sea admisible, respectivamente (Gráfico 2.10).

La caja contiene al 50% de las observaciones centrales y la longitud de su base es el recorrido intercuartílico. Los valores atípicos son los que están fuera de los segmentos horizontales. De la observación del diagrama de caja representado en el Gráfico 2.10 concluimos que por la izquierda no hay valores atípicos, ya que el bigote llega hasta el menor valor observado. Sin embargo, por la derecha encontramos dos valores atípicos (b y x_M). Ese bigote llega hasta el mayor valor observado no atípico (a).

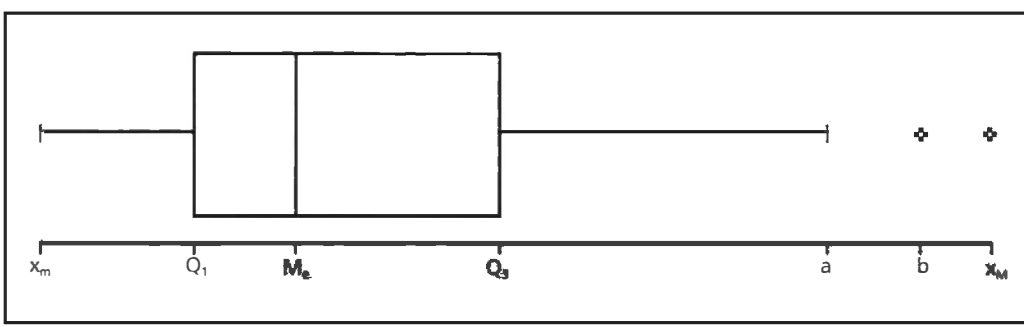


Gráfico 2.10. Diagrama de caja

Para juzgar la simetría o asimetría de un conjunto de datos a partir de este gráfico, se utilizan los siguientes criterios:

- Si la línea de la mediana está en el centro de la caja o cerca del mismo, constituye un indicio de simetría.

- Si la línea que parte de Q_3 es, aproximadamente, de la misma longitud que la que parte de Q_1 también es un indicio de simetría.
- Si la línea de la mediana está más cerca de Q_1 que del centro de la caja, es un indicio de que los datos son asimétricos a la derecha.
- Si la línea que parte de Q_3 es considerablemente más larga que la que lo hace de Q_1, también es un indicio de asimetría positiva.
- Son indicios de asimetría negativa que la línea de la Mediana esté más cerca de la línea de Q_3 que del centro de la caja y que la línea que parte de Q_3 sea considerablemente más corta que la que lo hace de Q_1.

Ejemplo 2.25. En la siguiente tabla se presenta la distribución por comunidad autónoma del número de empresas en España en 2009. Represente el diagrama de caja de la distribución y analice a partir de él la forma de la distribución, indicando si hay valores atípicos y cuáles son.

Tabla 2.39. Distribución del n.º de empresas a España por CC.AA.

Directorio central de empresas Empresas por CC. AA. (Total grupos CNAE93). Año 2009 Unidades: Número de empresas	
Andalucía	510.072
Aragón	93.283
Asturias (Principado de)	71.853
Balears (Illes)	91.826
Canarias	139.381
Cantabria	39.611
Castilla y León	170.626
Castilla-La Mancha	134.479
Cataluña	619.624
Comunitat Valenciana	362.844
Extremadura	67.181
Galicia	201.263
Madrid (Comunidad de)	511.804
Murcia (Región de)	95.636
Navarra (Comunidad Foral de)	43.282
País Vasco	172.152
Rioja (La)	23.525
Ceuta y Melilla	7.388

Fuente: INE

Para resolver calculamos, a partir de la distribución de los valores ordenados, los cuartiles: Q_1, Q_2 y Q_3.

La tabla con los valores ordenados es la que se presenta a continuación:

Empresas por CC. AA. (Total grupos CNAE93). Año 2009 Unidades: Número de empresas	
Ceuta y Melilla	7.388
Rioja (La)	23.525
Cantabria	39.611
Navarra (Comunidad Foral de)	43.282
Extremadura	**67.181**
Asturias (Principado de)	71.853
Balears (Illes)	91.826
Aragón	93.283
Murcia (Región de)	**95.636**
Castilla-La Mancha	**134.479**
Canarias	139.381
Castilla y León	170.626
País Vasco	172.152
Galicia	**201.263**
Comunitat Valenciana	362.844
Andalucía	510.072
Madrid (Comunidad de)	511.804
Cataluña	619.624

Fuente: INE

Teniendo en cuenta que disponemos de 18 observaciones, el primer cuartil ocupa el lugar 5, esto es, Q_1=67.181; la mediana es la media de las observaciones que ocupan los lugares nueve y diez, en este caso 115.057,5, y el tercer cuartil la que ocupa el lugar catorce, 201.263. Basándonos en estos valores representamos el diagrama de caja, teniendo en cuenta que el límite admisible inferior **(LI)** es negativo y, por lo tanto, no hay valores atípicos por la izquierda, y que el límite admisible superior **(LS)** resulta ser 402.386. Por lo tanto, son atípicos los valores correspondientes a Andalucía, Cataluña y Madrid. El gráfico de caja y bigotes es el que se muestra a continuación.

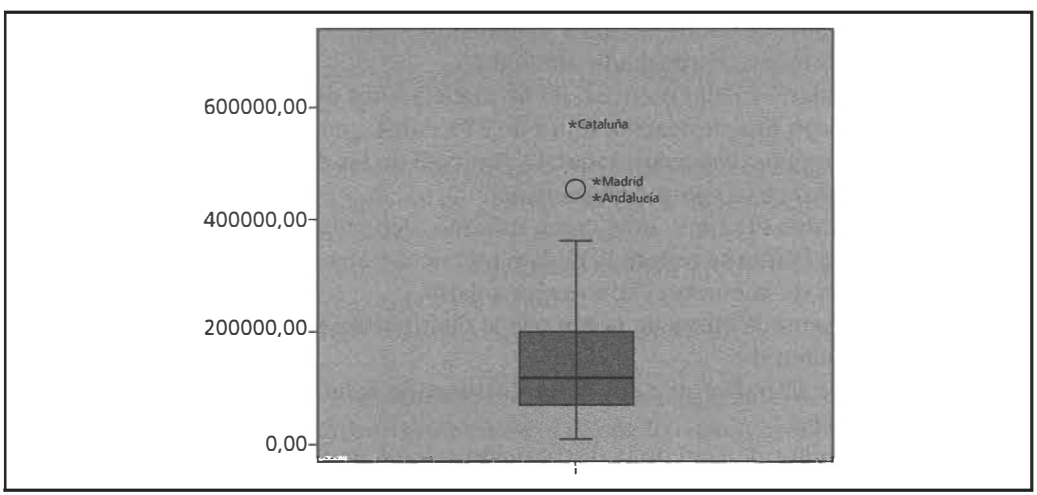


Gráfico 2.11. Diagrama de caja y bigotes para el número de empresas por comunidad autónoma

2.7. Ejercicios

Ejercicio 2.1. En la Tabla 2.40 figuran las 900 empresas de una región clasificadas según el número de empleados con que cuentan.

Tabla 2.40. Clasificación de empresa por n.º de empleados

N.º de empleados	N.º de empresas
40-60	50
60-80	200
80-90	300
90-100	100
100-160	250

a) Determine el número medio de empleados de las empresas de la región.
b) ¿Cuál es el porcentaje de empresas que cuentan con un número de empleados inferior al medio?
c) Si el grupo A lo forman el 25% de las empresas con mayor número de empleados, ¿qué número de empleados debe tener como mínimo una empresa para pertenecer al grupo A?
d) Represente gráficamente la distribución y, teniendo en cuenta la forma de la misma, proporcione una medida de asimetría de la distribución e interprete el resultado.

Ejercicio 2.2. En la Tabla 2.41 figuran los salarios mensuales, en euros, de los 900 trabajadores de una determinada empresa.

Tabla 2.41. Rango de salarios de una determinada empresa

Salario (€)	N.º de trabajadores
400-600	50
600-800	200
800-900	300
900-1.000	100
1.000-1.600	250

Se pide:

a) Determine el salario medio y proporcione una medida de su representatividad. Comente el resultado obtenido.
b) Si el salario medio mensual de los trabajadores de otra empresa es de 1.150 euros con una desviación típica de 175 euros, ¿cuál de las dos empresas presenta menor dispersión salarial? ¿En cuál de las dos empresas es más representativo el salario medio mensual?
c) Determine el salario que, como mínimo, debe tener un trabajador de la empresa a la que se refiere la tabla para encontrarse entre el 10% de los trabajadores de la empresa con mayor salario.
d) Represente la curva de Lorenz de la distribución salarial y comente el resultado obtenido.
e) Calcule el índice de Gini de la distribución salarial y comente el resultado obtenido.
f) Si el índice de Gini de la distribución salarial de los empleados de la empresa a la que nos hemos referido en el apartado b) es 0,65, ¿en cuál de las dos empresas hay mayor desigualdad en el reparto de la masa salarial?

Ejercicio 2.3. Se desea estudiar la capacidad hotelera de una importante ciudad turística de España a través del número de habitaciones de los hoteles de la ciudad. Para ello se dispone de la información contenida en la Tabla 2.42:

Tabla 2.42. Capacidad hotelera de una de las principales ciudades turísticas en España

Capacidad (n.° de habitaciones)	Hoteles
0-50	7
50-100	3
100-150	5
150-250	3
250 y más	2

Sabiendo, además, que la capacidad media es de 113,8 habitaciones por hotel, se pide:

a) Represente las distribuciones de frecuencias absolutas y absolutas acumuladas. Comente los gráficos obtenidos.
b) Determine el número de hoteles con capacidad superior a la media.
c) ¿Cuántas habitaciones tienen, como máximo, los hoteles de menor capacidad que representan el 50% del total de hoteles de la ciudad?
d) Represente la curva o poligonal de Lorenz y comente el gráfico obtenido.
e) Calcule el índice de Gini y comente el resultado obtenido.
f) ¿Cuántas habitaciones tienen, como máximo, los hoteles de menor capacidad que cubren el 20% del total de habitaciones de la ciudad?

Ejercicio 2.4. En la Tabla 2.43 se muestra la distribución porcentual de los habitantes de una pequeña localidad, atendiendo a su presupuesto mensual, expresado en euros, para un determinado año A.

Tabla 2.43. Distribución porcentual de los habitantes de una localidad (presupuesto mensual)

Presupuesto (€)	% habitantes
0-300	7,16
300-600	23,10
600-900	32,74
900-1.200	20,41
más de 1.200	16,59

Se pide:

a) En el año A, el presupuesto medio mensual entre los habitantes de esa localidad fue de 842,96 euros. ¿Qué porcentaje de los habitantes de la misma tuvieron un presupuesto inferior al medio?
b) Calcule una medida de la dispersión relativa en la distribución del presupuesto mensual y comente el resultado.
c) ¿A cuánto ascendió, en dicha localidad, el presupuesto mensual mediano en el año A?
d) ¿Cuál debió ser, como mínimo, el presupuesto mensual de un habitante que en el año A se encontraba entre el 25% de los habitantes de dicho lugar con mayor presupuesto mensual?
e) Analice la concentración en el reparto del total del presupuesto mensual entre los habitantes de la localidad.

Ejercicio 2.5. Supongamos que, en el año en curso, el salario medio mensual de los trabajadores de una empresa es de 1.275 euros con una desviación típica de 300 euros. Además, los salarios modal y mediano son de 800 y 975 euros, respectivamente. Se pide:

a) Si para el próximo año se prevé una subida salarial del 7%, determine las principales características de la distribución salarial, bajo el supuesto de que la plantilla sea la misma y que no haya promociones.
b) Si en lugar de la subida a la que nos hemos referido en el apartado anterior, la dirección de la empresa decide subir a todos sus empleados 200 euros, ¿cuáles serán las principales características de la nueva distribución salarial?
c) ¿Cuál de las dos subidas conduce a una situación de menor dispersión salarial?
d) Compare la concentración en la distribución salarial de este año con la correspondiente a la que resulta tras una subida salarial del 7%.

Ejercicio 2.6. De una compañía aérea A se sabe que, en el último mes, el 20% de sus vuelos no presentaron retraso y que, durante este periodo, 200 vuelos de la compañía han registrado retraso.

El análisis descriptivo, obtenido con el software R, correspondiente a la distribución de la variable RETRASO_A = Retraso en minutos de estos 200 vuelos es el que se presenta a continuación:

```
     RETRASO_A
  Min.    : 3.00
  1st Qu.:20.00
  Median :31.00
  Mean    :31.04
  3rd Qu.:44.00
  Max.    :59.00
  n       :200
  sd      :16.32779
  var     :266.5968
skewness :0.03586367
kurtosis :-1.180364
```

Para la compañía A,

a) ¿Cuál es el retraso medio de los vuelos con retraso? ¿Y del total de vuelos de la compañía?
b) Si el último vuelo realizado se encuentra entre el 25% de los vuelos con menos retraso de la distribución de los vuelos con retraso, ¿cuál es el máximo retraso que puede haber presentado?
c) Si el último vuelo realizado se encuentra entre el 20% de los vuelos con más retraso de la distribución de los vuelos totales, ¿cuál es el mínimo retraso que puede haber presentado?

Otra compañía aérea, B, presenta los siguientes resultados referidos a sus vuelos con retraso:

```
     RETRASO_B
  Min.    : 5.00
  1st Qu.:18.00
  Median :51.00
  Mean    :43.94
  3rd Qu.:64.00
  Max.    :75.00
  n       :80
  sd      :22.02659
  var     :485.1707
skewness :-1.035863
kurtosis :-1.008031
```

d) ¿En qué compañía, A o B, los retrasos presentan más dispersión relativa?
e) Indique en qué compañía un vuelo con 15 minutos de retraso presenta peor posición relativa.
f) Realice los diagramas de caja *(box-plot)* correspondientes a las distribuciones RETRASO_A y RETRASO_B.
g) Comente el tipo de asimetría y curtosis que presentan las distribuciones RETRASO_A y RETRASO_B.
h) Si tras una política de incentivos en la compañía A el retraso de los vuelos ha experimentado una reducción del 5%, ¿cuál será la nueva media aritmética de los vuelos con retraso y su desviación estándar? y ¿cuál será la media del total de vuelos?

Ejercicio 2.7. Las siguientes estadísticas son relativas a los precios por habitación doble y noche en sendos grupos de hoteles pertenecientes a los municipios de Torremolinos y Marbella en una determinada semana (Tabla 2.44).

Tabla 2.44. Tabla de precios por habitación doble y noche en Torremolinos y Marbella

Precio habitación doble/noche (€)	Hoteles en Torremolinos (%)
0-40	14.3
40-55	
55-65	28.6
65-150	
Total	

Precio habitación doble/noche (€)	Hoteles en Marbella
0-65	8
65-100	6
100-150	
Más de 150	3
Total	21

Fuente: www.booking.com y elaboración propia

Además, se sabe que:

— El 47,6% de los hoteles del grupo de Torremolinos tienen un precio por habitación doble inferior a 55 euros.
— El precio medio por habitación doble en el grupo de hoteles de Marbella es de 88,3 euros.

— La desviación típica del precio por habitación doble en el grupo de hoteles de Marbella es de 56,87 euros.

Se pide:

a) Indique la población, los elementos y el tipo de variable analizada que se deducen de la tabla estadística de Marbella.

b) Complete las dos tablas estadísticas.

c) Calcule el precio medio por noche de una habitación doble en el grupo de hoteles de Torremolinos.

d) ¿En cuál de los dos grupos de hoteles es más representativo el correspondiente precio medio?

e) Un turista tiene la posibilidad de reservar una habitación doble en un hotel del grupo de Torremolinos a un precio de 70 euros/noche o reservarla en otro del grupo de Marbella al precio de 85 euros/noche. ¿Qué hotel le ofrece un mejor precio relativo? Justifique su respuesta.

f) Los empresarios del grupo de Marbella han decidido incrementar sus precios, pero dudan entre efectuar una subida constante de 15 euros/noche en habitación doble o realizar un incremento de un 2,5% en el precio. ¿Cuál de las dos medidas conduce a una menor dispersión de los precios? Justifíquelo numéricamente.

CAPÍTULO 3
Análisis conjunto de dos variables

En este tema abordamos el análisis estadístico de un conjunto de datos relativos a dos variables que han sido observadas, simultáneamente, sobre los distintos elementos de una población. El objetivo básico del tema es proporcionar técnicas que permitan cuantificar el grado de relación o dependencia existente entre dichas variables y describir el comportamiento conjunto de las mismas, con la finalidad práctica de predecir, en el contexto general del que los datos proceden, el comportamiento de una de ellas (la que se considera dependiente o endógena) a partir de los valores observados de la otra (que se considera independiente o exógena).

3.1. Presentación de los datos

El resultado de la observación simultánea de dos variables sobre un mismo elemento de la población es un par de números. La información relativa a los **N** pares de observaciones se suele disponer en forma de tabla de doble entrada, que recibe el nombre de **tabla de correlación o de contingencia** (Tabla 3.1). Para formar dicha tabla, disponemos en el margen izquierdo de la misma, los distintos valores observados de una de las variables **(X)** y en el margen superior los de la otra **(Y).** Dentro de la tabla, en las casillas que se forman al combinar los valores de ambas variables, figura el número de veces que ha sido observado el par que corresponde a esa casilla.

Tabla 3.1. Distribución conjunta

X \ Y	y_1	y_2	...	y_j	...	y_s	...
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
...	...	...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	**N**

En la tabla anterior hemos representado con x_i, $i = 1,\ldots, r$, a los distintos valores de **X** y con y_j, $j = 1,\ldots, s$, a los distintos valores de **Y**. Además:

- **N** representa el número total de observaciones.
- n_{ij} es la **frecuencia conjunta** del par (x_i, y_j) o número de elementos en los que se ha observado el valor x_i de **X** y, simultáneamente, el valor y_j de **Y**.
- $n_{i.}$ es la **frecuencia marginal** de x_i o número de elementos en los que se ha observado el valor x_i de **X**.
- $n_{.j}$ es la **frecuencia marginal** de y_j o número de veces que se observa el valor y_j de **Y**.

Es claro que las frecuencias marginales se obtienen a partir de las conjuntas mediante

$$n_{i.} = \sum_{j=1}^{s} n_{ij}$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

Las frecuencias conjuntas y marginales, además de las relaciones anteriores, cumplen

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = N$$

Como en el caso de estadísticas de una variable, la frecuencia relativa **(f)** se define como la proporción del total **(N)** que representa cada frecuencia absoluta. Ahora, asociadas a los tres tipos de frecuencias absolutas que aparecen en la tabla de contingencia, tenemos

$$f_{ij} = \frac{n_{ij}}{N};\ f_{i.} = \frac{n_{i.}}{N};\ f_{.j} = \frac{n_{.j}}{N}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{i=1}^{r} f_{i.} = \sum_{j=1}^{s} f_{.j} = 1$$

Multiplicando por 100 las frecuencias relativas, se obtienen los porcentajes: p_{ij}, $p_{i.}$ y $p_{.j}$.

Observación. El tratamiento de los datos procedentes de la observación conjunta de dos atributos o de un atributo y una variable sería similar, con la diferencia de que el resultado de la observación serán pares cuyos componentes no son necesariamente de naturaleza numérica. En ese caso, en el margen (o márgenes) de la tabla correspondiente(s) al (a los) atributo(s) dispondremos las distintas modalidades que han sido observadas.

Por otra parte, en cuanto a las distribuciones bidimensionales en las que interviene alguna variable, hemos de señalar que los valores de la misma pueden venir agrupados en intervalos. En ese caso, en el margen correspondiente aparecerán los distintos intervalos o clases que contienen a las observaciones de dicha variable.

Los distintos valores o intervalos de la variable **X** junto con sus correspondientes frecuencias marginales, $(x_i, n_{i.})$, $i = 1,\ldots, r$, forman una distribución de frecuencias de

una variable: ***la distribución de frecuencias de* X** considerada al margen de la otra variable o con independencia de ella. Esta distribución se llama **distribución marginal de X.** De manera análoga, al considerar los distintos valores de **Y** acompañados de sus correspondientes frecuencias marginales, $(y_j, n_{.j})$, $j = 1,\ldots, s$, se obtiene la ***distribución de frecuencias de* Y** que, en este contexto, recibe el nombre de **distribución marginal de Y.**

Las distribuciones marginales asociadas a la variable bidimensional **(X, Y)** que figura en la Tabla 3.2 son las siguientes:

Tabla 3.2. Distribuciones marginales

X	Frecuencias	Y	Frecuencias
x_1	$n_{1.}$	y_1	$n_{.1}$
x_2	$n_{2.}$	y_2	$n_{.2}$
...	...	...	...
x_i	$n_{i.}$	y_j	$n_{.j}$
...	...	...	...
x_r	$n_{r.}$	y_s	$n_{.s}$

Otras distribuciones de frecuencias de una variable asociadas a una bidimensional son las denominadas **distribuciones condicionadas.**

En una distribución bidimensional, **(X,Y),** llamamos distribuciones condicionadas a las distribuciones de frecuencias de la variable **Y (X)** sobre cada uno de los subconjuntos de elementos en los que ha sido observado cada uno de los distintos valores de **X (Y).**

Concretamente, asociada al valor x_i de la variable **X** tenemos ***la distribución de frecuencias de* Y *condicionada por el valor* x_i *de* X**, $Y/X = x_i$, que, con la notación introducida en la Tabla 3.1, es la que viene dada en la siguiente tabla.

Tabla 3.3. Distribución condicionada de Y por el valor x_i de X

$Y/X=x_i$	$n_{j/i}$
y_1	n_{i1}
y_2	n_{i2}
...	...
y_j	n_{ij}
...	...
y_s	n_{is}
	$n_{i.}$

Observemos que hay una distribución condicionada de **Y** por cada uno de los distintos valores observados de **X.** En nuestro caso, tenemos **«r»** distribuciones condicionadas de **Y.**

De manera análoga se definen las distribuciones condicionadas de **X** por los distintos valores observados de **Y**. Dado el valor y_j de **Y**, ***la distribución condicionada de* X *por el valor* y_j *de* Y**, $X/Y=y_j$, es la que figura en la Tabla 3.4.

Tabla 3.4. Distribución condicionada de X por el valor y_j de Y

$Y/X=y_j$	$n_{j/i}$
x_1	n_{1j}
x_2	n_{2j}
...	...
x_i	n_{ij}
...	...
x_r	n_{rj}
	$n_{.j}$

Cuando, como en el caso que recoge la Tabla 3.1, la variable **Y** presenta «**s**» valores distintos, tenemos «**s**» distribuciones condicionadas de **X,** una por cada uno de los distintos valores de **Y.**

Las distribuciones marginales y condicionadas son distribuciones de una variable y su análisis descriptivo se basa en lo estudiado en el Capítulo 2.

Observación. Cuando se trabaja con pocas observaciones, en lugar de presentar los datos en forma de tabla de doble entrada, se presentan en dos columnas que contienen a todos los pares de valores observados (distintos o no), uno por cada elemento de la población que haya sido observado.

X	Y
x_1	y_1
x_2	y_2
...	...
x_i	y_i
...	...
x_N	y_N

Ejemplo 3.1. En la Tabla 3.5 figura la distribución de frecuencias conjunta del gasto medio semanal y el número de días que acuden mensualmente al supermercado las 1.200 familias de un determinado barrio de la ciudad.

Tabla 3.5. Distribución conjunta

Gasto (€) \ Número de días	5	7	8	9	11	13	
100-150	25	50	75	90	85	75	
150-175	70	25	40	55	45	65	
175-225	50	25	85	75	90	75	
225-300	20	25	15	15	15	10	
							1.200

Distribuciones marginales

Tabla 3.6. Distribución marginal del gasto medio semanal

Gasto medio semanal (€)	N.° de familias
100-150	400
150-175	300
175-225	400
225-300	100
	1.200

Tabla 3.7. Distribución marginal del número de días/mes

Número de días/mes	N.° de familias
5	165
7	125
8	215
9	235
11	235
13	225
	1.200

Algunas distribuciones condicionadas

Tabla 3.8. Distribución del gasto entre las familias que acuden 9 días al mes al supermercado

Gasto (€)/N.º de días/mes = 9 días	N.° de familias	$f_{gasto/y=9}$
100-150	90	90/235=0,383
150-175	55	55/235=0,234
175-225	75	75/235=0,319
225-300	15	15/235=0,064
	235	**1**

Tabla 3.9. Distribución del gasto entre las familias que acuden 7 días al mes al supermercado

Gasto (€)/N.º de días/mes = 7 días	N.° de familias	$f_{gasto/y=7}$
100-150	50	0,4
150-175	25	0,2
175-225	25	0,2
225-300	25	0,2
	125	**1**

3.2. Relaciones entre variables e independencia estadística

Definición. Dos variables **X** e **Y** son estadísticamente independientes si las distribuciones de frecuencias condicionadas satisfacen alguna de las siguientes condiciones:

(i) Fijado $i = 1,\ldots, r$, $f_{j/i} = f_{.j}$, para todo $j=1,\ldots, s$.
(ii) Fijado $j = 1,\ldots, s$, $f_{i/j} = f_{i.}$, para todo $i =1,\ldots, r$.

En el apartado i) afirmamos que ***todas las distribuciones condicionadas de* Y *son iguales y coinciden con la marginal de* Y**. En el apartado ii) afirmamos que ***todas las distribuciones condicionadas de* X *son iguales y coinciden con la marginal de* X.**

Observación. Para establecer la igualdad de las distribuciones condicionadas entre sí y con la marginal correspondiente, es necesario recurrir a las frecuencias relativas o a los porcentajes porque, en general, son distribuciones definidas sobre poblaciones (o conjuntos de elementos) con distinto tamaño.

Es fácil ver que las condiciones i) y ii) de la definición anterior son equivalentes y que a su vez equivalen a la siguiente condición:

(iii) Para cualesquiera i y j con $i=1, \ldots, r$ y $j=1, \ldots, s$, $f_{ij} = f_{i.}\, f_{.j}$

Aunque las condiciones i) y ii) son más intuitivas, esta última condición puede ser, en ocasiones, más fácil de manejar para el análisis de la independencia a partir de la tabla de correlación o contingencia.

Cuando se comprueba que dos variables, observadas conjuntamente sobre una población, son estadísticamente independientes, no tiene sentido un análisis posterior de su distribución conjunta y se analizan solo las distribuciones marginales de cada una de esas variables. En otro caso, es decir, si las variables están relacionadas o son dependientes, estaremos interesados en el análisis de la variación conjunta de esas dos variables, que englobaría el análisis de la covarianza, la correlación y la regresión.

3.3. Asociación entre variables cuantitativas

Las situaciones más comunes de asociación entre dos variables son las siguientes:

1. **Dependencia causal unilateral.** Una de las dos variables **(X)** influye en la otra **(Y),** pero no al revés. En esta situación, **X** recibe el nombre de **variable explicativa, independiente, causa** o **exógena,** mientras que de **Y** diremos que es la **explicada, dependiente, efecto** o **endógena.**
2. **Interdependencia.** La influencia entre las variables es recíproca y se produce en las dos direcciones.
3. **Dependencia indirecta.** Las variables **X** e **Y** presentan una sincronización en su variación debido a una tercera variable **(Z)** que influye en las dos.
4. **Covariación casual o espuria.** Se produce cuando las variables, de las que se sabe que no tienen ningún vínculo, varían de manera sincronizada; es decir, la sincronización es casual o accidental.

El **análisis estadístico de la covariación** o variación conjunta de dos variables no se ocupa de la detección del tipo de covariación existente entre un par de variables relacionadas o dependientes, sino que se limita a cuantificarla a través de medidas de asociación como la **covarianza** y el **coeficiente de correlación lineal simple.** Si, además, hay constancia de que la covarianza responde a una dependencia causal unilateral, es posible representar mediante una forma funcional a dicha covariación, lo que estaría dentro del denominado **análisis de regresión.** Este análisis tiene la finalidad práctica de describir

dicha relación y poder realizar predicciones sobre la variable que se considera dependiente o explicada a partir de valores conocidos para la independiente o explicativa.

El análisis estadístico de la covariación puede comenzar con la representación gráfica del **diagrama de dispersión o nube de puntos** de la distribución bidimensional **(X, Y).** Esta gráfica resulta al representar sobre unos ejes cartesianos los puntos de coordenadas **x** e **y** correspondientes a los diferentes elementos de la población o muestra. El conjunto de puntos que resulta se denomina **nube de puntos** o **diagrama de dispersión**. La inspección visual de dicho gráfico nos proporciona información sobre: si existe o no relación entre las variables; si la relación, en caso de que exista, es lineal o no; si las variables varían en el mismo sentido o no; etc.

En la Figura 3.1 aparecen diferentes gráficos que muestran las nubes de puntos correspondientes a cuatro situaciones distintas. El **Gráfico a)** muestra una relación lineal creciente entre las variables X e Y (valores elevados de Y se asocian a valores elevados de X y valores pequeños de Y se asocian a valores pequeños de X). Decimos que la asociación es lineal creciente o directa porque la nube de puntos se asemeja a una recta con pendiente positiva. El **Gráfico b)** muestra una relación lineal decreciente entre las variables X e Y (valores altos de X están asociados a valores bajos de Y y viceversa). Decimos que la asociación es lineal decreciente o inversa porque la nube de puntos se asemeja a una recta con pendiente negativa. Nótese que en la nube de puntos del **Gráfico b)** se aprecia una mayor dispersión que en la correspondiente al **Gráfico a).** El **Gráfico c)** muestra una relación no lineal (en este ejemplo, la nube de puntos se asemeja a una parábola) y el **Gráfico d)** parece representar ausencia de asociación entre las variables X e Y.

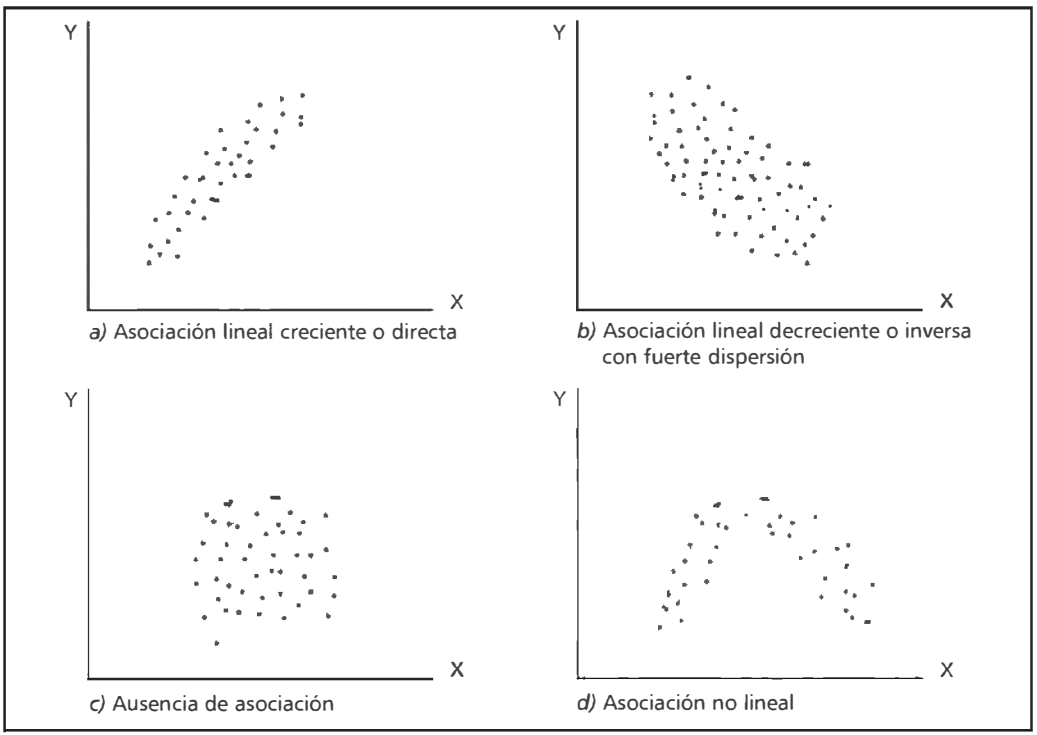


Figura 3.1. Nubes de puntos

Ejemplo 3.2. El número de días al mes que acuden al supermercado y el gasto medio semanal (en €) de 8 familias quedan recogidos en la Tabla 3.10 Represente la nube de puntos y extraiga conclusiones.

Tabla 3.10. Gasto y número de días que acuden al supermercado

Número de días/mes	Gasto medio semanal (€)
3	55
4	54
5	54
5	53
6	53
6	52
7	51
8	50

Solución

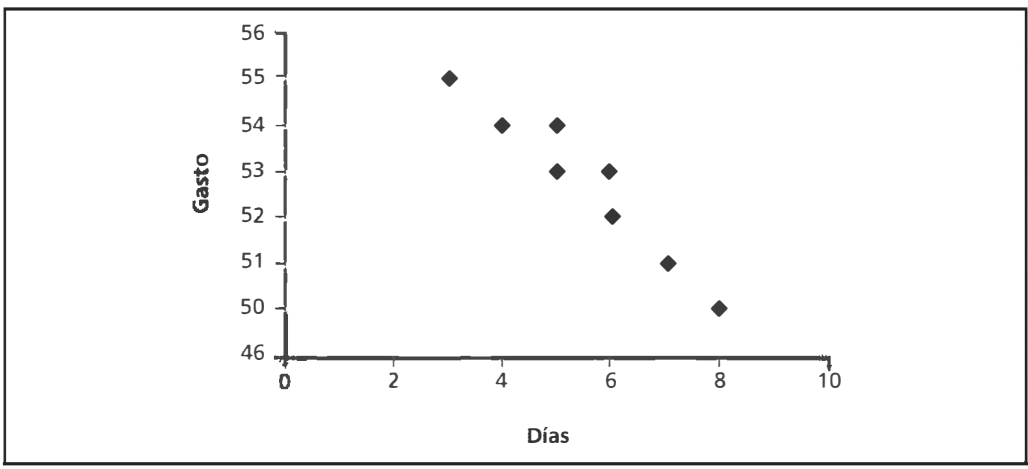


Gráfico 3.1. Diagrama de dispersión del Ejemplo 3.2

Como se puede apreciar en el gráfico anterior, la relación entre ambas variables es lineal y decreciente con poca dispersión.

3.3.1. Medidas de asociación: covarianza y coeficiente de correlación lineal

Las medidas de asociación son medidas numéricas con las que se pretende resumir mediante un número la intensidad de la covariación existente entre las dos variables. La **covarianza** y el **coeficiente de correlación lineal,** basado en ella, son medidas de asociación que ***indican el grado de covariación lineal entre las dos variables***.

Covarianza

La covarianza entre las dos variables **X** e **Y** que forman la bidimensional **(X,Y)** se representa mediante S_{xy} y se define como

$$S_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}$$

siendo **N** el número total observaciones de **(X, Y)**.

La expresión de la covarianza en términos de los distintos valores observados de **X** y de **Y** es

$$S_{xy} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - \bar{x}) \cdot (y_j - \bar{y}) \cdot n_{ij}}{N}$$

Es fácil ver que

$$S_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N} = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y}$$

y que

$$S_{xy} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot n_{ij}}{N} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{s} x_i \cdot y_j \cdot n_{ij}}{N} - \bar{x} \cdot \bar{y}$$

Estas dos últimas serán las fórmulas que utilizaremos habitualmente para calcular la covarianza, según que trabajemos con datos de frecuencia unitaria (todos los pares) o no (solo los pares distintos).

Observación. La utilización de la covarianza como medida del grado de relación lineal existente entre dos variables se basa en la siguiente consideración empírica: si en la nube de puntos o gráfico de dispersión trazamos unos ejes cartesianos con origen en el punto $(\bar{x}, \bar{y})$, el producto de las desviaciones $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ será positivo si el punto (x_i, y_i) está situado en los cuadrantes primero o tercero y en otro caso será negativo (véase la Figura 3.2).

Como se puede apreciar en la Figura 3.2, si entre las dos variables existe relación lineal positiva o directa, la mayor parte de los puntos estarán situados en los cuadrantes I y III y la covarianza será, por lo tanto, positiva. Si, por el contrario, la relación es lineal pero inversa o negativa, la mayor parte de los puntos estarán situados en los cuadrantes II y IV y la covarianza será negativa.

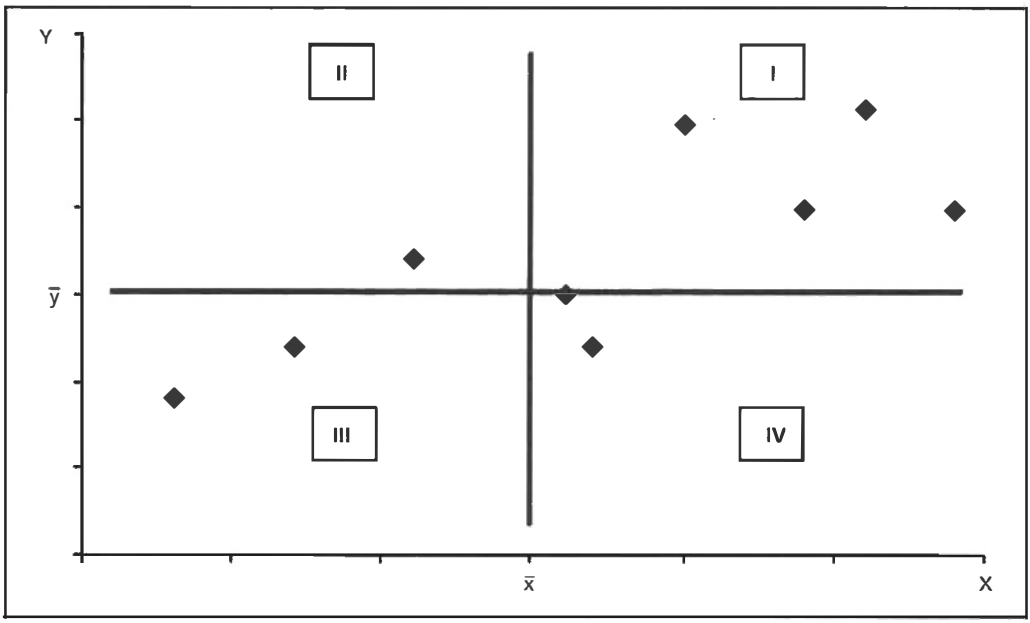


Figura 3.2. Descomposición de la nube de puntos

Por último, si no existe relación lineal entre las variables, los puntos se extenderán aleatoriamente en los cuatro cuadrantes, habrá compensaciones y la covarianza será cero o próxima a cero. Teniendo en cuenta estas consideraciones, ***un valor elevado de la covarianza (positivo o negativo) se considera indicio de relación lineal entre las variables, mientras que un valor pequeño, en términos absolutos, o nulo, refleja la ausencia de este tipo de relación***.

La covarianza viene expresada en el producto de las unidades de medida en las que vengan expresadas las variables **X** e **Y**, por lo tanto, se ve influida por cambios en las unidades de medida y no es apta para realizar comparaciones. Por otra parte, como su valor absoluto no está acotado superiormente, un valor no nulo de la covarianza no sirve para afirmar si la relación lineal que existe entre las variables es más o menos intensa.

Ejemplo 3.3. Obtenga la covarianza para la distribución conjunta del número de días al mes que acuden al supermercado y el gasto medio semanal (en €) de las 8 familias cuya información queda recogida en la Tabla 3.10.

Solución. Para calcular la covarianza entre Días **(X)** y Gasto **(Y)**, $\mathbf{S_{xy}}$, utilizamos la fórmula

$$\mathbf{S_{xy}} = \frac{\sum_{i=1}^{N} \mathbf{x_i} \cdot \mathbf{y_i}}{\mathbf{N}} - \bar{\mathbf{x}} \cdot \bar{\mathbf{y}}$$

Los cálculos necesarios para la obtención de la covarianza son los que se indican en la Tabla 3.11.

Tabla 3.11. Cálculos intermedios para la covarianza

Número de días/mes x_i	Gasto (€) y_i	$x_i\, y_i$
3	55	165
4	54	216
5	54	270
5	53	265
6	53	318
6	52	312
7	51	357
8	50	400
Totales **44**	**422**	**2.303**

Luego,

$$S_{xy} = \frac{\sum_{i=1}^{8} x_i \cdot y_i}{8} - \bar{x} \cdot \bar{y} = \frac{2303}{8} - \frac{44}{8} \cdot \frac{422}{8} = -2{,}25$$

El valor no nulo de la covarianza indica la existencia de relación lineal entre ambas variables. Si, además, tenemos en cuenta el signo (negativo, en este caso) concluiremos que dicha relación lineal es decreciente o negativa; es decir, a medida que el número de días al mes que acuden al supermercado es mayor, el gasto medio semanal disminuye. La covarianza anterior está expresada en €·días.

Coeficiente de correlación lineal

Si en la fórmula de la covarianza expresamos las desviaciones en unidades de desviación típica, dividiendo por S_x y por S_y, obtenemos el coeficiente de correlación lineal X e Y, es decir, r_{xy}

$$r_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N \cdot S_x \cdot S_y} = \frac{S_{xy}}{S_x \cdot S_y}$$

El coeficiente de correlación lineal es una medida relativa o adimensional del grado de relación lineal existente entre las dos variables y, por lo tanto, apta para comparar la intensidad de la relación lineal entre las variables de dos distribuciones bidimensionales distintas. Además, se puede demostrar que este coeficiente toma valores entre **-1** y **1**, siendo:

- $r_{xy} = 1$ si, y solo si, entre las dos variables existe una relación lineal perfecta y directa o positiva.
- $r_{xy} = -1$ si, y solo si, entre las dos variables existe una relación lineal perfecta e inversa o negativa.
- Si $r_{xy} = 0$, y solo si, entre las dos variables no existe relación lineal.

Ejemplo 3.4. Obtenga el coeficiente de correlación lineal entre el gasto medio semanal y el número de días al mes que acuden al supermercado correspondiente al grupo de familias cuya información se recoge en la Tabla 3.10.

Solución. Teniendo en cuenta la definición del coeficiente de correlación lineal,

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

además de lo ya calculado para la covarianza, es necesario calcular las desviaciones típicas de cada una de las variables. La tabla con la información necesaria para su cálculo es la que se muestra a continuación.

Tabla 3.12. Cálculos intermedios para el coeficiente de correlación lineal

Número de días/mes x_i	Gasto (€) y_i	$x_i\, y_i$	x_i^2	y_i^2
3	55	165	9	3.025
4	54	216	16	2.916
5	54	270	25	2.916
5	53	265	25	2.809
6	53	318	36	2.809
6	52	312	36	2.704
7	51	357	49	2.601
8	50	400	64	2.500
Totales **44**	**422**	**2.303**	**260**	**22.280**

$$S_x^2 = \frac{\sum_{i=1}^{8} x_i^2}{8} - \bar{x}^2 = \frac{260}{8} - \left(\frac{44}{8}\right)^2 = 2{,}25$$

$$S_y^2 = \frac{\sum_{i=1}^{8} y_i^2}{8} - \bar{y}^2 = \frac{22.280}{8} - \left(\frac{422}{8}\right)^2 = 2{,}4375$$

Por lo tanto, el coeficiente de correlación lineal, r_{xy}, resulta:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-2{,}25}{\sqrt{2{,}25} \cdot \sqrt{2{,}4375}} = -0{,}96$$

El valor del coeficiente de correlación lineal, que es muy próximo a 1 en valor absoluto, indica una fuerte relación lineal entre las variables. El signo negativo del mismo indica que la relación es decreciente; es decir, a medida que el número mensual de visitas al supermercado aumenta, el gasto medio semanal disminuye.

Observación. El hecho de que el coeficiente de correlación lineal sea cero o próximo a cero indica ausencia de relación lineal, no que las variables sean independientes. En el siguiente ejemplo se muestra una distribución de dos variables con covarianza nula y que, sin embargo, no son independientes.

Ejemplo 3.5. Comprobar que las variables **X** e **Y**, cuya distribución conjunta es la que figura en la Tabla 3.13, no son independientes y, sin embargo, tienen covarianza nula.

Tabla 3.13. Distribución conjunta del Ejemplo 3.5

X \ Y	-1	0	1	
-1	1	2	1	
0	2	1	1	
1	1	2	1	
				12

Solución. Para comprobar que las dos variables no son independientes, calcularemos dos distribuciones condicionadas de **X** y comprobaremos que no son iguales.

Tabla 3.14. Distribución condicionada de X/y = 0

X/y=0	$n_{x/y=0}$	$f_{x/y=0}$
-1	2	0,4
0	1	0,2
1	2	0,4
		1

Tabla 3.15. Distribución condicionada de X/y = –1

X/y=–1	$n_{x/y=-1}$	$f_{x/y=-1}$
–1	1	0,25
0	2	0,5
1	1	0,25
	4	1

Estas dos distribuciones condicionadas de **X** no son iguales, luego **las variables son dependientes.**

Para calcular la covarianza entre las dos variables, S_{xy}, utilizamos la fórmula

$$S_{xy} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} x_i \cdot y_j \cdot n_{ij}}{12} - \bar{x} \cdot \bar{y}$$

La distribución marginal de **X** viene dada por

Tabla 3.16. Distribución marginal de X

x_i	n_i
–1	4
0	0
1	4
	12

Para calcular la media de **X**, $\bar{x}$, añadimos una columna adicional que contenga los productos $x_i\ n_i$ (Tabla 3.17).

Tabla 3.17. Distribución marginal de X ampliada

x_i	n_i	$x_i n_i$
–1	4	–4
0	4	0
1	4	4
	12	**0**

De donde resulta que $\bar{x} = 0$ y, por lo tanto, que el producto $\bar{x} \cdot \bar{y} = 0$.

Para calcular la media de los productos, completamos la tabla original con los productos correspondientes a cada casilla (Tabla 3.18).

Tabla 3.18. Distribución conjunta de X e Y ampliada

X \ Y	–1	0	1
–1	1(1)	2(0)	1(-1)
0	2(0)	1(0)	1(0)
1	1(-1)	2(0)	1(1)
			12(0)

La media de los productos resulta al sumar todos los números que aparecen entre paréntesis y dividir por el total de observaciones (en nuestro caso, 12), esto es

$$\frac{\sum_{i=1}^{3}\sum_{j=1}^{3} x_i \cdot y_j \cdot n_{ij}}{12} = \frac{1 + 0 + (-1) + 0 + 0 + 0 + (-1) + 0 + 1}{12} = 0$$

Luego la covarianza es nula.

Ejemplo 3.6. Calcule la covarianza y el coeficiente de correlación lineal entre las variables X = «gasto medio semanal» e Y= «número mensual de días que acuden al supermercado» cuya distribución conjunta figura en la Tabla 3.19.

Tabla 3.19. Distribución conjunta de X e Y

Visitas mensuales (días) / Gasto (€)	5	7	8	9	11	13
100-150	25	50	75	90	85	75
150-175	70	25	40	55	45	65
175-225	50	25	85	75	90	75
225-300	20	25	15	15	15	10

A partir de las distribuciones marginales de **X** e **Y**, obtenemos las tablas 3.20 y 3.21 que se muestran a continuación.

Tabla 3.20. Distribución marginal de X (ampliación)

Gasto (€)	N.° de familias	x_i	$x_i\, n_i$
100-150	400	125	50.000
150-175	300	162,5	48.750
175-225	400	200	80.000
225-300	100	262,5	26.250
Totales	**1.200**		**205.000**

Luego

$$\bar{x} = \frac{205.000}{1.200} = 170{,}83 \text{ euros}$$

Tabla 3.21. Distribución marginal de Y (ampliación)

Visitas mensuales (días)	N.° de familias	$y_j\, n_j$
5	165	825
7	125	875
8	215	1.720
9	235	2.115
11	235	2.585
13	225	2.925
Totales	**1.200**	**11.045**

Luego

$$\bar{y} = \frac{11.045}{1.200} = 9{,}204 \text{ días/mes}$$

Para calcular la media de los productos, procedemos como en el ejemplo anterior, pero haciendo uso de las marcas de clase para la variable **X**.

Tabla 3.22. Distribución conjunta de X e Y (ampliación)

X \ Y	5	7	8	9	11	13
125	25(15.625)	50(43.750)	75(75.000)	90(101.250)	85(116.875)	75(121.875)
162,5	70(56.875)	25(28.437,5)	40(52.000)	55(80.437,5)	45(80.437,5)	65(137.312,5)
200	50(50.000)	25(35.000)	85(136.000)	75(135.000)	90(198.000)	75(195.000)
262,5	20(26.250)	25(45.937,5)	15(31.500)	15(35.437,5)	15(43.312,5)	10(34.125)
Totales	148.750	153.125	294.500	352.125	438.625	488.312,5

Luego la covarianza entre las dos variables resulta

$$S_{xy} = \frac{1}{1.200}\sum_{i=1}^{4}\sum_{j=1}^{6} x_i \cdot y_j - \bar{x}\cdot\bar{y} =$$

$$= \frac{1}{1.200}(148.750 + 153.125 + 294.500 + 352.125 + 438.625 + 488.312,5) -$$

$$-170,83 \cdot 9,204 = \frac{1.875.437,5}{1.200} - 1.572,32 = -9,46$$

Al ser distinta de cero, podemos concluir que existe relación lineal entre las variables y, teniendo en cuenta que el signo es negativo, diremos que la misma es inversa o decreciente; es decir, al aumentar el número mensual de visitas al supermercado, disminuye el gasto medio semanal.

Para cuantificar el grado de relación lineal existente entre las variables, calculamos el coeficiente de correlación lineal. Para ello, son necesarias las desviaciones típicas de ambas variables, que vamos a calcular a continuación (Tablas 3.23 y 3.24).

Tabla 3.23. Distribución marginal de X (ampliación)

Gasto (€)	N.° de familias	x_i	$x_i\, n_i$	$x^2_i\, n_i$
100-150	400	125	50.000	6.250.000
150-175	300	162,5	48.750	7.921.875
175-225	400	200	80.000	16.000.000
225-300	100	262,5	26.250	6.890.625
	1.200		**205.000**	**37.062.500**

$$S_x^2 = \frac{\sum_{i=1}^{4} x_i^2 \cdot n_i}{1.200} - \bar{x}^2 = \frac{37.062.500}{1.200} - 170,83^2 = 1.702,5278$$

$$S_x = \sqrt{1.702,5278} = 41,26$$

Tabla 3.24. Distribución marginal de Y (ampliación)

Número mensual de visitas (días)	N.° de familias	$y_j\, n_j$	$y^2_j\, n_j$
5	165	825	4.125
7	125	875	6.125
8	215	1.720	13.760
9	235	2.115	19.035
11	235	2.585	28.435
13	225	2.925	38.025
Totales	**1.200**	**11.045**	**109.505**

$$S_y^2 = \frac{\sum_{j=1}^{6} y_j^2 \cdot n_j}{1.200} - \bar{y}^2 = \frac{109.505}{1.200} - 9{,}204^2 = 6{,}5406$$

$$S_y = \sqrt{6{,}5406} = 2{,}55755$$

Luego el coeficiente de correlación lineal, r_{xy}, resulta

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{-9{,}46}{41{,}26 \cdot 2{,}56} = -0{,}0896$$

De dicho valor, además de lo que se concluyó a partir de la covarianza, se deduce que la relación lineal existente entre las variables es muy débil, prácticamente inexistente, ya que el valor del coeficiente es muy próximo a cero.

3.3.2. Regresión lineal: estimación, bondad de ajuste y predicción

La regresión tiene por objeto expresar mediante una forma funcional la relación de dependencia causal que se presenta entre dos variables estadísticas **X** e **Y,** asumiendo para una de ellas el papel de independiente, exógena o explicativa **(X)** y para la otra **(Y)** el papel de dependiente, endógena o explicada. El propósito es describir el tipo de relación que existe entre las variables y poder hacer predicciones para la variable explicada a partir de observaciones de la explicativa.

La regresión lineal, en particular, consiste en expresar la relación existente entre las variables **X** e **Y** mediante una recta.

Concretamente, la recta de regresión que explica **Y** a partir de **X**

$$y_i^* = a + bx_i$$

es, de entre todas las rectas del plano, la que mejor se adapta a la nube de puntos (véase la Figura 3a) según el criterio de los mínimos cuadrados; es decir, aquella que hace mínima a la suma $\sum_{i=1}^{N} e_i^2$, siendo e_i los errores cometidos al comparar, para cada x_i, los valores observados, y_i, con los valores que proporciona la recta, y_i^*

$$e_i = y_i - y_i^* = y_i - (a + bx_i),\ i = 1,\ldots, N$$

La imposición de este criterio (mínimos cuadrados ordinarios) permite determinar los parámetros **a** y **b** de la recta de regresión a partir de los datos, obteniéndose

$$\text{i)}\ a = \bar{y} - b\bar{x}$$

$$\text{ii)}\ b = \frac{S_{xy}}{S_x^2}$$

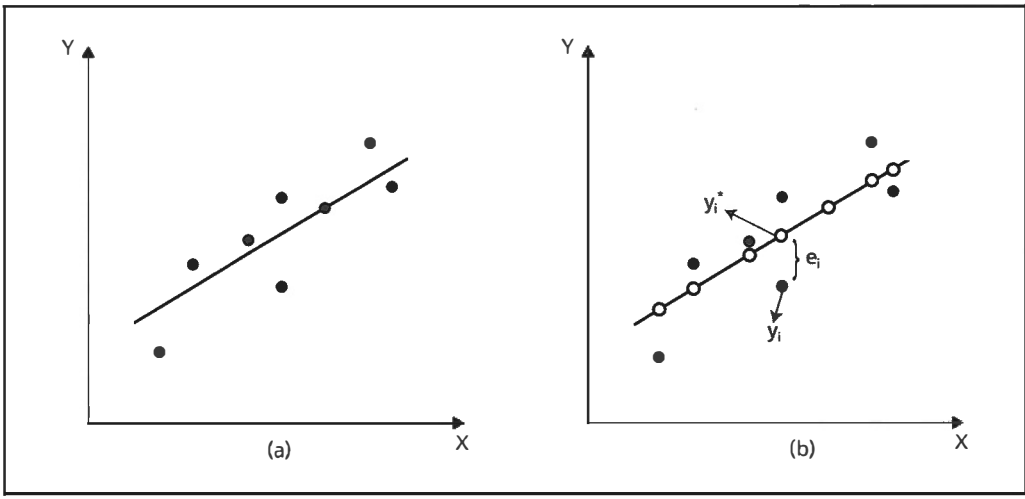


Figura 3.3. Nube de puntos y recta de regresión

Interpretación de los parámetros a y b

- **a:** se le denomina **término independiente** y es la **ordenada en el origen de la recta de regresión;** es decir, el punto de corte de la recta con el eje de ordenadas. Representa **el valor estimado para la variable explicada (Y) cuando la explicativa (X) toma el valor cero.** Tiene la misma unidad de medida que Y y puede tener sentido o no en el contexto del enunciado.

- **b:** se le denomina **coeficiente de regresión** y es la **pendiente de la recta de regresión**. Representa **el crecimiento (b > 0) o disminución (b < 0) que experimenta el valor estimado para la variable explicada (Y) por un crecimiento de una unidad en la variable explicativa (X).** El signo indica si la relación lineal es creciente o decreciente. Si no hay relación lineal entre las variables, ya vimos que $S_{xy} = 0$, por lo que, **b = 0.** En este caso particular, la recta de regresión es horizontal. La unidad de medida de la pendiente viene condicionada por las unidades de medida de las variables X e Y.

Algunas propiedades descriptivas del método de mínimos cuadrados ordinarios

- La recta de regresión que explica Y a partir de X pasa por el punto $(\bar{x}, \bar{y})$. Esta propiedad está implícita en la condición i) expuesta anteriormente.

- La suma de los errores es siempre cero

$$\sum_{i=1}^{N} e_i = 0$$

- Teniendo en cuenta la propiedad anterior, es fácil deducir que la suma de los valores estimados de Y coincide con la suma de los valores observados de dicha variable. Esto es

$$\sum_{i=1}^{N} y_i^* = \sum_{i=1}^{N} y_i$$

Ejemplo 3.7. Especifique y estime, a partir de los datos de la Tabla 3.10, un modelo de regresión lineal para explicar el gasto medio semanal en función del número mensual de visitas al supermercado.

Solución. La variable explicativa **(X)** debe ser el número mensual de visitas al supermercado, expresado en días. La variable explicada **(Y)** será el gasto medio semanal, expresado en euros.

El modelo lineal que especificamos es

$$y_i = a + bx_i + e_i$$

Para estimar los valores de los parámetros **a** y **b** a partir de los datos utilizamos las expresiones i) y ii)

$$\text{i)}\ \ a = \bar{y} - b\bar{x}$$

$$\text{ii)}\ \ b = \frac{S_{xy}}{S_x^2}$$

En el ejercicio anterior hemos calculado los valores de $\bar{x}$, $\bar{y}$, S_{xy} y S_x^2, siendo

$$\bar{x} = \frac{44}{8}\ \text{días}$$

$$\bar{y} = \frac{422}{8}\ \text{euros}$$

$$S_{xy} = -2{,}25\ \text{euros} \cdot \text{días}$$

$$S_x^2 = 2{,}25\ \text{días}^2$$

Luego

$$b = \frac{-2{,}25}{2{,}25} = -1\ \text{euros/días}$$

$$a = \frac{422}{8} - (-1) \cdot \frac{44}{8} = 58{,}25\ \text{euros}$$

Por lo tanto, el modelo lineal estimado a partir de los datos es

$$y_i = y_i^* + e_i, \text{ siendo } y_i^* = 58{,}25 - x_i$$

El valor del coeficiente de regresión **(–1)** indica la disminución media que experimenta el gasto medio semanal por un incremento unitario en el número mensual de visitas

al supermercado; es decir, por cada día en que se incremente el número mensual de visitas al supermercado, el gasto medio semanal disminuye, por término medio, un euro.

El valor del término independiente **(58,25 euros)** indica el valor estimado para el gasto medio semanal cuando no se acude al supermercado.

En la Tabla 3.25 se presentan los valores estimados de Y (Gasto) a través de la recta de regresión. Por comparación con los valores observados, podemos obtener los errores que figuran, también, en dicha tabla.

Tabla 3.25. Valores observados de X e Y, valores estimados de Y y errores

Número de días/mes x_i	Gasto (€) y_i	y^*_i	$e_i=y_i-y^*_i$
3	55	55,25	-0,25
4	54	54,25	-0,25
5	54	53,25	0,75
5	53	53,25	-0,25
6	53	52,25	0,75
6	52	52,25	-0,25
7	51	51,25	-0,25
8	50	50,25	-0,25
44	**422**	**422**	**0**

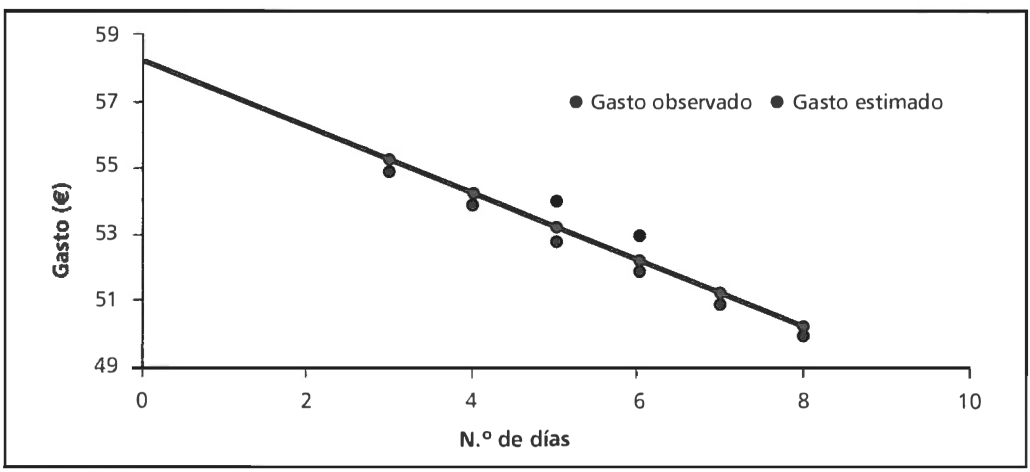


Gráfico 3.2. Nube de puntos y recta ajustada del Ejemplo 3.7

Bondad del ajuste y predicción

Una vez estimada la recta de regresión, necesitamos una medida que nos describa hasta qué punto el ajuste realizado es aceptable o no. Nótese que la recta trata de ajustarse a una determinada nube de puntos, pero si esta no toma una forma similar a una línea recta, o bien su dispersión es muy elevada, el ajuste derivado de la línea de regresión no será aceptable.

¿Hasta qué punto podemos considerar bueno el ajuste de una línea recta? Una respuesta a esta cuestión se puede obtener a partir del cálculo del **coeficiente de determinación.** Para deducir su fórmula, utilizaremos la descomposición de la varianza de la variable Y.

El criterio de mínimos cuadrados que se utiliza para estimar la recta de regresión permite la siguiente descomposición de la varianza de la variable dependiente del siguiente modo:

$$S_y^2 = S_{y^*}^2 + S_e^2$$

Es decir, la variabilidad total de la variable dependiente, medida a través de su varianza, se descompone en la variabilidad explicada por la independiente a través del modelo y la variabilidad residual o parte de la variabilidad de la dependiente que no viene explicada por el modelo. Basándonos en esta descomposición, podemos afirmar que el modelo será tanto mejor, cuanto mayor sea la variabilidad explicada $S_{y^*}^2$ o, equivalentemente, cuanto menor sea la variabilidad no explicada por el mismo (S_e^2).

Ambas cantidades presentan el inconveniente de no ser adimensionales y de carecer de una cota superior independiente de los datos, que nos permita decir si un ajuste es mejor o peor que otro a partir de la comparación de las correspondientes varianzas. Sin embargo, si en la expresión de la descomposición dividimos en ambos miembros por S_y^2 resulta la expresión

$$1 = \frac{S_{y^*}^2}{S_y^2} + \frac{S_e^2}{S_y^2}$$

El primer sumando, $\frac{S_{y^*}^2}{S_y^2}$, recibe el nombre de **coeficiente de determinación** del ajuste y se representa por R^2

$$R^2 = \frac{S_{y^*}^2}{S_y^2}$$

Este coeficiente tiene las siguientes propiedades:

- **Es adimensional,** al ser cociente de varianzas que vienen en la misma unidad. Representa la proporción (o el porcentaje, en caso de que multipliquemos por 100) de las variaciones de Y que vienen explicadas por X a través del modelo lineal estimado.
- De la igualdad anterior se concluye que $0 \leq R^2 \leq 1$.
- El coeficiente de determinación R^2 correspondiente a una regresión lineal está relacionado con el coeficiente de correlación lineal mediante la igualdad $R^2 = r_{xy}^2$.

El ajuste será tanto mejor cuanto más próximo a 1 esté el coeficiente de determinación. Si dicho coeficiente es nulo, el modelo estimado no explica nada.

Cuando la recta de regresión ajustada se considera buena ($\mathbf{R^2}$ suficientemente próximo a **1**), se puede utilizar para realizar predicciones sobre la variable endógena a partir de valores observados de la exógena. Dado un valor de la variable exógena, al que representamos mediante $\mathbf{x_p}$, la predicción de la variable endógena vendrá dada por

$$\mathbf{y_p^* = a + bx_p}$$

Dentro de la predicción distinguimos entre interpolación y extrapolación. Si representamos con $\mathbf{x_m}$ al valor más pequeño de la variable exógena y con $\mathbf{x_M}$ al valor más grande, la predicción que se realiza para un valor, $\mathbf{x_p}$, comprendido entre los dos límites anteriores, recibe el nombre de **interpolación,** mientras que la predicción hecha con un valor de la exógena fuera del recorrido $\mathbf{[x_m - x_M]}$ se denomina **extrapolación.** En el Gráfico 3.3 se muestran los tipos de predicción para el Ejemplo 3.7. En este caso, $\mathbf{x_m = 3}$, $\mathbf{x_M = 8}$ y la línea recta muestra predicciones para valores de la exógena que van desde 0 hasta 9 días. Por tanto, para valores de X entre 3 y 8 hablamos de interpolaciones, y para valores de X fuera de ese intervalo, es decir, para x = 0, 1, 2 y 9, de extrapolaciones.

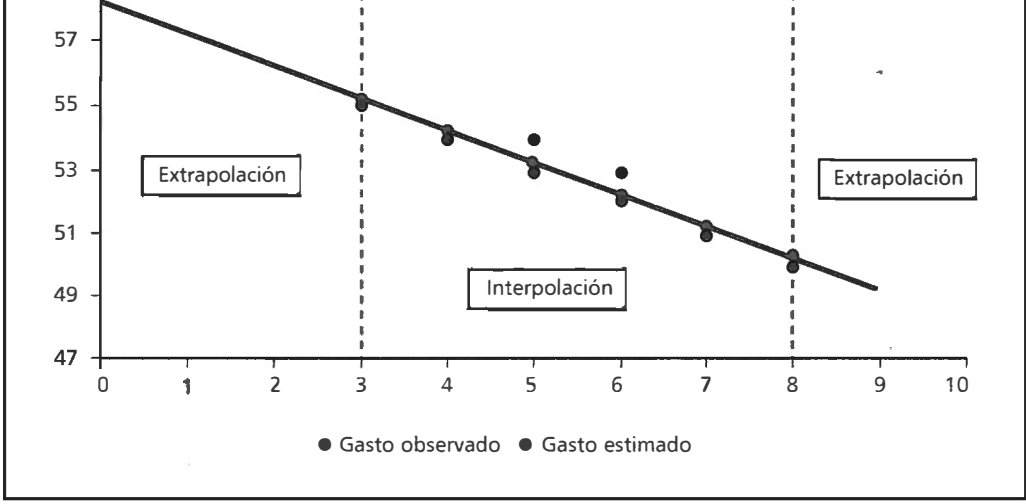


Gráfico 3.3. Tipos de predicción sobre el Ejemplo 3.7

La fiabilidad de la predicción dependerá de:

a) La bondad del ajuste: en igualdad de condiciones, cuanto más elevado sea el coeficiente de determinación, más fiable es la predicción.

b) El tipo de predicción: en igualdad de condiciones, una interporlación siempre resulta más fiable que una extrapolación. En el caso de una extrapolación, cuanto más nos alejamos de la nube de puntos, menos fiable es la predicción, ya que mayor es el riesgo de que la estructura de los datos, fuera del recorrido utilizado para estimar, sea menos parecida a la que se presenta dentro del mismo.

c) El grado de conocimiento de $\mathbf{x_p}$: para realizar la predicción necesitamos conocer el valor de la variable exógena. En ocasiones, dicho valor no es conocido y es necesario fijar uno como posible. Por tanto, la fiabilidad de la predicción, dependerá también, del grado de confianza que depositemos en ese valor fijado para la exógena.

Ejemplo 3.8. Analice la bondad del modelo estimado en el Ejemplo 3.7.

Solución. La bondad del modelo la medimos mediante el coeficiente de determinación **(R^2).**

Para el ajuste lineal, $\mathbf{R^2 = r^2}$, por lo tanto, basándonos en el valor del coeficiente de correlación lineal que se calculó en el Ejemplo 3.4, $\mathbf{R^2 = 0{,}96^2 = 0{,}92}$, que está muy próximo a **1,** lo que indica que el ajuste es bastante bueno. El **92%** de las variaciones en el gasto medio semanal familiar vienen explicadas por las variaciones en el número mensual de visitas al supermercado a través del modelo estimado.

Ejemplo 3.9. Utilice el modelo estimado en el Ejemplo 3.7 para estimar el gasto medio semanal de una familia que a lo largo del mes ha acudido 9 días al supermercado. Comente la fiabilidad de la predicción.

Solución

$$\mathbf{y_i^* = 58{,}25 - 9 = 49{,}25 \text{ euros}}$$

Mediante el modelo ajustado, estimamos que el gasto medio semanal de una familia que a lo largo del mes ha acudido 9 días al supermercado es de 49,25 euros.

Se trata de una extrapolación, para un valor próximo al recorrido de la exógena, a partir de un modelo con un coeficiente de determinación elevado. Por tanto, la predicción puede considerarse bastante fiable.

3.4. Ejercicios

Ejercicio 3.1. La Tabla 3.26 recoge el número de vehículos importados por España durante el año 2010 (expresado en unidades). En la misma se clasifican las importaciones de vehículos según país de procedencia y tipo de vehículo.

Tabla 3.26. Importación de vehículos en España en 2010

Procedencia / Tipos	Turismos	Camiones	Tractores
Alemania	116.693	5.081	11.908
EE.UU.	1.201	145	348
Francia	146.859	6.791	2.080
Italia	43.463	7.011	3.793
Japón	37.952	1.561	1.243
Países Bajos	2.070	220	1.244
Reino Unido	57.896	2.601	554
Suecia	2.364	17	0
Otros países	168.849	21.463	7.427

Fuente: Comercio Exterior, Principales Resultados, INE (2010)

Se pide:

a) Indique cuál es la población a la que la tabla se refiere, cuál es el tamaño poblacional, cuáles son los caracteres analizados y la naturaleza de los mismos.

b) Obtenga la distribución de vehículos importados procedentes de los distintos países sin tener en cuenta la tipología de los mismos. ¿De qué país o grupo de países se importó una mayor cantidad de vehículos en términos porcentuales?

c) Obtenga la distribución porcentual, según procedencia, de los tractores importados.

d) Determine la distribución de frecuencias absolutas conjuntas para los vehículos importados, bajo el supuesto de que la procedencia y el tipo de los mismos fuesen independientes.

Ejercicio 3.2. Una sociedad de valores que opera en la Bolsa de Madrid clasifica a sus inversores considerando dos caracteres: Perfil de riesgo del inversor (X) y Sexo (Y) del mismo. La Tabla 3.27 muestra la distribución porcentual por sexo de los inversores teniendo en cuenta su perfil inversor.

Tabla 3.27. Distribución porcentual del perfil inversor por sexo

X \ Y	Poco arriesgado	Arriesgado	Muy arriesgado
Hombre	82,09	59,46	22,64
Mujer	17,91	40,54	77,36

Se pide:

a) ¿Cómo se denomina a las distribuciones que se presentan en la tabla anterior?

b) Interprete el dato correspondiente a un hombre con perfil inversor muy arriesgado.

c) Sabiendo que la sociedad de valores tiene una cartera de clientes formada por 10.930 inversores de los cuales un 61,30% tienen un perfil de poco arriesgado en sus inversiones bursátiles y un 33,85% tienen un perfil arriesgado, obtenga el número de inversores hombres y mujeres que posee la sociedad.

d) Elabore la tabla de frecuencias relativas conjuntas de los inversores atendiendo al sexo y al perfil inversor.

Ejercicio 3.3. Se desea especificar un modelo de regresión lineal simple (modelo RLS) para explicar el consumo de combustible en calefacción de casas unifamiliares (no adosadas) durante el mes de enero. Se dispone de una muestra de 15 casas construidas por la misma empresa en diferentes urbanizaciones. En cada una de estas casas se han observado las siguientes variables:

- **CON:** Consumo mensual de combustible en calefacción (en litros).
- **TEM:** Temperatura atmosférica promedio diaria (en °C).
- **ESP:** Espesor de la capa de aislamiento de la vivienda (en cm).

Los valores observados de estas variables son los que figuran en la Tabla 3.28.

Tabla 3.28. Valores de las variables CON, TEM y ESP

CON	TEM	ESP
1.100	14	7,5
1.456	8	7,5
656	14	25
164	25	15
376	21	15
924	13	15
1.468	6	15
1.204	2	25
952	8	25
484	20	7.5
124	22	25
812	16	15
1.764	7	7,5
1.292	14	7,5
208	17	25

A partir de dicha información se pide:

a) Vector de medias, matriz de varianzas y covarianzas y matriz de coeficientes de correlación para el conjunto formado por esas tres variables.

b) Indique cuál de las dos variables, TEM o ESP, proporciona un modelo RLS con mayor capacidad explicativa del comportamiento de la variable CON.

c) Para explicar el comportamiento de la variable CON, utilizando el programa R, se han ajustado las siguientes rectas:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1381.17     299.23   4.616 0.000484 ***
ESP           -32.56      17.22  -1.891 0.081068 .
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1783.54     176.95  10.079 1.64e-07 ***
TEM           -66.52      11.64  -5.717 7.09e-05 ***
```

Especifique la ecuación correspondiente a cada una de las rectas estimadas e indique qué porcentaje de la variación total observada en CON queda explicado mediante cada una de ellas.

d) Utilizando la recta con mayor capacidad explicativa, indique cuál será el consumo estimado de una vivienda situada en una localidad donde TEM = 15°C y ESP = 15 cm y el de una vivienda situada en una localidad con TEM = -5°C y ESP = 25 cm. Razone cuál de las dos predicciones es más fiable.

e) Obtenga la predicción del valor de CON para la tercera vivienda observada. Explique la discrepancia entre el valor observado de CON y la predicción. ¿Cómo se denomina a dicha predicción? ¿Es fiable?

Ejercicio 3.4. Con 72 datos mensuales anteriores a enero de un determinado año (A) que, sobre viajeros y grado de ocupación por plazas para Andalucía y España, ofrece la Encuesta de Ocupación Hotelera (EOH) que elabora el INE, se han estimado las rectas de regresión mínimo-cuadráticas para explicar, mediante un modelo lineal, las variaciones en el grado de ocupación a partir de las variaciones en el número de viajeros. En el Gráfico 3.4 se presentan las nubes de puntos y las rectas ajustadas $\left(y_A^* \text{ e } y_E^*\right)$ para Andalucía y España, respectivamente.

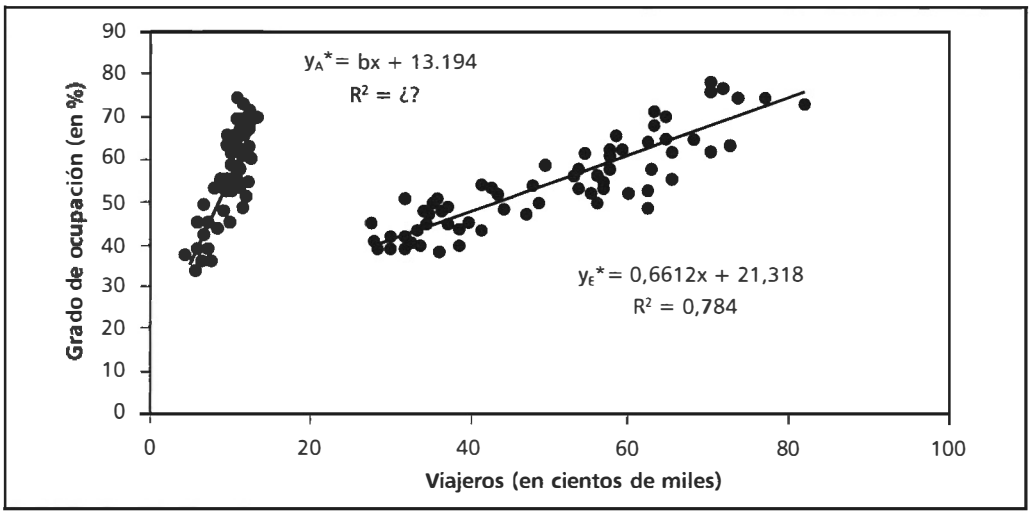


Gráfico 3.4. Variaciones en el grado de ocupación hotelera

Se pide:

a) Indique el significado, tanto matemático como en el contexto del enunciado, de los parámetros de la recta estimada para España, $y_E^* = 0,6612x + 21,318$.

b) Indique el porcentaje de las variaciones en el grado de ocupación hotelera en España que no vienen explicadas por las variaciones en el número de viajeros a través del modelo lineal estimado.

c) Si en el periodo utilizado para la estimación, por término medio, mensualmente visitaron España 5.097.323 viajeros, ¿cuál fue, por término medio, el grado de ocupación mensual en España durante ese periodo?

d) Determine, haciendo uso de los datos auxiliares que se le facilitan a continuación, la pendiente de la recta estimada para Andalucía y el coeficiente de determinación de dicho ajuste, sabiendo que el grado de ocupación hotelera viene expresado en tantos por ciento y el número de viajeros en cientos de miles.

$$\sum_i y_i = 3855,85;\ \sum_i x_i = 689,41626;\ y_i^2 = 215.900,59;$$
$$\sum_i x_i^2 = 6.992,4543;\ \sum_i x_i y_i = 38.569,275$$

e) Estime el grado de ocupación hotelera en enero del año A, tanto para España como para Andalucía, teniendo en cuenta que el número de viajeros en enero del año A en España y Andalucía ascendió a 3.356.718 y 656.711 viajeros, respectivamente.

Ejercicio 3.5. Haciendo uso de los datos anuales sobre las ventas de una empresa en el periodo 1999-2006, se desea analizar la relación entre las ventas en Andalucía (en millones de euros), que figura en la Tabla 3.29 con el nombre de V_AND, y las ventas en España (en millones de euros), que figura con el nombre de V_ESP.

Se sabe además que, para el periodo considerado:

- La media anual de las ventas en Andalucía fue de 12,26 millones de euros.
- La media anual de las ventas en España fue de 64,96 millones de euros.
- El coeficiente de correlación lineal simple entre las dos variables ascendió a 0,9937.

Tabla 3.29. Relación entre las ventas de Andalucía con respecto a España (en millones de euros)

Años	V_AND^2	V_ESP^2
1999	113,9	3.432,7
2000	124,7	3.514,4
2001	127,9	3.588,5
2002	130,1	3.584,3
2003	140,5	3.910,2
2004	157,1	4.466,4
2005	185,8	4.988,5
2006	241,1	6.729,2
Total	**1.221,07**	**34.214,12**

a) Estime una recta de regresión que permita explicar el comportamiento anual de las ventas en Andalucía a partir de los datos anuales correspondientes a España.

b) Explique el significado, matemático y económico, de los parámetros estimados en el apartado anterior.

c) ¿Qué porcentaje de las variaciones en las ventas anuales correspondientes a Andalucía no viene explicado por las variaciones en las ventas anuales correspondientes a toda España a través del modelo lineal estimado?

d) ¿A cuánto asciende el error cometido en la estimación, a partir de la recta de regresión, de las ventas en Andalucía en el año 2006?

e) Obtenga una predicción de las ventas en Andalucía para el año 2010, suponiendo que las ventas en España para ese año fuesen de 62.530.000 euros. Comente la fiabilidad de la predicción.

Ejercicio 3.6. Se desea explicar, mediante un modelo lineal, el comportamiento del grado de ocupación por plazas en Andalucía (%) en los doce meses de un determinado año. Con este fin, nos planteamos utilizar las observaciones mensuales correspondientes a dos posibles variables explicativas: número mensual de pernoctaciones (en millones) o número mensual de viajeros (en millones de personas). Los gráficos correspondientes a las nubes de puntos y a las dos posibles líneas de regresión figuran a continuación:

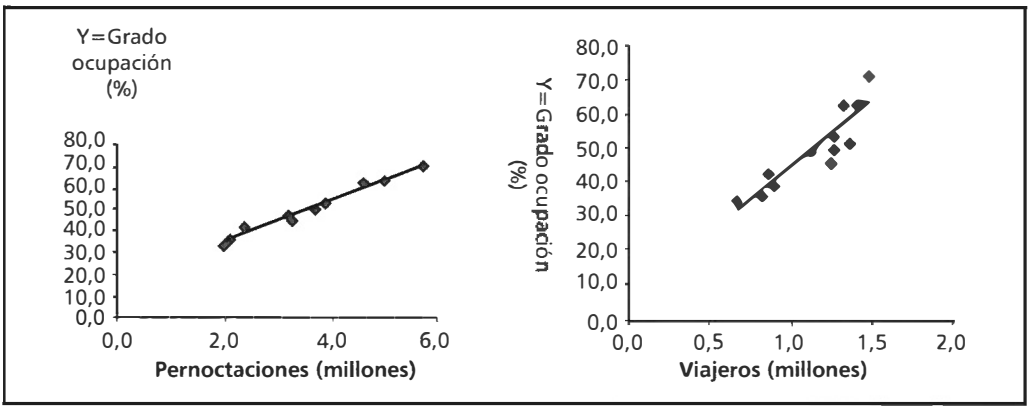


Gráfico 3.5. Comportamiento del grado de ocupación por plazas en Andalucía

Se sabe además que

$$\sum y_i = 598{,}6;\ \sum x_i = 41{,}3;\ \sum y_i x_i = 2.209{,}5;\ \sum y_i^2 = 31.270{,}3;\ \sum x_i^2 = 158{,}1$$

donde X son las pernoctaciones en millones.

Por otra parte, en cuanto a la recta de regresión que explica el grado de ocupación hotelera (Y) en función del número de viajeros, se han hecho los cálculos correspondientes y se sabe que:

i) El grado de ocupación hotelera asciende a 6,3% cuando el número de viajeros es cero.

ii) Cuando el número de viajeros aumenta en un millón de personas el grado de ocupación hotelera aumenta en 38,4 puntos porcentuales.

iii) El 17,8% de las variaciones en el grado de ocupación hotelera no vienen explicadas por el número de viajeros a través del modelo lineal.

Se pide:

a) Usando solo la información de los gráficos, ¿qué variable cree usted que podría explicar mejor el grado de ocupación hotelera? Justifique su respuesta.

b) Estime la recta de regresión que explica el grado de ocupación hotelera en función de las pernoctaciones.

c) Utilice la información de los puntos i) y ii), para determinar los parámetros de la recta que explica el grado de ocupación hotelera en función del número de viajeros.

d) Se sabe que el coeficiente de determinación (R^2) de la recta de regresión que explica el grado de ocupación hotelera en función de las pernoctaciones asciende a 0,994. ¿Cuál es el valor del Coeficiente de Determinación (R^2) de la otra recta de regresión?

e) Teniendo en cuenta el apartado anterior y sabiendo que el número de pernoctaciones para el mes de enero del año siguiente al que corresponden las observaciones asciende a 2.200.000 pernoctaciones y que el número de viajeros para dicho mes asciende a 800.000 personas, obtenga una predicción del grado de ocupación hotelera para dicho mes utilizando el modelo más adecuado.

Ejercicio 3.7. Con los datos que, sobre la Renta Nacional Disponible (RND) en 10^{12} euros y sobre los Gastos internos totales en actividades de I+D (GI+D) en 10^9 euros, proporciona el Instituto Nacional de Estadística para el periodo 2000-2009, se ha elaborado la Tabla 3.30.

Tabla 3.30. Renta Nacional disponible y Gastos internos en I+D para el período 200-2009

Años	GI+D	RND
2000	5,7	0,54
2001	6,2	0,58
2002	7,2	0,62
2003	8,2	0,67
2004	8,9	0,71
2005	10,2	0,76
2006	11,8	0,82
2007	13,3	0,87
2008	14,7	0,89
2009	14,6	0,85

Con los datos de la tabla se ha estimado la recta de regresión que permite explicar el Gasto en I+D a partir de la Renta Nacional Disponible en el periodo considerado. En el Gráfico 3.6 se han representado los datos y la recta estimada.

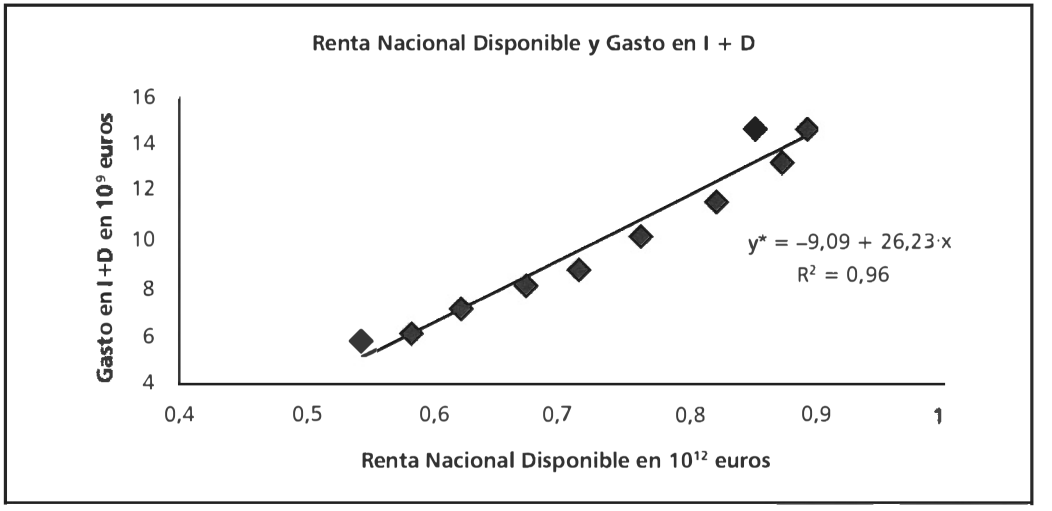


Gráfico 3.6. Recta de regresión del Gasto en I+D sobre la Renta Nacional Disponible

Se pide:

a) Dado que el valor de la pendiente es 26,23, demuestre que el valor del término independiente es –9,09.
b) Interprete los valores de los parámetros del modelo lineal estimado en el contexto del enunciado.
c) Utilice la recta estimada para hacer una predicción sobre el gasto en I+D en 2010 sabiendo que la renta nacional disponible para dicho año ascendió a 854.880 millones de euros.
d) Analice la fiabilidad de la predicción obtenida en el apartado anterior e indique qué porcentaje de las variaciones en el gasto anual en I+D no viene explicado por las variaciones en la renta anual nacional disponible a través del modelo estimado.

Ejercicio 3.8. Partiendo de los datos anuales, desde 2003 hasta 2011, de los precios de las acciones de BBVA, del índice IBEX 35 y de los tipos de interés de la Deuda del Estado a 3 años, se han calculado las correspondientes rentabilidades de las acciones BBVA (R_BBVA), del índice IBEX 35 (R_IBEX35) y de la Deuda del Estado a 3 años (R_DEUDA_E). Los datos, para los años 2004-2011, figuran en la Tabla 3.31.

Tabla 3.31. Rentabilidad de las acciones de BBVA, IBEX35 y Deuda del Estado a Bonos (2004-2011)

Años	R_BBVA	R_IBEX35	R_DEUDA_E
2004	27,04	19,82	1,39
2005	21,87	18,89	-10,20
2006	28,35	22,05	31,25
2007	5,46	18,78	16,98
2008	-34,75	-23,77	-5,87
2009	-13,71	-15,14	-55,57
2010	-2,83	1,06	16,65
2011	-18,55	-2,19	37,92

Nota: Rentabilidad de un activo = Ln $(\text{precio}_t / \text{precio}_{t-1}) \cdot 100$

Con los datos anteriores, se desea estimar la recta de regresión, conocida como modelo CAPM *(Capital Asset Pricing Model)*:

$$\mathbf{y_i = a + bx_i + e_i}$$

donde $\mathbf{y_i}$ mide el exceso de rentabilidad de las acciones BBVA con respecto a la rentabilidad de la Deuda del Estado (R_BBVA - R_DEUDA_E) y $\mathbf{x_i}$ mide el exceso de rentabilidad del índice IBEX 35 (R_IBEX35 - R_DEUDA_E). Los datos correspondientes a estas variables son los que figuran en la siguiente Tabla 3.32.

Tabla 3.32. Variables del exceso de rentabilidad de las acciones

Años	Y_i	X_i
2004	25,69917	18,47833
2005	32,06951	29,08936
2006	-2,901294	–9,194047
2007	-11,52368	1,801860
2008	-28,88106	–17,90856
2009	41,85659	40,42401
2010	-19,47858	–15,58075
2011	-56,46988	–40,10441

a) Represente la nube de puntos junto a la recta de regresión del modelo CAPM.

b) Estime la recta de regresión antes mencionada.

c) Interprete el significado del término independiente y del coeficiente de regresión. ¿Le parece razonable la estimación obtenida del término independiente?

d) Obtenga una medida de la bondad del ajuste e interprete su significado.

e) Para el año 2012 se estima que la rentabilidad del mercado (R_IBEX35) será del 10% y la rentabilidad del activo libre de riesgo del 15%. Obtenga una predicción de la rentabilidad de las acciones de BBVA para el año 2012.

CAPÍTULO 4
Números índices

Los números índices son la herramienta descriptiva que se utiliza para comparar dos situaciones tomando una de ellas como referencia. Su conocimiento es necesario para poder abordar, entre otros, la descripción de la evolución temporal de variables cuantitativas.

Comenzamos el tema con la definición de índices simples, en serie o en cadena, estudiando la relación entre ambos y definiendo el concepto asociado de tasa de variación. Posteriormente, pasaremos a estudiar los índices sintéticos o complejos, que se utilizan para resumir en un único indicador la variación porcentual que, en conjunto, experimentan varias variables cuantitativas. Distinguiremos entre índices complejos no ponderados y ponderados, según queramos dar a todas las variables la misma importancia en la construcción del índice o no. Entre los diversos métodos existentes para la construcción de índices complejos, estudiaremos los más simples: el método de la media aritmética y el método de la media agregativa.

A continuación, tras abordar los problemas del cambio de base y el enlace de series de números índices, se estudian el Índice de Precios al Consumo (IPC) y el Índice de Precios al Consumo Armonizado, por su utilidad para el seguimiento de la situación económica del país, al cuantificar la inflación sufrida por el consumidor general, así como algunos índices bursátiles, por su especial relevancia en el ámbito financiero.

Por último, trataremos el problema de la deflación de series económicas, fundamental para la interpretación de series que reflejan el valor de un bien o conjunto de bienes en una unidad monetaria, ya que la deflación permite apreciar la evolución real de la serie eliminando de la misma el efecto de la inflación.

4.1. Números índices simples

Los números índices simples son relativos a una magnitud variable y se obtienen al expresar los valores de la misma como porcentajes de uno de ellos que se toma como referencia. Al valor de la variable que se toma como referencia se le denomina **base del índice**. Dependiendo de si dicha referencia es fija o no, podemos hablar de: índices **en serie** (referencia fija) e índices **en cadena** (referencia variable).

Números índices simples en serie

Si x_t y x_0 son dos valores de una variable **X,** el valor del número índice en serie que le corresponde al valor x_t tomando como referencia o base fija el valor x_0 se representa mediante $I_0^t(X)$ y se define como

$$I_0^t(X) = \frac{x_t}{x_0} \cdot 100$$

Observación. Aunque las observaciones pueden ser transversales o temporales, es más frecuente la utilización de los números índices cuando se trabaja con datos temporales y la terminología está adaptada a ese tipo de datos. En este sentido, **la lectura que hacemos del símbolo $I_0^t(X)$ es «valor del índice de X para el periodo t tomando como base o referencia el valor en el periodo 0».**

Por otra parte, con los porcentajes se elimina la unidad de medida por lo que los números índices facilitan las comparaciones entre distintas variables.

Ejemplo 4.1. En la siguiente tabla se presenta el número de mujeres activas en España desde el tercer trimestre de 2009 hasta el tercer trimestre de 2010. Obtenga los índices simples con base en el tercer trimestre de 2009 que reflejan la evolución porcentual del número de mujeres activas.

Tabla 4.1. Mujeres activas

Trimestres	Activos (Miles de mujeres)
2009-III	10.089,4
2009-IV	10.139,3
2010-I	10.213,3
2010-II	10.250,5
2010-III	10.265,2

Fuente: EPA (INE)

Los índices pedidos son los que se presentan en la Tabla 4.2.

Tabla 4.2. Mujeres activas (índices en serie)

Trimestres	Activos (Miles de mujeres)	Índices Base en 2009-TIII
2009-III	10.089,4	100
2009-IV	10.139,3	100,4946
2010-I	10.213,3	101,2280
2010-II	10.250,5	101,5967
2010-III	10.265,2	101,7424

Fuente: Elaboración propia a partir de la EPA (INE)

Estos índices reflejan la variación porcentual que experimentan los distintos valores de la variable con respecto al valor que se ha tomado como referencia. A la vista de la tabla, podemos decir, por ejemplo, que el n.º de mujeres activas en España en el tercer trimestre de 2010 es un 1,74% superior al existente en el mismo trimestre del año anterior.

En el ejemplo anterior hemos calculado los índices correspondientes a cada valor tomando siempre como referencia o base el valor del tercer trimestre de 2009. De acuerdo con la definición, los índices que, como los anteriores, se obtienen para una serie de datos respecto de una **base (o periodo de referencia) fija** se llaman **índices en serie.**

Números índices simples en cadena

Cuando el índice correspondiente a cada dato se calcula tomando como referencia o base el dato inmediato anterior, se obtienen los **índices en cadena**. Para series de observaciones temporales, estos índices reflejan la variación porcentual que experimenta la variable entre cada dos observaciones consecutivas.

Concretamente, si x_t y x_{t-1} son los valores observados de una variable X en dos instantes consecutivos, el índice en cadena que le corresponde al valor x_t se representa mediante $IC^t(X)$ y se define como

$$IC^t = \frac{x_t}{x_{t-1}} \cdot 100$$

Ejemplo 4.2. Los índices en cadena correspondientes a los datos en los que se basa el Ejemplo 4.1 son los que figuran en la Tabla 4.3.

Tabla 4.3. Mujeres activas (índices en cadena)

Trimestres	Activos (Miles de mujeres)	Índices en cadena
2009-III	10.089,4	...
2009-IV	10.139,3	100,49
2010-I	10.213,3	100,73
2010-II	10.250,5	100,36
2010-III	10.265,2	100,14

Fuente: Elaboración propia a partir de la EPA (INE)

En este ejemplo los índices en cadena reflejan la variación porcentual intertrimestral del n.º de mujeres activas en España. Por ejemplo, el valor del índice que corresponde al segundo trimestre de 2010 es 100,36. Dicho valor refleja que en el segundo trimestre de 2010 el n.º de mujeres activas fue un 0,36% superior al dato del trimestre anterior.

Relación entre índices en serie e índices en cadena

1. **Los índices en cadena se pueden obtener a partir de los índices en serie.**

$$IC^t(X) = \frac{x_t}{x_{t-1}} \cdot 100 = \frac{\frac{x_t}{x_0} \cdot 100}{\frac{x_{t-1}}{x_0} \cdot 100} = \frac{I_0^t(X)}{I_0^{t-1}(X)} \cdot 100$$

2. Los índices en serie se pueden obtener a partir de los índices en cadena.

$$I_0^t(X) = \frac{x_t}{x_0} \cdot 100 = \frac{x_t}{x_{t-1}} \cdot \frac{x_{t-1}}{x_{t-2}} \cdot \frac{x_{t-2}}{x_{t-3}} \cdot \ldots \cdot \frac{x_2}{x_1} \cdot \frac{x_1}{x_0} \cdot 100 =$$

$$\frac{IC^t(X)}{100} \cdot \frac{IC^{t-1}(X)}{100} \cdot \frac{IC^{t-2}(X)}{100} \cdot \ldots \cdot \frac{IC^2(X)}{100} \cdot \frac{IC^1(X)}{100} \cdot 100$$

Ejemplo 4.3. Las siguientes series recogen los índices en cadena trimestrales para el número de parados en los sectores de la construcción y servicios en España.

Tabla 4.4. Parados en construcción y servicios

Trimestres	IC Parados Construcción	IC Parados Servicios
2009-I	...	...
2009-II	94,37	101,33
2009-III	88,64	95,84
2009-IV	98,79	100,70
2010-I	97,87	106,35
2010-II	87,71	95,91
2010-III	87,40	96,04

Fuente: Elaboración propia a partir de la EPA (INE)

a) Determine, a partir de la tabla, la variación porcentual que experimentó el número de parados en el sector servicios en el periodo que va desde 2009-III hasta 2010-III y en el periodo que va desde 2009-I hasta 2010-III.

b) Sabiendo que en el segundo trimestre de 2010 el n.º de parados en el sector de la construcción ascendió a 527,6 miles de personas, obtenga la serie trimestral para el número de parados en el sector de la construcción, expresada en miles de personas, desde el primer trimestre de 2009 hasta el tercer trimestre de 2010.

Solución. En el apartado a) debemos calcular

$$I_{2009\text{-III}}^{2010\text{-III}}(\text{parados servicios}) \text{ e } I_{2009\text{-I}}^{2010\text{-III}}(\text{parados servicios})$$

Teniendo en cuenta la relación entre índices simples y en cadena, obtenemos que

$$I_{2009\text{-III}}^{2010\text{-III}}(\text{parados servicios}) = \frac{IC^{2010\text{-III}}}{100} \cdot \frac{IC^{2010\text{-II}}}{100} \cdot \frac{IC^{2010\text{-I}}}{100} \cdot \frac{IC^{2009\text{-IV}}}{100} 100 =$$

$$= 0{,}9604 \cdot 0{,}9591 \cdot 1{,}0635 \cdot 1{,}007 \cdot 100 = 98{,}65$$

Por lo tanto, la variación porcentual que corresponde al periodo 2009-III-2010-III es de –1,35%. Del mismo modo

$$I_{2009\text{-I}}^{2010\text{-III}}(\text{parados servicios}) =$$

$$= \frac{IC^{2010\text{-III}}}{100} \cdot \frac{IC^{2010\text{-II}}}{100} \cdot \frac{IC^{2010\text{-I}}}{100} \cdot \frac{IC^{2009\text{-IV}}}{100} \cdot \frac{IC^{2009\text{-III}}}{100} \cdot \frac{IC^{2009\text{-II}}}{100} \cdot 100 =$$

$$= 0{,}9604 \cdot 0{,}9591 \cdot 1{,}0635 \cdot 1{,}007 \cdot 0{,}9584 \cdot 1{,}0133 \cdot 100 = 95{,}8$$

Por lo tanto, la variación porcentual que corresponde al periodo 2009-I-2010-III es de –4,2%.

Para el apartado b) lo más cómodo es obtener la serie de índices en serie con base en 2009-I para el n.º de parados en el sector de la construcción y, luego, utilizar el dato absoluto del segundo trimestre de 2010 para obtener la serie pedida. Dichos índices son los que se muestran en la Tabla 4.5.

Tabla 4.5. Parados en el sector de la construcción (Índices en serie)

Trimestres	IC Parados Construcción	Índices base 2009-I (Parados Const.)
2009-I	...	100,00
2009-II	94,37	94,37
2009-III	88,64	83,65
2009-IV	98,79	82,64
2010-I	97,87	80,88
2010-II	87,71	70,94
2010-III	87,40	62,00

Fuente: Elaboración propia a partir de la EPA (INE)

Como

$$I_{2009\text{-I}}^{2010\text{-III}}\left(\text{parados construcción}\right)=\frac{x_{2010\text{-II}}}{x_{2009\text{-I}}}\cdot 100=70{,}94$$

utilizando el dato para 2010-II, se obtiene que

$$\frac{527{,}6}{x_{2009\text{-I}}}\cdot 100=70{,}94$$

de donde se concluye que

$$x_{2009\text{-I}}=743{,}727 \text{ miles de parados}$$

Por último, utilizando el índice para cada trimestre y el valor en términos absolutos para el primer trimestre de 2009, se obtienen los restantes datos, que se muestran en la Tabla 4.6.

Tabla 4.6. Miles de parados en el sector de la construcción

Trimestres	IC Parados Construcción	Índices base 2009-I (Parados Const.)	Miles de parados (Construcción)
2009-I	...	100,00	743,727
2009-II	94,37	94,37	701,855
2009-III	88,64	83,65	622,128
2009-IV	98,79	82,64	614,616
2010-I	97,87	80,88	601,526
2010-II	87,71	70,94	527,600
2010-III	87,40	62,00	461,111

Fuente: Elaboración propia a partir de la EPA (INE)

4.2. Tasas de variación

Si x_{t_1} es el valor de una variable **X** en el instante o periodo de tiempo t_1 y x_{t_2} el valor de la misma en un instante o periodo posterior t_2, la **tasa de variación** de **X** en t_2 con respecto a t_1, $\mathbf{Tasa}_{t_1}^{t_2}(x)$, se define como

$$\mathbf{Tasa}_{t_1}^{t_2}(x)=\frac{x_{t_2}-x_{t_1}}{x_{t_1}}\cdot 100$$

Observemos que

$$\mathbf{Tasa}_{t_1}^{t_2}(x)=I_{t_1}^{t_2}(x)-100$$

Dicha tasa será positiva, negativa o nula, según que el valor de **X** en t_2 sea mayor, menor o igual que el correspondiente a t_1.

Es obvio que la tasa de variación entre dos observaciones consecutivas de **X**, $\mathbf{Tasa}_{t-1}^{t}(x)$, a la que representaremos con $\mathbf{Tasa}^{t}(x)$, se calcula a partir del índice en cadena mediante

$$\mathbf{Tasa}^{t}(x)=\frac{x_t-x_{t-1}}{x_{t-1}}\cdot 100 = IC^{t}(x)-100$$

Observación. En general, hablaremos de **tasa de variación interanual**, **intertrimestral** o **intermensual**, para referirnos a la tasa de variación entre observaciones consecutivas y correspondientes a años, trimestres o meses, respectivamente.

También se suele utilizar la expresión **tasa de variación interanual correspondiente a un determinado mes (o trimestre)** para referirse a la variación porcentual que experimenta la variable en un determinado mes (o trimestre) con respecto al valor correspondiente al mismo mes (o trimestre) del año inmediato anterior, cuando se trabaja con series de datos mensuales o trimestrales correspondientes a distintos años.

Ejemplo 4.4. En la Tabla 4.7 se presenta el gasto total en viajes turísticos de los residentes en España, por tipo de destino, para los años del periodo 2001-2004. Se pide:

a) ¿Cuál fue el incremento porcentual del gasto en viajes turísticos de los residentes en España entre los años 2001 y 2004?

b) Determine la variación porcentual del gasto en viajes turísticos con destino el extranjero de los residentes en España que corresponde a cada año con respecto al año 2001.

c) Determine las tasas de variación interanual (%) para el gasto por viajes turísticos con destino España de los residentes en España que corresponden a los años del periodo 2001-2004, sabiendo que el incremento porcentual que dicho gasto experimentó en el año 2001 con respecto al año 2000 fue de un 16,2%.

Tabla 4.7. Gasto total en viajes turísticos de los residentes en España

Datos absolutos (10^6 Euros)	Año 2001	Año 2002	Año 2003	Año 2004
Viajes turísticos	12.815,2	12.093,0	12.743,7	14.568,5
Destino España	9.896,0	9.274,6	9.828,5	11.154,3
Destino extranjero	2.919,2	2.814,4	2.915,2	3.414,2

Fuente: IET, Movimientos turísticos de los españoles (Familitur)

Solución

a) $$I_{2001}^{2004}(\text{Gasto en viajes turísticos}) = \frac{14.568,5}{12.815,2} \cdot 100 = 113,68$$

Por lo tanto,

$$\text{Tasa}_{2001}^{2004}(\text{Gasto}) = 13,68\%$$

b) Obtenemos la serie de índices simples con base 2001 para el gasto en viajes turísticos con destino el extranjero y, a partir de ella, restando 100 a cada índice se obtienen las tasas pedidas. Dicha serie y las correpondientes tasas se presentan en la Tabla 4.8.

Tabla 4.8. Gasto en viajes turísticos con destino el extranjero (índices y tasas)

	2001	2002	2003	2004
Gasto (destino extranjero) 10^6 €	2.919,2	2.814,4	2.915,2	3.414,2
$I_{2001}^{\text{Año}}$	100	96,41	99,86	116,96
$\text{Tasa}_{2001}^{\text{Año}}$	0	-3,59	–0,14	16,96

c) Las tasas de variación interanual, que se calculan a partir de los índices en cadena, son las que se muestran en la Tabla 4.9.

Tabla 4.9. Gasto en viajes turísticos con destino España (índices en cadena)

	2001	2002	2002	2004
Gasto (destino España) en 10^6 €	9.896,0	9.274,6	9.828,5	11.154,3
$IC^{\text{Año}}$	116,2	93,72	105,97	113,49
Tasa interanual	16,2	–6,28	5,97	13,49

Ejemplo 4.5. En la Tabla 4.10 figura el número de hipotecas inmobiliarias constituidas para fincas rústicas entre los meses de Junio de 2004 y Diciembre de 2004. Determine las tasas intermensuales de variación correspondientes al n.º de hipotecas para fincas rústicas. Sabiendo que el n.º de hipotecas para fincas rústicas en el mes de agosto de 2005 ascendió a 3.940, obtenga la tasa de variación interanual correspondiente al mes de agosto de 2005.

Tabla 4.10. Número de hipotecas rústicas

	2004-06	2004-07	2004-08	2004-09	2004-10	2004-11	2004-12
Hipotecas	4.155	3.836	3.380	4.212	4.119	3.927	3.801

Fuente: Boletín Mensual de Estadística, INE

Solución

Tabla 4.11. Índices en cadena y tasas intermensuales (hipotecas rústicas)

	2004-06	2004-07	2004-08	2004-09	2004-10	2004-11	2004-12
Hipotecas	4.155	3.836	3.380	4.212	4.119	3.927	3.801
IC(Mes)	...	92,32	88,11	124,62	97,79	95,34	96,79
Tasa (Mes)	...	–7,68	–11,89	24,62	–2,21	–4,66	–3,21

Teniendo en cuenta que en agosto de 2005 se constituyeron 3.940 hipotecas rústicas, podemos calcular la tasa de variación interanual correspondiente a agosto del año 2005.

$$\textbf{Tasa interanual(2005-08)} = \frac{\textbf{Hipotecas (2005-08)}}{\textbf{Hipotecas (2004-08)}} \cdot 100 - 100 =$$

$$= \frac{3.940}{3.380} \cdot 100 - 100 =$$

$$= 116{,}57 - 100 = 16{,}57\%$$

4.2.1. Tasa media de variación

Llamamos **tasa media de variación** de la variable **X** en el periodo **[t,t+k]** (o, **tasa media de crecimiento acumulativo** en el periodo **[t,t+k]**) a la tasa T_k que permite obtener la observación en el instante o periodo **t+k**, x_{t+k}, partiendo de la observación x_t, en el instante o periodo **t**, y aplicando, entre instantes o periodos consecutivos, un crecimiento porcentual constante e igual a T_k.

Teniendo en cuenta la definición anterior y representando con «y» a las observaciones auxiliares intermedias, se tiene que

$$y_{t+1} = x_t + \frac{T_k}{100} \cdot x_t = \left(\frac{100 + T_k}{100}\right) \cdot x_t$$

$$y_{t+2} = y_{t+1} + \frac{T_k}{100} \cdot y_{t+1} = \left(\frac{100+T_k}{100}\right) \cdot y_{t+1} = \left(\frac{100+T_k}{100}\right)^2 \cdot x_t$$

$$y_{t+3} = y_{t+2} + \frac{T_k}{100} \cdot y_{t+2} = \left(\frac{100+T_k}{100}\right) \cdot y_{t+2} = \left(\frac{100+T_k}{100}\right)^3 \cdot x_t$$

.

.

.

$$y_{t+k} = y_{t+k-1} + \frac{T_k}{100} \cdot y_{t+k-1} = \left(\frac{100+T_k}{100}\right) \cdot y_{t+k-1} = \left(\frac{100+T_k}{100}\right)^k \cdot x_t$$

Como queremos que se cumpla $y_{t+k} = x_{t+k}$, debe ser

$$x_{t+k} = \left(\frac{100 + T_k}{100}\right)^k \cdot x_t$$

y, por lo tanto

$$\sqrt[k]{\frac{x_{t+k}}{x_t}} = \frac{100 + T_k}{100}$$

o, lo que es equivalente

$$T_k = 100 \cdot \sqrt[k]{\frac{x_{t+k}}{x_t}} - 100$$

Ejemplo 4.6. Basándose en los datos de la Tabla 4.10, determine la tasa media de variación intermensual que corresponde al n.º de hipotecas para fincas rústicas constituidas entre Junio de 2004 y Diciembre de 2004. Determine también, haciendo uso del dato relativo a agosto de 2005, el crecimiento medio mensual acumulativo para el periodo agosto 2004-agosto 2005.

Solución

$$T_6 = 100 \cdot \sqrt[6]{\frac{3.801}{4.155}} - 100 = 100 \cdot \sqrt[6]{0,9148} - 100 =$$
$$= 100 \cdot 0,9853 - 100 = 98,53 - 100 = -1,47\%$$

Para determinar el crecimiento medio mensual acumulativo para el periodo agosto 2004-agosto 2005, calculamos la tasa correspondiente a ese periodo de 13 meses

$$T_{12} = 100 \cdot \sqrt[12]{\frac{3.940}{3.380}} - 100 = 100 \cdot \sqrt[12]{1,16568} - 100 =$$
$$= 100 \cdot 1,012857 - 100 = 101,29 - 100 = 1,29\%$$

4.3. Números índices complejos

Los índices complejos o compuestos se emplean para resumir de forma conjunta las fluctuaciones de un conjunto de **N** variables relacionadas entre sí. Los dos métodos más simples para obtener este tipo de índices son: **el método de la media aritmética** y **el método de la media agregativa**.

Consideremos la variable compleja $\mathbf{X=(X_1,\ldots,X_N)}$ que en los periodos base **(T=0)** y actual **(T=t)** ha tomado los valores que se presentan en la siguiente tabla.

Periodo base	Periodo actual	Índices simples
x_{10}	x_{1t}	$I_0^t(x_1) = \frac{x_{1t}}{x_{10}} \cdot 100$
...	...	...
x_{io}	x_{it}	$I_0^t(x_i) = \frac{x_{it}}{x_{i0}} \cdot 100$
...	...	...
x_{N0}	x_{Nt}	$I_0^t(x_N) = \frac{x_{Nt}}{x_{N0}} \cdot 100$

a.1) El índice complejo que se obtiene mediante el método de la media aritmética simple, $\bar{I}_0^t(\mathbf{x})$, se define como

$$\bar{I}_0^t(\mathbf{x})=\frac{I_0^t(x_1)+I_0^t(x_2)+\ldots+I_0^t(x_N)}{N}$$

a.2) El índice complejo que se obtiene mediante el método de la media agregativa simple, $(I_A)_0^t(\mathbf{x})$, se define como

$$(I_A)_0^t(\mathbf{x})=\frac{x_{1t}+x_{2t}+\ldots+x_{Nt}}{x_{10}+x_{20}+\ldots+x_{N0}}\cdot 100$$

Observemos que este **último tipo de índice complejo solo tiene sentido si las distintas variables vienen expresadas en la misma unidad de medida.**

Cuando se quiere incorporar al índice complejo la importancia relativa que cada una de las magnitudes simples tiene dentro del conjunto formado por todas ellas, se introducen pesos o ponderaciones que varían de una magnitud a otra y que pueden ser dependientes o no del instante en el que se calcula el índice, los índices complejos así obtenidos se denominan **índices complejos ponderados.**

Los casos más simples de índices complejos ponderados son los siguientes:

b.1) El índice complejo que se obtiene mediante el método de la media aritmética ponderada, $(I_w)_0^t(\mathbf{x})$, con pesos o ponderaciones $\{w_1, w_2, \ldots, w_N\}$, se define como

$$(\bar{I}_w)_0^t(\mathbf{x})=\frac{I_0^t(x_1)\cdot w_1+I_0^t(x_2)\cdot w_2+\ldots+I_0^t(x_N)\cdot w_N}{w_1+w_2+\ldots+w_N}$$

b.2) El índice complejo que se obtiene mediante el método de la media agregativa ponderada, $(I_{Aw})_0^t(\mathbf{x})$, con pesos o ponderaciones $\{w_1, w_2, \ldots, w_N\}$, se define como

$$(\bar{I}_{Aw})_0^t(\mathbf{x})=\frac{x_{1t}\cdot w_1+x_{2t}\cdot w_2+\ldots+x_{Nt}\cdot w_N}{x_{10}\cdot w_1+x_{20}\cdot w_2+\ldots+x_{N0}\cdot w_N}\cdot 100$$

Ejemplo 4.7. En la Tabla 4.12 se recoge el número mensual de hipotecas inmobiliarias constituidas para fincas de naturaleza urbana, atendiendo a la clasificación de dichas fincas en viviendas, solares y otras, desde enero hasta agosto de 2010. Utilice la información contenida en dicha tabla para obtener una serie de números índices complejos, mediante el método de la media agregativa simple, que refleje la evolución intermensual del n.º total de hipotecas constituidas para fincas de naturaleza urbana.

Tabla 4.12. Número de hipotecas inmobiliarias constituidas

Hipotecas para fincas de naturaleza urbana			
	Viviendas	Solares	Otras
2010-01	53.747	3.625	28.164
2010-02	54.813	3.830	24.875
2010-03	53.513	3.543	24.175
2010-04	50.342	3.034	22.352
2010-05	55.755	3.083	25.487
2010-06	55.143	3.007	25.140
2010-07	55.570	2.277	22.348
2010-08	50.223	3.095	20.833

Fuente: INE, Boletín Mensual de Estadística

Solución

Obtendremos, en primer lugar, el total de hipotecas para fincas urbanas que corresponde a cada mes y, a partir del mismo, elaboraremos los índices en cadena que reflejarán las variaciones intermensuales (Tabla 4.13).

Tabla 4.13. Índices complejos en cadena

Hipotecas para fincas de naturaleza urbana						
	Vivienda	Solares	Otras	Total	IC(Total)	Tasas de variación (Intermensuales)
2010-01	53.747	3.625	28.164	85.536	...	...
2010-02	54.813	3.830	24.875	83.518	97,64	-2,36
2010-03	53.513	3.543	24.175	81.231	97,26	-2,74
2010-04	50.342	3.034	22.352	75.728	93,23	-6,77
2010-05	55.755	3.083	25.487	84.325	111,35	11,35
2010-06	55.143	3.007	25.140	83.290	98,77	-1,23
2010-07	55.570	2.277	22.348	80.195	96,28	-3,72
2010-08	50.223	3.095	20.833	74.151	92,46	-7,54

Fuente: Elaboración propia a partir del Boletín Mensual de Estadística (INE)

4.3.1. Índices complejos de precios, cantidades y valor

En este epígrafe nos centraremos en el estudio de algunos índices complejos ponderados concretos (**Laspeyres, Paasche y Fisher**) para precios, cantidades y valor. La notación que vamos a utilizar es la que queda recogida en el siguiente cuadro.

Cuadro 4.1. Notación

Bien i				
	Precio	**Cantidad**	**Valor**	**Ponderación**
Tiempo t	p_{it}	q_{it}	$v_{it}=p_{it}\,q_{it}$	w_{it}
Tiempo 0	p_{io}	q_{io}	$v_{io}=p_{io}\,q_{io}$	w_{io}
Índices simples	$p_0^t(i)$	$q_0^t(i)$	$v_0^t(i)$	

Índices complejos ponderados de precios

Precios de Laspeyres $\left(P_{L,0}^t\right)$: es un índice de precios complejo ponderado, obtenido por el método de la media aritmética ponderada de los índices simples. Las ponderaciones son fijas en el tiempo y coinciden con el valor de los distintos bienes en el instante que se toma como referencia o base. Concretamente

$$P_{L,0}^t = \frac{\sum_{i=1}^{N} p_0^t(i)\cdot w_{io}}{\sum_{i=1}^{N} w_{io}} = \frac{\frac{p_{1t}}{p_{10}}\cdot p_{10}\cdot q_{10} + \frac{p_{2t}}{p_{20}}\cdot p_{20}\cdot q_{20} + \ldots + \frac{p_{Nt}}{p_{N0}}\cdot p_{N0}\cdot q_{N0}}{p_{10}\cdot q_{10} + p_{20} + \ldots + p_{N0}\cdot q_{N0}}\cdot 100 =$$

$$= \frac{\sum_{i=1}^{i=N} p_{it}\cdot q_{io}}{\sum_{i=1}^{i=N} p_{io}\cdot q_{io}}\cdot 100$$

Observación. El índice de precios de Laspeyres también puede verse como un índice complejo ponderado obtenido por el método de la media agregativa ponderada de los precios de los distintos bienes, utilizando como ponderación para cada bien la cantidad correspondiente al mismo en el instante tomado como referencia o base.

Precios de Paasche $\left(P^t_{P,0}\right)$**:** es un índice de precios complejo ponderado obtenido por el método de la media aritmética ponderada de los índices simples, pero las ponderaciones son variables en el tiempo y coinciden con el valor de los distintos bienes en el instante para el que se calcula el índice a precios del instante que se toma como referencia o base. Esto es

$$P^t_{P,0} = \frac{\sum_{i=1}^{N} p_0^t(i)\cdot w_{it}}{\sum_{i=1}^{N} w_{it}} = \frac{\frac{p_{1t}}{p_{10}}\cdot p_{10}\cdot q_{1t} + \frac{p_{2t}}{p_{20}}\cdot p_{20}\cdot q_{2t} + \ldots + \frac{p_{Nt}}{p_{N0}}\cdot p_{N0}\cdot q_{Nt}}{p_{10}\cdot q_{1t} + p_{20}\cdot q_{2t} + \ldots + p_{N0}\cdot q_{Nt}}\cdot 100 =$$

$$= \frac{\sum_{i=1}^{i=N} p_{it}\cdot q_{it}}{\sum_{i=1}^{i=N} p_{io}\cdot q_{it}}\cdot 100$$

Observación. Como en el caso anterior, este índice puede verse también como un índice complejo ponderado de precios obtenido mediante el método de la media agregativa ponderada, donde las ponderaciones son las cantidades de los distintos bienes en el instante para el que se calcula el índice.

Precios de Fisher $\left(P^t_{F,0}\right)$**:** se define como la «media geométrica» de los índices de precios de Laspeyres y Paasche; es decir

$$\left(P^t_{F,0}\right) = \sqrt{P^t_{L,0}\cdot P^t_{P,0}}$$

Índices complejos ponderados de cantidades

Cantidades de Laspeyres: índice complejo ponderado de cantidades, obtenido por el método de la media aritmética ponderada de los índices simples. Las ponderaciones son, para los distintos bienes, el valor del bien en el periodo base.

$$Q^t_{L,0} = \frac{\sum_{i=1}^{N} q_0^t(i)\cdot w_i}{\sum_{i=1}^{N} w_i} = \frac{\frac{q_{1t}}{q_{10}}\cdot p_{10}\cdot q_{10} + \frac{q_{2t}}{q_{20}}\cdot p_{20}\cdot q_{20} + \ldots + \frac{q_{Nt}}{q_{N0}}\cdot p_{N0}\cdot q_{N0}}{p_{10}\cdot q_{10} + p_{20}\cdot q_{20} + \ldots + p_{N0}\cdot q_{N0}}\cdot 100 =$$

$$= \frac{\sum_{i=1}^{i=N} p_{i0}\cdot q_{it}}{\sum_{i=1}^{i=N} p_{io}\cdot q_{io}}\cdot 100$$

Cantidades de Paasche: índice complejo ponderado de cantidades, obtenido por el método de la media aritmética ponderada de los índices simples. Las ponderaciones son, para los distintos bienes, el valor del bien en tiempo actual con cantidades del tiempo base.

$$Q^t_{P,0} = \frac{\sum_{i=1}^{N} q^t_0(i) \cdot w_{it}}{\sum_{i=1}^{N} w_{it}} = \frac{\frac{q_{1t}}{q_{10}} \cdot p_{1t} \cdot q_{10} + \frac{q_{2t}}{q_{20}} \cdot p_{2t} \cdot q_{20} + \ldots + \frac{q_{Nt}}{q_{N0}} \cdot p_{Nt} \cdot q_{N0}}{p_{1t} \cdot q_{10} + p_{2t} \cdot q_{20} + \ldots + p_{Nt} \cdot q_{N0}} \cdot 100 =$$

$$= \frac{\sum_{i=1}^{i=N} p_{it} \cdot q_{it}}{\sum_{i=1}^{i=N} p_{it} \cdot q_{i0}} \cdot 100$$

Cantidades de Fisher: índice complejo ponderado que se define como la media geométrica de los índices de cantidades de Laspyres y Paasche; es decir,

$$Q^t_{F,0} = \sqrt{Q^t_{L,0} \cdot Q^t_{P,0}}$$

Índice complejo de valor

Los valores de un conjunto de N bienes se expresan en unidades monetarias y, por tanto, son comparables. Si los valores están expresados en una unidad común, el índice complejo de valor se obtiene por el método de la media agregativa simple

$$v^t_0 = \frac{p_{1t}q_{1t} + p_{2t}q_{2t} + \ldots + p_{Nt}q_{Nt}}{p_{10}q_{10} + p_{20}q_{20} + \ldots + p_{N0}q_{N0}} \cdot 100 = \frac{\sum_{i=1}^{i=N} p_{it}q_{it}}{\sum_{i=1}^{i=N} p_{io}q_{io}} \cdot 100$$

Relaciones entre el índice complejo de valor y los índices de precios y cantidades de Laspeyres, Paasche y Fisher

Cuando se trabaja con un único bien, en términos absolutos, valor es igual a precio por cantidad

$$v_{it} = p_{it} \cdot q_{it}$$

Sin embargo, en términos porcentuales (índices), lo que se cumple es

$$v^t_0(i) = \frac{v_{it}}{v_{i0}} \cdot 100 = \frac{p_{it}}{p_{i0}} \cdot \frac{q_{it}}{q_{i0}} \cdot 100 = \frac{P^t_0(i) \cdot Q^t_0(i)}{100}$$

Análogamente, cuando se trabaja con un conjunto de N bienes y sus índices complejos asociados de precios, cantidades y valor, se cumplen las siguientes relaciones:

$$v^t_0 = \frac{P^t_{L,0} \cdot Q^t_{P,0}}{100} = \frac{P^t_{P,0} \cdot Q^t_{L,0}}{100} = \frac{P^t_{F,0} \cdot Q^t_{F,0}}{100}$$

Ejemplo 4.8. Supongamos que los precios y cantidades relativos a tres bienes A, B y C en el periodo 2007-2009 son los que figuran en la Tabla 4.14.

Tabla 4.14. Precios y cantidades de los bienes A, B y C

	A		B		C	
Años	Precio	Cantidad	Precio	Cantidad	Precio	Cantidad
2007	2	8	4	7	3	25
2008	3	7	6	7	4	30
2009	4	8	8	7	4	35

Determine los índices complejos de precios y cantidades para el conjunto de los tres bienes, en el año 2009 tomando como referencia o base 2007, mediante los métodos de Laspeyres, Paasche y Fisher. Obtenga el índice complejo de valor para 2009 tomando como base 2007 y compruebe que se cumplen las relaciones antes enunciadas.

Solución. El resultado de aplicar las fórmulas correspondientes a los índices pedidos es el que se muestra en la Tabla 4.15.

Tabla 4.15. Índices complejos para los bienes de la Tabla 4.14

	Laspeyres		Paasche		Fisher	
Años	Precios	Cantidades	Precios	Cantidades	Precios	Cantidades
2007	100,00	100,00	100,00	100,00	100,00	100,00
2008	139,50	110,92	138,64	110,24	139,07	110,58
2009	157,98	125,21	153,02	121,28	155,48	123,23

En cuanto al índice complejo para el valor en 2009 tomando como referencia el valor de 2007, se puede comprobar que el mismo asciende a 191,6 y que, además de a partir de los datos originales, se puede obtener a partir de cualquiera de las igualdades antes enunciadas.

$$v_{2007}^{2009} = \frac{P_{L,2007}^{2009} \cdot Q_{P,2007}^{2009}}{100} = \frac{P_{P,2007}^{2009} \cdot Q_{L,2007}^{2009}}{100} = \frac{157{,}98 \cdot 121{,}28}{100} = \frac{153{,}02 \cdot 125{,}21}{100} = 191{,}6$$

$$v_{2007}^{2009} = \frac{P_{F,2007}^{2009} \cdot Q_{F,2007}^{2009}}{100} = \frac{155{,}48 \cdot 123{,}23}{100} = 191{,}6$$

4.4. Enlace y cambio de base

4.4.1. Cambio de base

Para facilitar comparaciones o realizar determinados cálculos puede resultar conveniente modificar la base de una serie de índices ya calculados. Para ello, basta tener en cuenta que:

- El índice toma el valor 100 en el instante o periodo que se toma como nueva base.
- Los restantes valores del índice deben transformarse en la misma proporción que el correspondiente a la nueva base.

— La relación entre los índices de la nueva serie, con base en T=i, y los de la antigua, con base en T=0, es

$$I_i^j = I_0^j \cdot \frac{100}{I_0^i}$$

Periodo	Índice (Base en T=0)	Índice (Base en T=i)
0	100	I_I^0
1	I_0^1	I_i^1
...	.	.
j	I_0^j	I_I^j
...	.	.
i	I_0^i	100
...		.
t	I_0^t	I_i^t

Ejemplo 4.9. Transformar la serie de índices con base 1999 que se muestra en la Tabla 4.16 en otra en la que los índices estén expresados en base 2005.

Tabla 4.16. Índices en serie con base en el dato del año 1999

Años	Índices (Base 1999)
1999	100
2000	108
2001	115
2002	120
2003	128
2004	132
2005	145
2006	154

Solución. La relación entre los índices con base en 1999 y los que tienen base en 2005 es la siguiente

$$I_{2005}^t = I_{1999}^t \cdot \frac{100}{I_{1999}^{2005}}$$

Teniendo en cuenta la relación anterior, se obtiene el contenido de la Tabla 4.17.

Tabla 4.17. Índices en serie con base en el dato del año 2005

Años	Índices (Base 2005)
1999	68,96
2000	74,48
2001	79,31
2002	82,76
2003	88,27
2004	91,03
2005	100
2006	106,21

4.4.2. Enlace de series de índices

Cuando disponemos de dos o más series de números índices relativas a una misma magnitud variable (o conjunto de ellas) que han sido observadas en periodos de tiempo que se solapan, podemos obtener, a partir de ellas, una única serie **(serie enlazada)** en la que se puede apreciar la evolución del fenómeno en el periodo que resulta de unir los correspondientes a las distintas series.

El enlace de las distintas series de índices **se realiza expresándolas en la misma base.** Para ello, se utiliza, si es necesario, la información relativa a los valores del número índice en los periodos de tiempo que son comunes a las distintas series.

Ejemplo 4.10. Enlazar las dos series de índices de precios que se dan a continuación, obteniendo una única con base en 1999, que refleje la evolución porcentual del índice de precios en el periodo completo (Tabla 4.18).

Tabla 4.18. Índices de precios

Años	IP (Base 1999)	IP (Base 2003)
1999		
2000		
2001	100	
2002	105	
2003	112	100
2004	120	110
2005	130	123
2006		133
2007		140

Solución. Tenemos que calcular los índices de precios con base 1999 para los años 2004 a 2007. Para ello basta tener en cuenta la relación

$$I^t_{1999} = I^t_{2003} \cdot \frac{I^{2003}_{1999}}{100}$$

La serie enlazada, correspondiente al periodo 1999-2007, es la que se muestra en la Tabla 4.19.

Tabla 4.19. Índices de precios con base en 1999

Años	IP (Base 1999)
1999	100
2000	105
2001	112
2002	120
2003	130
2004	143
2005	159,9
2006	172,9
2007	182

Alternativamente, podríamos haber enlazado las series anteriores fijando la base en el año 2003. En ese caso, la relación que habría que aplicar para expresar los índices anteriores a 2003 en base 2003 sería

$$I^t_{2003} = I^t_{1999} \cdot \frac{100}{L^{2003}_{1999}}$$

y la serie enlazada es la siguiente (Tabla 4.20).

Tabla 4.20. Índices de precios con base en 2003

Años	IP (Base 2003)
1999	76,92
2000	80,77
2001	86,15
2002	92,31
2003	100
2004	110
2005	123
2006	133
2007	140

4.5. Deflación de una serie de valores monetarios. Índice implícito de precios

La **inflación (deflación)** es un fenómeno económico que se caracteriza por una **subida (caída)** generalizada y persistente de los precios monetarios de los bienes y servicios que se producen y se consumen en el país. Tanto la inflación como la deflación son fenómenos que alteran o modifican el poder adquisitivo del dinero, también llamado **poder de compra.**

Formalmente, para eliminar el efecto de la inflación en una serie de valores monetarios que son agregados de precios y cantidades para un conjunto de N bienes, es necesario dividir el valor de cada periodo por el índice de precios de Paasche correspondiente a ese periodo para dicho conjunto de bienes, expresado en tanto por uno. A dicha operación se le llama **deflación de la serie** y a la serie así obtenida se le denomina serie de valores a precios constantes. La terminología que se emplea es la siguiente:

- **Valor monetario, valor en términos corrientes o nominales en unidades monetarias del año, mes, etc.** para referirnos al valor sin eliminar el efecto de la inflación o variación de los precios.
- **Valor real, a precios constantes o en unidades monetarias del instante o periodo en el que el índice de precios tenga la base**, para referirnos al valor una vez eliminado el efecto de la inflación.

$$\frac{\text{Valor a precios corrientes (t)}}{\frac{\text{Precios de Paasche}_0^t}{100}} =$$

$$= \text{Valor a Precios constantes (t) en u.m.del periodo 0}$$

En España, la inflación se cuantifica generalmente entre comas, mediante el **Índice de Precios al Consumo (IPC)** que estudiaremos en el apartado siguiente. Se trata de un índice de Laspeyres y es el más utilizado para eliminar el efecto que la inflación tiene sobre las series de valores monetarios. La deflación permite la comparación de valores monetarios correspondientes a distintos años en términos de poder de compra o poder adquisitivo.

La relación entre ambos valores es en este caso

$$\text{Valor real}^t = \frac{\text{Valor corriente}^t}{\text{IPC}_{t_0}^t} \cdot 100$$

El valor real así obtenido viene expresado en unidades monetarias constantes del periodo t_0.

Ejemplo 4.11. En la Tabla 4.21 figura la evolución del consumo de un determinado producto, expresado en 10^3 € corrientes, durante el periodo 1999-2003, así como los correspondientes índices de precios

Tabla 4.21. Evolución del consumo

Años	Consumo (10^3 €)	IP
1999	80	107
2000	110	113
2001	90	122
2002	100	128
2003	120	136

Se pide:

a) Indique el incremento porcentual del consumo, tanto en términos corrientes como en términos constantes, entre 1999 y 2003.

b) Obtenga las variaciones interanuales del consumo, en términos corrientes, durante el periodo considerado.

Solución. A partir del consumo en términos corrientes y del índice de precios (Tabla 4.21), se obtiene el consumo en términos reales que se muestra en la Tabla 4.22.

Tabla 4.22. Consumo en términos nominales y en términos reales

Años	Consumo (10^3 €)	IP	Consumo (10^3 €) en T.R.
1999	80	107	74,766
2000	110	113	97,345
2001	90	122	73,770
2002	100	128	78,125
2003	120	136	88,235

a) En términos corrientes

$$TV_{1999\text{-}2003} = \frac{120}{80} \cdot 100 - 100 = 50\%$$

En términos constantes

$$TV_{1999\text{-}2003} = \frac{88{,}235}{74{,}766} \cdot 100 - 100 = 18{,}015\%$$

b) Las variaciones porcentuales interanuales las obtenemos a partir de los índices en cadena, que permiten comparar el dato de cada año con el del año inmediato anterior. Estos índices y las tasas, que se obtienen restando al índice 100, son los que se muestran en la Tabla 4.23.

Tabla 4.23. Índices en cadena y tasas de variación interanual

Años	$IC^{año}$	Tasas
1999	--	--
2000	137,5	37,5
2001	81,82	–18,18
2002	111,11	11,11
2003	120	20

4.6. Índice de precios al consumo (IPC) e índice de precios al consumo armonizado (IPCA)

El **Índice de Precios al Consumo (IPC)** es un índice complejo que permite estudiar la evolución y los cambios mensuales en el conjunto de los precios de los bienes y servicios que consumen las familias de un país.

La obtención del IPC requiere de dos operaciones distintas que son complementarias:

- Determinación cualitativa y cuantitativa de la cesta de la compra correspondiente a la familia media del país. Esta cesta contiene la lista de los bienes y servicios consumidos por la familia media durante un periodo de tiempo. En España el proceso para determinar la composición de la cesta de la compra y su estructura de ponderaciones utiliza como fuente fundamental de información la nueva Encuesta de Presupuestos Familiares (EPF), que desde el año 2006 sustituye a la Encuesta Continua de Presupuestos Familiares y es de periodicidad anual.

- Proceso de recogida periódica de precios al consumidor para los bienes y servicios que intervienen en el índice.

Actualmente, el periodo de referencia del **IPC español** es el año 2011; es decir, **es un índice con base en el año 2011.**

En el IPC se llama **periodo de referencia de los precios** al periodo con cuyos precios se comparan los precios corrientes, es decir, al periodo elegido para el cálculo de los índices elementales. Con la fórmula de cálculo empleada para el **IPC base 2011 (Laspeyres encadenado),** el periodo de referencia de los precios **varía cada año y es el mes de diciembre del año inmediatamente anterior al considerado.**

Según figura en la Nota de Prensa del INE sobre el IPC base 2011 **(http://www.ine.es/prensa/np701.pdf),** la estructura de ponderaciones de la base 2011 se ha elaborado a partir de la EPF como fuente principal de información. Además, se ha utilizado información de otras fuentes, tales como la evolución del consumo privado de la Contabilidad Nacional, la evolución de precios del IPC y otras fuentes de diferentes sectores.

Figura 4.1. Ponderaciones del IPC

Ponderaciones de grupos (tanto por cien)			
Grupo	2011	2012	%
01. Alimentos y bebidas no alcohólicas	12,16	18,26	0,6
02. Bebidas alcohólicas y tabaco	2,87	2,89	0,7
03. Vestido y calzado	8,59	8,34	–2,9
04. Vivienda	11,70	12,00	2,6
05. Menaje	6,84	6,67	–2,5
06. Medicina	3,21	3,14	–2,1
07. Transporte	14,74	15,16	2,9
08. Comunicaciones	3,98	3,85	–3,3
09. Ocio y cultura	7,64	7,54	–1,3
10. Enseñanza	1,38	1,42	2,8
11. Hoteles, cafés y restaurantes	11,52	11,46	–0,5
12. Otros bienes y servicios	9,37	9,26	–1,2
TOTAL	**100**	**100**	

Fuente: Instituto Nacional de Estadística (INE)

La cesta de la compra para el IPC con base 2011 está formada por **489** artículos distribuidos por grandes grupos. A modo de ejemplo, en la Figura 4.2 puede verse la configuración de la cesta de la compra para el IPC con base 2006, formada por 491 artículos.

Las aplicaciones del IPC son numerosas y de gran importancia en los ámbitos económico, jurídico y social. Entre ellas cabe destacar su utilización como medida de la inflación. También se aplica en la revisión de los contratos de arrendamiento de inmuebles, como referencia en la negociación salarial, en la fijación de las pensiones, en la actualización de las primas de seguros y otros tipos de contrato, y como deflactor en la Contabilidad Nacional (**http://www.ine.es/daco/daco43/metoipc06.pdf,** Metodología general IPC Base 2006, INE).

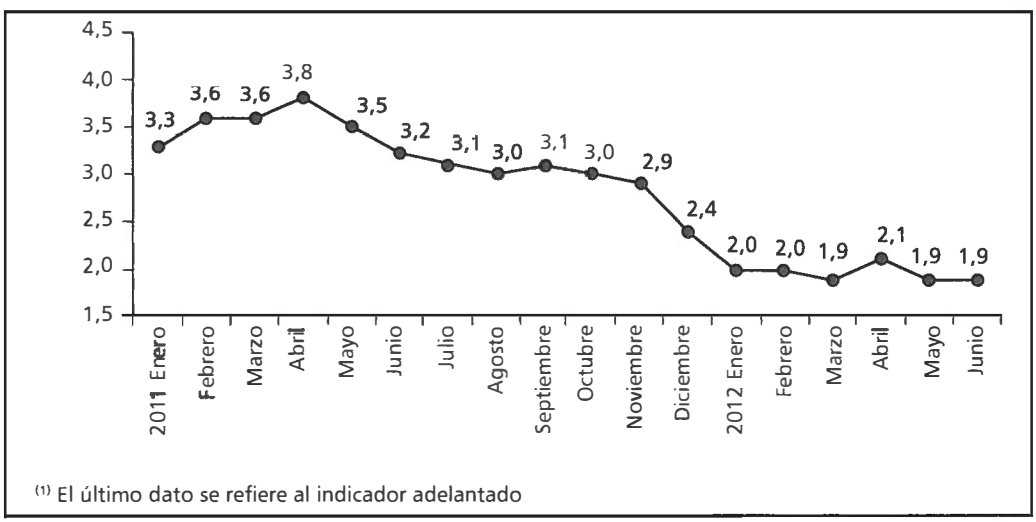
Figura 4.2. Distribución de la cesta de la compra según los grupos

Número de artículos IPC Base 2006	
Grupo	**Número de artículos**
01. Alimentos y bebidas no alcohólicas	176
02. Bebidas alcohólicas y tabaco	12
03. Vestido y calzado	67
04. Vivienda	18
05. Menaje	60
06. Medicina	13
07. Transporte	31
08. Comunicaciones	3
09. Ocio y cultura	43
10. Enseñanza	7
11. Hoteles, cafés y restaurantes	23
12. Otros bienes y servicios	38
TOTAL	**491**

Fuente: INE

En el Gráfico 4.1 se muestra la evolución de las tasas de variación interanuales del IPC que corresponden a los distintos meses tabulados.

El **Índice de Precios de Consumo Armonizado (IPCA)** es un indicador estadístico cuyo objetivo es proporcionar una medida común de la inflación que permita realizar comparaciones entre los países de la Unión Europea (UE), y entre estos y otros países que no pertenecen a la UE. Es, como el IPC, un índice de precios de Laspeyres encadenado y, en cada país, cubre las parcelas que superan el uno por mil del total de gasto de la cesta de la compra nacional (INE, Metodología IPCA). En cuanto a las ponderaciones de los precios en los distintos grupos, en la Figura 4.3 se ofrecen las correspondientes a 2012 para el IPCA.



(1) El último dato se refiere al indicador adelantado

Fuente: Nota de Prensa INE (junio 2012)

Gráfico 4.1. Evolución anual del IPC, Base 2011. Índice General

Figura 4.3. Ponderaciones del índice de Precios al Consumo Armonizado (IPCA)

Ponderaciones IPCA-2005=100 (conjunto nacional) por clases/subclases y periodo Unidades tanto por mil	
	Ponderación IPCA 2012
Alimentos y bebidas no alcohólicas (01)	184.326
Bebidas alcohólicas y tabaco (02)	29.469
Artículos de vestir y calzado (03)	84.457
Vivienda, agua, electricidad, gas y otros combustibles (04)	123.714
Mobiliario. enseres domésticos y gastos corrientes de conservación de la vivienda (05)	66.576
Sanidad (06)	31.607
Transporte (07)	144.565
Comunicaciones (08)	38.089
Ocio y cultura (09)	76.733
Educación (10)	14.345
Restaurantes y hoteles (11)	138.19
Otros bienes y servicios (12)	67.927

Fuente: INE

4.7. Índices bursátiles

Un índice bursátil es un tipo de índice que refleja el cambio en el tiempo del valor de un conjunto de acciones de empresas que cotizan en bolsa. Los más habituales son los índices de valor calculados mediante la fórmula:

$$v_0^t = \frac{p_{1t} \cdot q_{1t} + p_{2t} \cdot q_{2t} + \ldots + p_{Nt} \cdot q_{Nt}}{p_{10} \cdot q_{10} + p_{20} \cdot q_{20} + \ldots + p_{N0} \cdot q_{N0}} \cdot 100 = \frac{\sum_{i=1}^{i=N} p_{it} \cdot q_{it}}{\sum_{i=1}^{i=N} p_{io} \cdot q_{io}} \cdot 100,$$

donde,

- p_{it} es el precio o cotización de las acciones de la empresa i en tiempo t y p_{io} el correspondiente en tiempo 0.
- q_{it} es el número de acciones de la empresa i en tiempo t y el correspondiente en tiempo 0.

Los principales índices bursátiles en España son: el Índice General de la Bolsa de Madrid (IGBM) y el **IBEX 35**.

El IGBM[1]

Este índice no está integrado por un número fijo de empresas, sino que su número podrá variar semestralmente, en función de que las distintas compañías cumplan o no un conjunto de requisitos estipulados previamente.

[1] Información en **http://www.bolsamadrid.es/esp/contenido.asp?enlace=indices/igbm/sectores.htm**.

La ponderación de cada valor se calcula en función de la capitalización bursátil del último día del semestre anterior, entendiendo por tal el número de acciones de la empresa por su cotización.

Para el cálculo diario del índice, se parte de los índices diarios de cada valor perteneciente al IGBM. El Índice Valor, en base 100, es el cociente entre la cotización del día de cálculo y el llamado precio de referencia, que es el precio de cierre del día anterior a la modificación de la composición del IGBM, 31/12 y el 30/06 en las reuniones ordinarias del Comité encargado de la elaboración y actualización de este índice.

A continuación, se obtienen los Índices Subsectoriales como la suma de los Índices Valores ponderados por su peso en el Subsector. A partir de estos Índices Subsectoriales se obtiene el Índice del Sector, igualmente como la suma de los Índices Subsectoriales ponderando por el peso de cada Subsector en el Sector.

El IGBM se obtiene como la suma ponderada, por el peso de cada sector en el IGBM, de los Índices Sectoriales.

El IBEX 35[2]

El Índice se compone de los 35 valores cotizados en el Sistema de Interconexión Bursátil de las cuatro Bolsas Españolas (Barcelona, Bilbao, Madrid y Valencia) que sean los más líquidos durante el periodo de control; es decir, en el intervalo de los seis meses anteriores a la fecha de la revisión. Para ver fórmula de cálculo y otra información relativa al mismo **http://www.ibex35.com/.**

4.8. Ejercicios

Ejercicio 4.1. En la Tabla 4.24 figura información relativa al número de viajeros según la categoría del establecimiento (número de estrellas de oro).

Tabla 4.24. Información relativa al n.º de viajeros según categoría del establecimiento

	Hoteles: estrellas de oro				
	5 Estrellas	4 Estrellas	3 Estrellas	2 Estrellas	1 Estrella
	Índices	**Viajeros (Miles)**	**TV(ref. 2005)**	**Viajeros (Miles)**	**Índices**
2003	100	20.490,5	-8,4	6.626,3	100
2004	117,1	23.618,0	-5,3	6.855,0	101,6
2005	130,9	25.548,3	0	6.992,9	101,5
2006	165,3	31.385,6	17,9	7.398,8	107,8

Fuente: Encuesta de ocupación hotelera. INE

Se pide:

a) ¿Cuál es la tasa de variación del número de turistas en hoteles de 4 estrellas en el periodo 2004-2006?

b) ¿Cuál es la variación porcentual interanual del número de turistas en hoteles de 5 estrellas en el año 2005?

c) Obtenga la serie de números índices simples correspondiente al número de viajeros en hoteles de 3 estrellas, tomando como referencia el dato del año 2003.

[2] Información en **http://www.ibex35.com/.**

d) Obtenga las series de números índices simples, tomando como referencia el dato del año 2003, para el número de viajeros en hoteles de 4 y 2 estrellas, respectivamente.

e) Construya una serie de índices complejos no ponderados, mediante el método de la media aritmética simple, que refleje la evolución porcentual, con respecto al año 2003, del número total de viajeros alojados en hoteles de Andalucía en los años del periodo considerado. Una vez elaborada, proporcione una estimación de la variación porcentual interanual del número de viajeros alojados en hoteles de Andalucía en el año 2004.

f) Basándose en la serie de índices complejos calculada en el apartado anterior, estime las tasas de variación interanual correspondientes al número de viajeros alojados en hoteles de Andalucía para los años del periodo 2004-2006.

Ejercicio 4.2. En la Tabla 4.25 figuran los datos anuales para el periodo 1999-2006 relativos al gasto medio diario de dos familias (A y B) residentes en Andalucía (expresado en euros). En dicha tabla figura también la serie para esos años del IPC de Andalucía.

Tabla 4.25. Datos anuales para el período 1999-2006 relativos al gasto medio diario de dos familias residentes en Andalucia

	Gasto medio diario (A)	Gasto medio diario (B)	IPC Andalucía
1999	34,2	43,2	93,6
2000	35,1	46,4	96,5
2001	37,7	47,5	100,0
2002	41,8	47,4	103,4
2003	42,3	52,0	106,5
2004	47,4	52,5	109,7
2005	49,4	55,2	113,3
2006	49,6	57,1	117,1

Se pide

a) Calcule, para el periodo considerado, la tasa de variación del gasto medio diario en términos corrientes de la familia A y compárela con la correspondiente a la familia B.

b) Obtenga la serie del gasto medio diario de la familia B, expresada en euros constantes del año 1999.

c) Calcule para el periodo considerado, la tasa de variación del gasto medio diario en términos constantes de la familia A. Interprete el resultado obtenido.

Ejercicio 4.3. En la Tabla 4.26 se detallan, en términos corrientes, los ingresos anuales de un supermercado durante el periodo 2001- 2005, así como el IPC facilitado por el INE:

Tabla 4.26. Ingresos anuales de un supermercado e IPC (2001-2005)

Años	Ingresos (10^4 euros)	IPC
2001	40	100
2002	45	103,54
2003	44	106,68
2004	50	109,93
2005	55	113,63

Con esta información, se pide:

a) Obtenga los índices en cadena correspondientes a los ingresos anuales en términos corrientes y dé las tasas de variación interanuales de los mismos para los años del periodo 2002-2005.
b) Obtenga la tasa media de variación interanual de los ingresos en el periodo considerado, en términos corrientes.
c) Obtenga la serie de los ingresos anuales para el periodo considerado en términos reales o constantes del año 2003.
d) Exprese la variación porcentual experimentada por los ingresos en el periodo 2002-2005, tanto en términos corrientes como en términos constantes. Comente los resultados obtenidos.

Ejercicio 4.4. Se dispone de la siguiente información (Tabla 4.27) del Índice de Precios al Consumo en España (I.N.E.).

Tabla 4.27. Índice de Precios al Consumo en España (2000-2001)

Año 2000	IPC	Año 2001	IPC
Enero	94,8	**Enero**	98,3
Febrero	94,6	**Febrero**	98,3
Marzo	95,4	**Marzo**	99,1
Abril	96,1	**Abril**	99,9
Mayo	96,3	**Mayo**	100,3
Junio	96,5	**Junio**	100,5
Julio	96,1	**Julio**	99,8
Agosto	96,4	**Agosto**	99,9
Septiembre	97,1	**Septiembre**	100,4
Octubre	98,0	**Octubre**	100,9
Noviembre	98,5	**Noviembre**	101,1
Diciembre	98,7	**Diciembre**	101,4

a) Calcule la tasa de variación interanual de los precios en el mes de junio de 2001.
b) Calcule la tasa de variación intermensual de la inflación en el mes de septiembre de 2001.
c) Calcule la inflación acumulada en el mes de octubre de 2001.
d) Actualice a fecha 31 de diciembre de 2001 el precio de un producto que en mayo de 2000 ascendía a 130 euros.

Ejercicio 4.5. En la Tabla 4.28 figuran los índices en cadena que proporcionan la evolución de los ingresos anuales de una empresa.

Tabla 4.28. Índices en cadena de una empresa (2006-2010)

Años	Índices en cadena
2006	---
2007	102,50
2008	98,60
2009	105,00
2010	108,30

Se pide:

a) Obtenga la variación porcentual interanual de los ingresos de dicha empresa en 2008.

b) Obtenga la tasa de variación de los ingresos en el periodo 2006-2010.

c) Calcule la tasa media de variación interanual acumulativa de los ingresos para el periodo 2006-2010.

d) Sabiendo que los precios en el año 2010 crecieron un 2,98% con respecto al año anterior, ¿cuál fue, en términos reales, la variación porcentual interanual de los ingresos de la empresa en el año 2010?

Ejercicio 4.6. Teniendo en cuenta la información que, sobre la evolución mensual del Índice de precios hoteleros, figura en la Tabla 4.29, responda a las cuestiones que se le plantean.

Tabla 4.29. Evaluación mensual del Índice de precios hoteleros

	Índice de precios hoteleros					
	Total categorías	Cinco estrellas de oro	Cuatro estrellas de oro	Tres estrellas de oro	Dos estrellas de oro	Una estrella de oro
2006-09	115,7	109,6	111,5	117,8	121,2	125,7
2006-10	112,0	110,7	108,7	112,1	118,5	124,8
2006-11	111,5	109,0	109,1	110,4	121,4	128,2
2006-12	113,0	113,4	110,6	111,2	123,7	129,4
,,,	,,,	,,,	,,,	,,,	,,,	,,,
2007-09	119,3	112,2	114,1	122,2	127,4	134,2
2007-10	113,9	113,6	¿.....?	113,7	124,4	131,3
2007-11	113,4	108,2	111,1	113,0	125,7	134,2

Fuente: Encuesta de Ocupación en alojamientos turísticos, INE

a) Calcule, para los hoteles de cinco estrellas, la inflación acumulada en noviembre de 2007.

b) Calcule el incremento porcentual de los precios en los hoteles de tres estrellas en el periodo octubre de 2006-noviembre de 2007.

c) Sabiendo que el incremento interanual de los precios en hoteles de cuatro estrellas en el mes de octubre de 2007 fue de un 1,01% calcule el índice de precios en dicho mes.

d) Calcule la variación porcentual que experimentaron los precios hoteleros entre septiembre y noviembre de 2007.

e) Obtenga, mediante el método de la media aritmética simple, un índice complejo para el mes de noviembre de 2007 con base en septiembre del mismo año, que refleje la variación de los precios hoteleros en su conjunto. Compare la variación porcentual de los precios que se deriva de este índice con la obtenida en el apartado anterior.

CAPÍTULO 5

Análisis descriptivo de las componentes de una serie temporal

Una serie temporal es una sucesión de observaciones temporales de una variable. Al ser el tiempo una variable que conlleva un orden natural, las observaciones deberán escribirse siguiendo dicho orden. A las **series temporales** también se las denomina **series cronológicas o históricas.**

Tabla 5.1. Serie temporal

t	y_t
1	y_1
2	y_2
.	.
.	.
.	.
t	y_t
.	.
.	.
.	.
T-1	y_{T-1}
T	y_T

Ejemplos

- Número mensual de parados registrados en las Oficinas de Empleo en España en el periodo 1998-2011.
- Serie trimestral del coste salarial total por hora de trabajo.
- Número mensual de hipotecas constituidas para fincas urbanas.
- Número trimestral de pernoctaciones en establecimientos hoteleros de la provincia de Málaga en el periodo 2000-2009.

Entre los **objetivos del análisis estadístico** de una serie temporal destacamos:

- La **descripción** de la evolución pasada de la variable.
- La **predicción** de los valores futuros de la misma a corto y medio plazo.

La forma más simple de comenzar el análisis es mediante la representación gráfica de la serie. En los ejes de coordenadas, reservamos el eje de abscisas para representar

los distintos instantes o periodos de tiempo a los que se refieren las observaciones **(t)** y el eje de ordenadas para los valores de la variable en dichos instantes o periodos **(y_t)**. La representación gráfica de la serie se obtiene al unir mediante trazos rectos los puntos **(t, y_t)** correspondientes a observaciones consecutivas en el tiempo.

Ejemplo 5.1. La representación gráfica de la serie temporal correspondiente a los miles de asalariados en el sector público en España en el periodo 1996-2009 (Tabla 5.2) es la que se presenta en el Gráfico 5.1.

Tabla 5.2. Asalariados en el sector público en España (miles)

Años	Asalariados en el sector público (miles)
1996	2.322,8
1997	2.363,5
1998	2.328,2
1999	2.357,7
2000	2.441,475
2001	2.505,7
2002	2.591,6
2003	2.707,7
2004	2.800,375
2005	2.864,15
2006	2.882,175
2007	2.913,025
2008	2.958,65
2009	3.062,05

Fuente: Boletín Mensual de Estadística (INE)

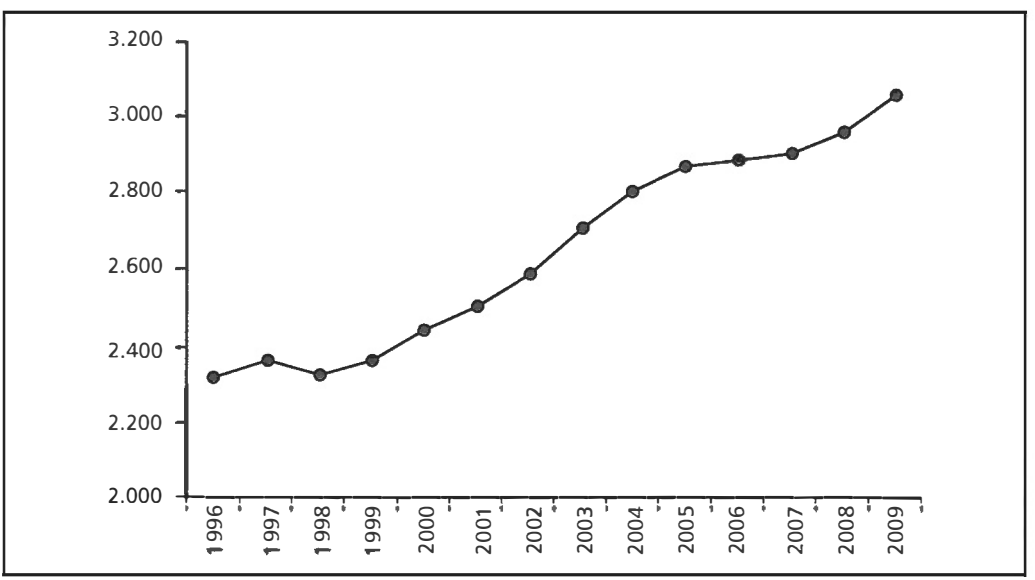


Gráfico 5.1. Miles de asalariados en el sector público

La simple inspección de la gráfica de una serie temporal permite concluir que la variable, en la mayor parte de los casos, no permanece constante, sino que sufre variaciones a lo largo del tiempo. Dependiendo de cómo sean estas variaciones, las series temporales pueden agruparse en dos categorías básicas: **evolutivas y estacionarias.** En el primer caso el nivel medio de la variable cambia sustancialmente a lo largo del tiempo, mientras que en el segundo, el mismo permanece prácticamente constante.

Ejemplo 5.2.

- **Serie Evolutiva.** La serie correspondiente al ejemplo anterior.
- **Serie Estacionaria.** Serie de las temperaturas máximas anuales en Madrid durante el periodo 1997-2008 (Tabla 5.3 y Gráfico 5.2).

Tabla 5.3. Temperaturas máximas anuales en Madrid (Retiro)

Años	Temperatura Máxima en Madrid (Retiro)
1997	36,5
1998	36,5
1999	37,3
2000	36,4
2001	36,4
2002	36,3
2003	38,6
2004	38,4
2005	38
2006	37,6
2007	36,9
2008	38

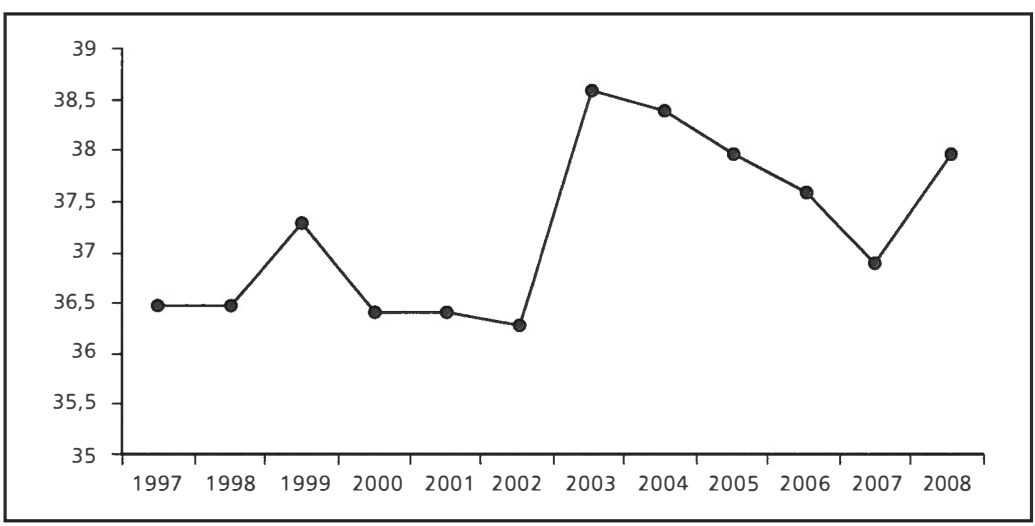


Gráfico 5.2. Temperatura máxima en Madrid

La descripción analítica de la serie se puede abordar siguiendo distintos procedimientos o enfoques entre los que destacan:

- **El enfoque causal:** basado en la consideración de que las variaciones en la serie son consecuencia de las que se producen en otras variables de las que aquella depende. El objetivo será determinar tales variables y expresar mediante un modelo la forma en la que las mismas influyen en la serie.
- **El enfoque clásico:** basado en la idea de que en toda serie empírica hay presentes unas componentes teóricas o hipotéticas, que se pueden aislar para conocer su evolución temporal y de manera tal que las variaciones que a lo largo del tiempo experimenta la variable son el resultado de la interacción entre sus componentes y de la evolución individual de las mismas. **Este será el enfoque que nosotros adoptaremos para el análisis.**

Componentes no observables de una serie temporal

- **Tendencia** o tendencia secular **(T_t).**
- **Fluctuaciones estacionales** o componente estacional **(E_t).**
- **Fluctuaciones cíclicas** o componente cíclica **(C_t).**
- **Variaciones accidentales** o componente irregular **(I_t).**

Tendencia (T_t). Refleja la evolución de la serie a largo plazo.

- Para apreciarla es preciso que la serie temporal abarque bastantes años.
- La tendencia expresa si, a largo plazo, la variable adopta un movimiento persistente (creciente, decreciente o estacionario) que obedezca a una ley determinada. Expresa, por lo tanto, si la serie es estacionaria o evolutiva.

Ejemplo 5.3. En el Gráfico 5.3 se ha representado la serie temporal de las pernoctaciones de extranjeros en Málaga **(Y_t).** La tendencia (variable no observable), que se ha representado en el gráfico mediante una línea recta, es creciente.

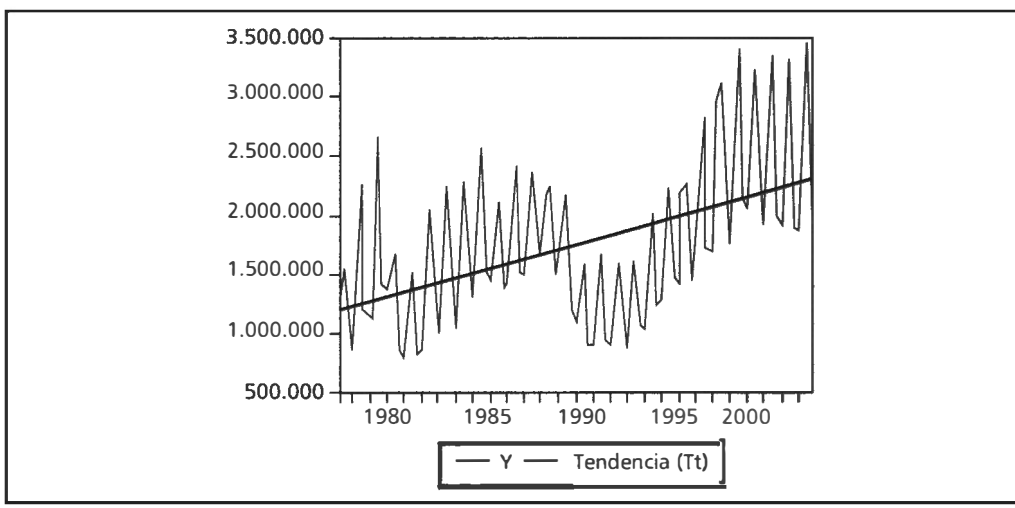


Gráfico 5.3. Pernoctaciones de extranjeros en Málaga (Tendencia)

Fluctuaciones estacionales (E_t). Son movimientos a corto plazo que se repiten periódicamente con periodicidad igual o inferior al año.

- Generalmente se deben a factores climatológicos, a las costumbres de la población, etc.
- Solo se presentan en series de observaciones con periodicidad inferior al año (mensual, trimestral, etc.)
- Son comunes en series de temperaturas, consumo de algunos productos, precios de productos agrícolas y ganaderos y, por supuesto, en las relativas al turismo.

Ejemplo 5.4. Volviendo al ejemplo de las pernoctaciones de extranjeros en Málaga **(Y_t),** en el Gráfico 5.4, y centrándonos en un pequeño tramo de la serie (2000-2003), podemos apreciar claramente las fluctuaciones estacionales: sistemáticamente, el segundo y tercer trimestre muestran un mayor volumen de pernoctaciones que el primero y el cuarto.

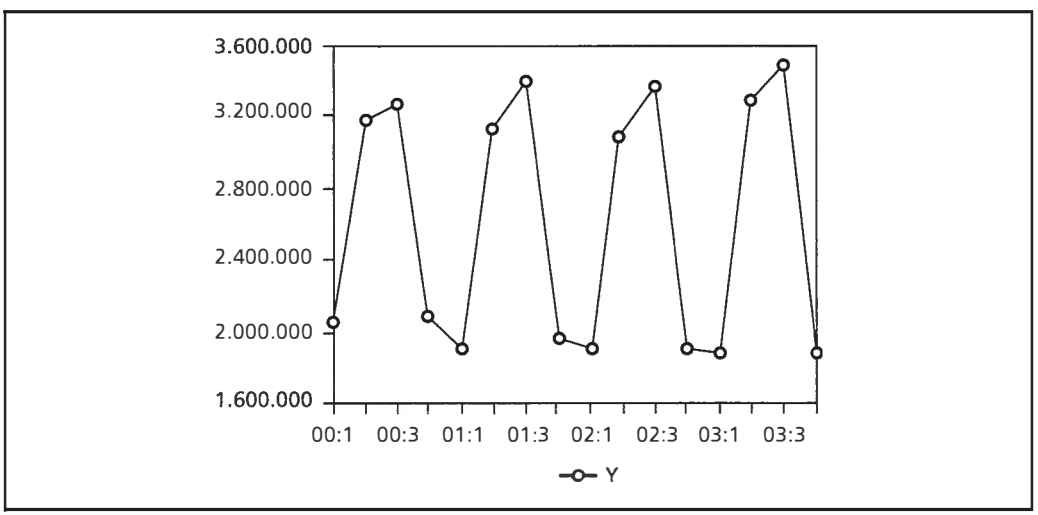


Gráfico 5.4. Pernoctaciones de extranjeros en Málaga (Componente Estacional)

Fluctuaciones cíclicas (C_t). Son movimientos ondulatorios a medio plazo que se repiten, pero no tan regularmente como lo hacen las variaciones estacionales.

- Esta componente tiene un marcado carácter económico por ser, en general, consecuencia de la alternancia entre periodos de prosperidad y depresión propia del sistema económico.
- Es la componente más difícil de aislar. Por ello, en muchas ocasiones, si se detecta y no es posible aislarla, se filtra la serie para eliminar las variaciones estacionales, cuando estas existen, y a la componente resultante se le denomina Tendencia-Ciclo.

Ejemplo 5.5. En el siguiente Gráfico 5.5 podemos apreciar las componentes Tendencia y Tendencia-Ciclo de las pernoctaciones de extranjeros en Málaga **(Y_t).**

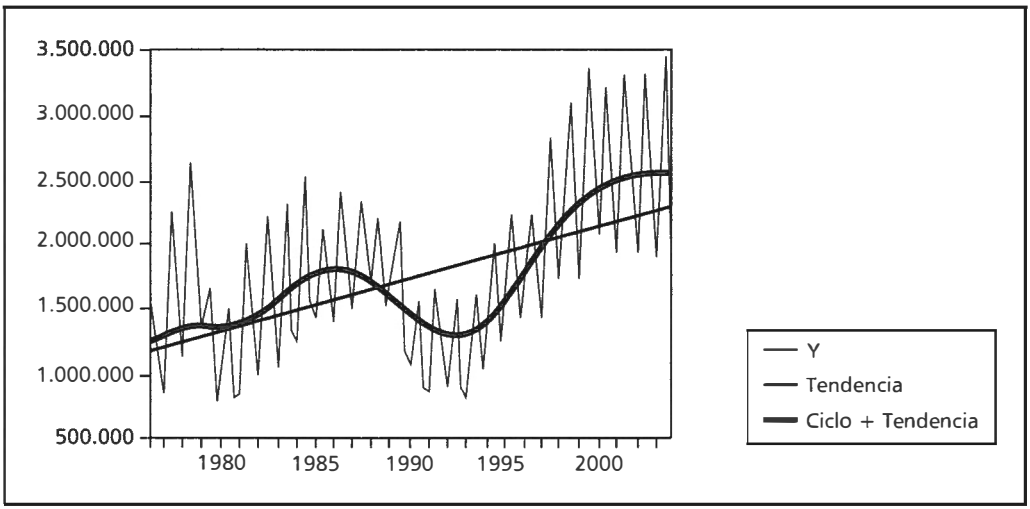


Gráfico 5.5. Pernoctaciones de extranjeros en Málaga (Tendencia y Tendencia-Ciclo)

Variaciones accidentales o irregulares (I_t). La componente irregular de una serie recoge las variaciones que quedan fuera del análisis en las tres componentes anteriores. Incluye tanto los movimientos provocados por factores fortuitos con un efecto significativo (efectos de una huelga, de una catástrofe natural...), como movimientos de menor cuantía que pueden estar ocasionados por múltiples factores no identificados.

Ejemplo 5.6. En el Gráfico 5.6 podemos ver la componente «irregular» de las pernoctaciones de extranjeros en Málaga **(Y_t).**

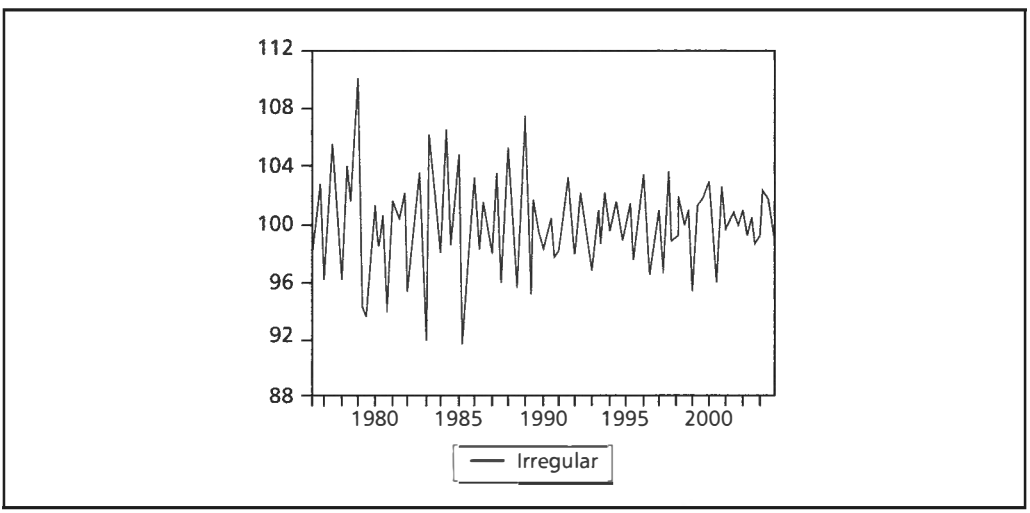


Gráfico 5.6. Pernoctaciones de extranjeros en Málaga (Componente Irregular)

En una serie temporal pueden estar presentes todas las componentes o solo algunas de ellas. Por ejemplo, en una serie de datos anuales no hay fluctuaciones estacionales.

Así, si sumamos los datos trimestrales de pernoctaciones de cada año, podemos obtener una serie anual en la que ya no hay variaciones estacionales (Gráfico 5.7).

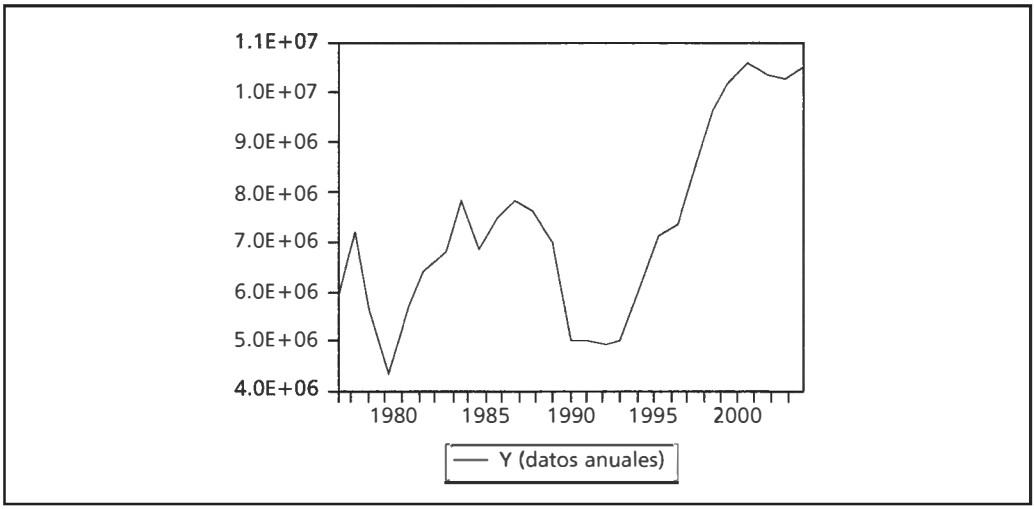


Gráfico 5.7. Pernoctaciones de extranjeros en Málaga (datos anuales)

Finalmente, un tema relevante es el de la elección de la forma en la que las componentes se combinan entre sí para dar lugar a la serie original **(Y_t).**

Las hipótesis más frecuentes son:

- **Hipótesis aditiva:** $Y_t = T_t + E_t + C_t + I_t$
- **Hipótesis multiplicativa:** $Y_t = T_t \cdot E_t \cdot C_t \cdot I_t$
- **Hipótesis mixta:** $Y_t = (T_t + C_t + I_t) \cdot E_t$

Bajo la hipótesis aditiva, las cuatro componentes están expresadas en la misma unidad de medida que la serie Y_t. En cambio, bajo la hipótesis multiplicativa, solo una componente (generalmente la tendencia) está expresada en la unidad de la serie y las restantes son adimensionales y se expresan en tanto por uno o porcentaje respecto a la tendencia. En el caso de la hipótesis mixta solo la estacionalidad es adimensional.

En muchas ocasiones, el análisis de una serie temporal está encaminado al estudio de solo una de las componentes; por ejemplo, la tendencia, y, para ello, es necesario estimarla, eliminando previamente las restantes componentes que en la serie estén presentes. Por ello, es necesario estudiar formas de aislar las componentes de una serie, así como formas de «filtrar» la serie para eliminar ciertas componentes. Las componentes que analizaremos a continuación son la tendencia y las variaciones estacionales.

5.1. Estimación de la tendencia lineal

La regresión lineal, explicada con detalle en el Capítulo 3, es uno de los métodos que se pueden emplear para la estimación de la tendencia. En adelante, supondremos que dicha tendencia es lineal, esto es, sigue una línea recta (creciente o decreciente). En el Gráfico 5.8 se representan las pernoctaciones de extranjeros en Málaga (datos anuales) junto a la **tendencia lineal estimada.**

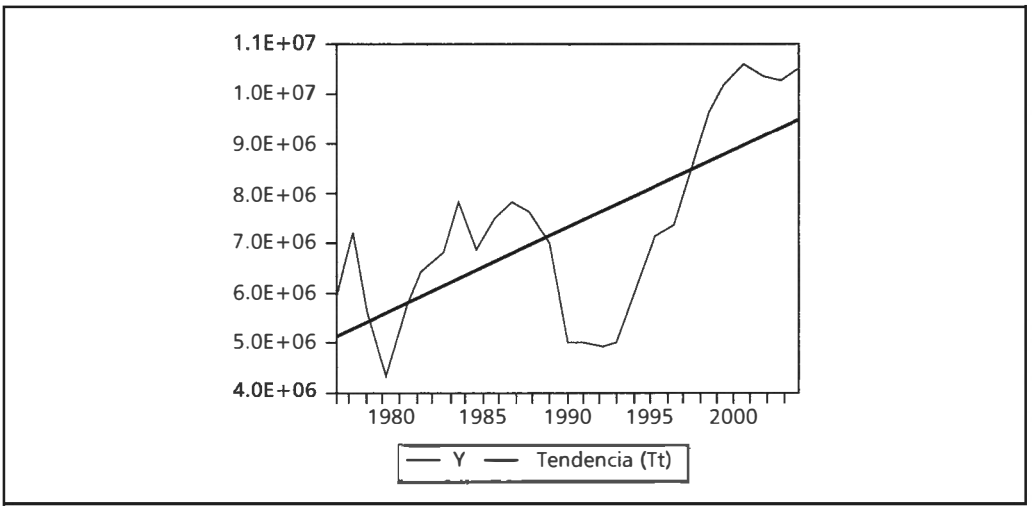


Gráfico 5.8. Pernoctaciones de extranjeros en Málaga

Pero, ¿cómo estimamos la recta que va a representar a la tendencia de la serie?

Distinguiremos dos casos:

- **Caso 1.** En la serie original (y_t) no se observan variaciones estacionales, ya sea porque las observaciones se toman con periodicidad anual o superior o porque la variable a la que la serie se refiere no está afectada por este tipo de movimientos.
- **Caso 2.** En la serie original (y_t) están presentes las variaciones estacionales.

5.1.1. Estimación de la tendencia lineal cuando no hay estacionalidad

El problema es formalmente análogo al de la regresión lineal entre dos variables que ya hemos estudiado en el Capítulo 3. Suponemos que la serie original (y_t) puede descomponerse en la tendencia (y_t^*), que supondremos que es una función lineal de **t**, y en un término residual (e_t) que, bajo la hipótesis de integración aditiva de las componentes, englobará a las variaciones cíclicas y a las irregulares o residuales. Esto es,

$$\left.\begin{array}{l} \textbf{Hipótesis 1: } y_t = T_t + C_t + I_t \\ \textbf{Hipótesis 2: } y_t = y_t^* + e_t, \text{ siendo } y_t^* = a + bt \\ \textbf{Hipótesis 3: } T_t = y_t^* \end{array}\right\} \quad y_t = \underbrace{a + bt}_{\text{Tendencia}} + \overbrace{e_t}^{C_t + I_t}$$

Observación. Hemos de señalar que aunque, así visto, el problema es formalmente análogo al de la regresión lineal, donde el papel de variable explicativa **(X)** ahora lo hace **t**, existe una diferencia conceptual importante: entre las variables **t** e y_t no existe dependencia causal, lo que sí ocurría en el contexto de la regresión.

La **estimación de los parámetros a y b** que determinan la tendencia lineal, basada en el criterio de mínimos cuadrados ordinarios, resulta ser, como ya vimos en el Capítulo 3.

$$\begin{cases} b = \dfrac{S_{ty_t}}{S_t^2} = \dfrac{\dfrac{1}{T}\sum_{t=1}^{T} t \cdot y_t - \bar{t} \cdot \bar{y}_t}{\dfrac{1}{T}\sum_{t=1}^{T} t^2 - (\bar{t})^2} \\ a = \bar{y}_t - b\bar{t} \end{cases}$$

La **interpretación de esos coeficientes** es la misma que en regresión:

- **a** es la **ordenada en el origen** o **valor estimado para la tendencia cuando t=0.** Se interpreta como el valor medio de la variable en ese instante.
- **b** es la **pendiente de la tendencia lineal** y se interpreta como la **variación que, por término medio, experimenta la variable por un incremento unitario en el tiempo.**

La **medida de bondad del ajuste de la tendencia lineal** es (como vimos en el Capítulo 3) el coeficiente de determinación, R^2, que podemos calcular a partir de su relación con el coeficiente de correlación lineal entre las variables **t** e y_t; es decir

$$R^2 = r_{ty_t}^2 = \frac{S_{ty_t}^2}{S_t^2 \cdot S_{y_t}^2}$$

o mediante su relación con las varianzas de y_t, y_t^* y e_t, es decir,

$$R^2 = \frac{S_{y_t^*}^2}{S_{y_t}^2} = 1 - \frac{S_{e_t}^2}{S_{y_t}^2}$$

R^2 representa, por lo tanto, la proporción de las variaciones de la serie que quedan englobadas en la tendencia a través de esa especificación lineal, mientras que el cociente

$$\frac{S_{e_t}^2}{S_{y_t}^2}$$

representa al resto; es decir, la proporción de las variaciones en la serie debidas a las variaciones cíclicas, a las irregulares o a la parte de tendencia que no queda recogida por el modelo lineal. Basándonos en lo anterior diremos que la tendencia lineal, así obtenida, es tanto mejor cuanto más próximo a **1** sea su coeficiente de determinación, R^2.

La recta estimada, cuando se ajusta bien a la serie, se puede utilizar para describir el comportamiento de la variable a largo plazo y para efectuar predicciones sobre la variable a partir de las predicciones de la tendencia.

Ejemplo 5.7. La Tabla 5.4 contiene los datos anuales relativos al número de asalariados en el sector público en España, en millones de personas, para los años del periodo 1996-2009. Se pide:

a) Represente gráficamente la serie original.
b) Estime, utilizando el método de mínimos cuadrados ordinarios, la tendencia lineal de la serie de asalariados en el sector público y explique el significado de los parámetros.

c) Represente en el mismo gráfico la serie original y la tendencia ajustada.
d) Obtenga una medida del grado de ajuste de la tendencia lineal estimada a la serie original.
e) Realice una predicción sobre el número de asalariados en España en el sector público en el año 2010 y diga qué fiabilidad le merece.

Tabla 5.4. Asalariados en el sector público

Años	Asalariados en el sector público (10^6)
1996	2,32
1997	2,36
1998	2,33
1999	2,36
2000	2,44
2001	2,51
2002	2,59
2003	2,71
2004	2,8
2005	2,86
2006	2,88
2007	2,91
2008	2,96
2009	3,06

Solución

a) La representación gráfica de esta serie es la que se muestra en el siguiente gráfico.

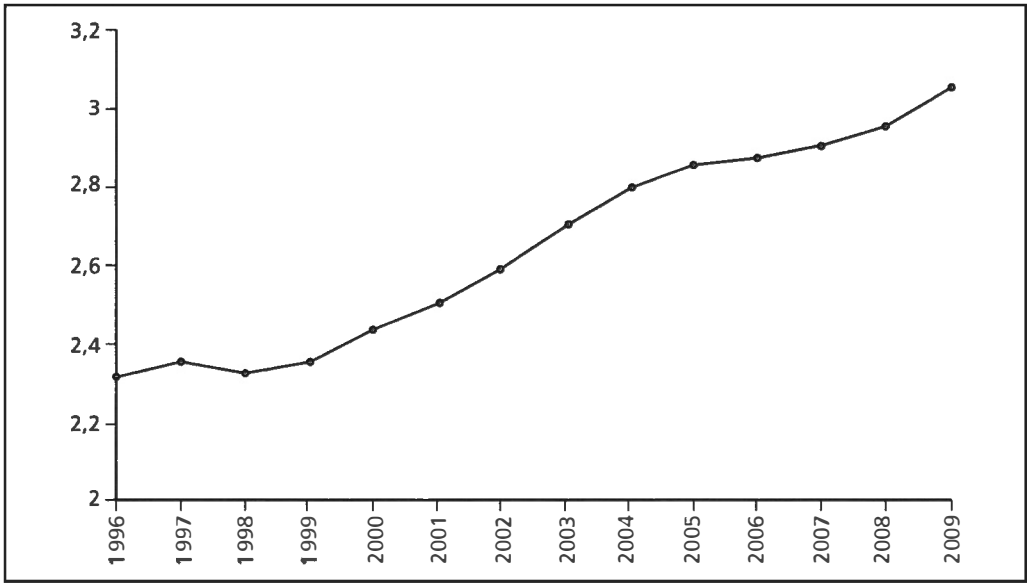


Gráfico 5.9. Millones de asalariados en el sector público

b) El modelo es

$$Y_t = Y_t^* + e_t = a + b\,t + e_t$$

Para estimar los parámetros **a** y **b** a partir de los datos, utilizamos las fórmulas

$$\begin{cases} b = \dfrac{S_{ty_t}}{S_t^2} \\ a = \bar{y}_t - b \cdot \bar{t} \end{cases}$$

Para obtener los valores de: $\bar{t}$, $\bar{y}_t$, S_{ty_t} y S_t^2, que son necesarios para la determinación de los parámetros «a» y «b», completamos la Tabla 5.4 con las columnas necesarias para los cálculos intermedios (Tabla 5.5); es decir, una columna para t^2, otra para $t \cdot y_t$ y una última para $(y_t)^2$. Basándonos en dicha tabla, obtenemos

$$\bar{t} = \frac{28.035}{14} = 2.002,5$$

$$S_t^2 = \frac{\sum t^2}{14} - \bar{t}^2 = \frac{56.140.315}{14} - 2.002,5^2 = 16,25$$

$$\bar{y}_t = \frac{37,09}{14} = 2,6493$$

$$S_{ty_t} = \frac{\sum t \cdot y_t}{14} - \bar{t} \cdot \bar{y}_t = \frac{74.286,81}{14} - 2.002,5 \cdot 2,6493 = 0,9775$$

Tabla 5.5. Millones de asalariados en el sector público (cálculos)

t	y_t	t^2	y_t^2	$t \cdot y_t$
1996	2,32	3.984.016	5,3824	4.630,72
1997	2,36	3.988.009	5,5696	4.712,92
1998	2,33	3.992.004	5,4289	4.655,34
1999	2,36	3.996.001	5,5696	4.717,64
2000	2,44	4.000.000	5,9536	4.880
2001	2,51	4.004.001	6,3001	5.022,51
2002	2,59	4.008.004	6,7081	5.185,18
2003	2,71	4.012.009	7,3441	5.428,13
2004	2,8	4.016.016	7,84	5.611,2
2005	2,86	4.020.025	8,1796	5.734,3
2006	2,88	4.024.036	8,2944	5.777,28
2007	2,91	4.028.049	8,4681	5.840,37
2008	2,96	4.032.064	8,7616	5.943,68
2009	3,06	4.036.081	9,3636	6.147,54
Totales **28.035**	**37,09**	**56.140.315**	**99,1637**	**74.286,81**

A partir de los cálculos anteriores, concluimos que el valor del parámetro «b» es:

$$b = \frac{S_{ty_t}}{S_t^2} = \frac{0,9775}{16,25} = 0,06$$

y, teniendo en cuenta que

$$a = \bar{y}_t - b \cdot \bar{t}$$

obtenemos que el valor del parámetro «a» es

$$a = \bar{y}_t - b \cdot \bar{t} = 2{,}6493 - 0{,}06 \cdot 2.002{,}5 = -117{,}5$$

Por lo tanto, la tendencia lineal estimada es

$$y_t^* = -117{,}5 + 0{,}06 \cdot t$$

o.t. = 0; u.t. = 1 año

donde **o.t. = 0 y u.t. = 1 año** indican que el origen del tiempo se ha tomado en el año cero y que la unidad temporal con la que se trabaja es el año.
El **significado** matemático del parámetro **«b»** es el de pendiente de la recta. Es decir, el incremento que experimenta **Y*** cuando **t** aumenta en una unidad. En este caso, el incremento es de 0,06 millones por cada unidad de tiempo.
En cuanto al significado económico, esto es, atendiendo a lo que representa la variable, según la tendencia estimada, el número de asalariados en el sector público en el periodo temporal observado aumenta cada año 0,06 millones, por término medio.
El **significado** matemático del parámetro **«a»** es el de valor de Y_t^* cuando **t=0,** es decir, el punto de corte de la recta con el eje de ordenadas. En nuestro caso, $Y_0^* = -117{,}5$. En el contexto del enunciado no admite una interpretación, dado que no tiene sentido hablar de un número negativo de asalariados en el sector público.

c) Como medida del grado de ajuste de la tendencia estimada calculamos el coeficiente de determinación, R^2. En nuestro caso, con los cálculos ya realizados la forma más simple de obtenerlo es a partir de su relación con el coeficiente de correlación lineal. Esto es

$$R^2 = \frac{S_{tY_t}^2}{S_t^2 \cdot S_{Y_t}^2} = \frac{\left(\sum \frac{t \cdot Y_t}{T} - \bar{t} \cdot \bar{Y}_t\right)^2}{\left(\frac{\sum t^2}{T} - \bar{t}^2\right)\left(\frac{\sum Y_t^2}{T} - \bar{Y}_t^2\right)}$$

Como

$$S_{t,Y_t} = 1{,}578 \qquad S_t^2 = 16{,}25$$

y

$$S_{Y_t}^2 = \frac{\sum Y_t^2}{T} - \bar{Y}_t^2 = \frac{99{,}1637}{14} - 1{,}6493^2 = 0{,}0643$$

resulta que

$$R^2 = \frac{0{,}9775^2}{16{,}25 \cdot 0{,}0643} = 0{,}92$$

y, al ser **muy próximo a 1,** se concluye que la tendencia lineal estimada se ajusta bastante bien (o representa bastante bien) a la serie original.

d) En el siguiente gráfico se han representado la serie original y la tendencia lineal estimada. En el mismo se observa el grado de ajuste de la recta estimada a la serie original.

Gráfico 5.10. Millones de asalariados en el sector público y tendencia lineal estimada

e) La predicción para el año 2010, basada en la tendencia lineal estimada, se calcula sustituyendo en la expresión de la recta el valor de la variable **t** por 2010. Concretamente

$$\mathbf{Y^*_{2010} = -117{,}5 + 0{,}06 \cdot 2010 = 3{,}1 \text{ millones de asalariados}}$$

¿Es fiable dicha predicción?

La fiabilidad de dicha predicción viene avalada por el elevado coeficiente de determinación y por la cercanía del año para el que se hace la predicción al intervalo de años que hemos utilizado en la estimación de la tendencia.

Como norma a seguir, una predicción hecha con una tendencia lineal diremos que es fiable si el valor del coeficiente $\mathbf{R^2}$ es elevado (Si $R^2 \geq 0{,}75$, el ajuste de la tendencia lineal es bueno y si $R^2 \geq 0{,}95$, el ajuste es muy bueno) y el valor de **t** está entre los utilizados para la estimación o próximo a ellos.

En nuestro caso, el valor de **t** es próximo a los utilizados y el valor del coeficiente $\mathbf{R^2}$ es elevado, luego la predicción es fiable.

Observación. Como paso previo a la estimación de la tendencia, podríamos haber cambiado el origen del tiempo, para facilitar los cálculos. En este sentido, los cambios más usuales son:

- **Situar el origen en el primer año de la serie (t= 0 en 1996)**
- **Situar el origen en el centro de los datos (t= 0 en 2002,5)**

Los cálculos intermedios necesarios para la estimación de la tendencia cuando se lleva a cabo un cambio como los anteriormente descritos, así como los resultados obtenidos para los datos del ejemplo anterior, se presentan a continuación.

t= 0 en 1996

Tabla 5.6. Cálculos intermedios para la estimación de la tendencia

t	y_t	t^2	$t \cdot y_t$	$(y_t)^2$
0	2,32	0	0	5,3824
1	2,36	1	2,36	5,5696
2	2,33	4	4,66	5,4289
3	2,36	9	7,08	5,5696
4	2,44	16	9,76	5,9536
5	2,51	25	12,55	6,3001
6	2,59	36	15,54	6,7081
7	2,71	49	18,97	7,3441
8	2,8	64	22,4	7,8400
9	2,86	81	25,74	8,1796
10	2,88	100	28,8	8,2944
11	2,91	121	32,01	8,4681
12	2,96	144	35,52	8,7616
13	3,06	169	39,78	9,3636
Totales 91	**37,09**	**819**	**255,17**	**99,1637**

Teniendo en cuenta los datos de la tabla anterior, resulta

$$\bar{t} = \frac{\sum t}{T} = \frac{91}{14} = 6{,}5$$

$$\bar{y}_t = \frac{\sum y_t}{T} = \frac{37{,}09}{14} = 2{,}6493$$

$$S_{ty_t} = \frac{\sum t \cdot y_t}{T} - \bar{t} \cdot \bar{y}_t = \frac{255{,}17}{14} - 6{,}5 \cdot 2{,}6493 = 1{,}006$$

$$S_t^2 = \frac{\sum t^2}{T} - \bar{t}^2 = \frac{819}{14} - 6{,}5^2 = 16{,}25$$

Luego los valores estimados de los parámetros de la recta que representa a la tendencia son, en este caso

$$b = \frac{S_{ty_t}}{S_t^2} = \frac{1{,}006}{16{,}25} = 0{,}06$$

y, teniendo en cuenta que $a = \bar{y}_t - b \cdot \bar{t}$, determinamos el valor del parámetro «**a**», que resulta ser

$$a = 2{,}6493 - 0{,}06 \cdot 6{,}5 = 2{,}2593$$

Por lo tanto, la tendencia lineal estimada es

$$y_t^* = 2{,}2593 + 0{,}06 \cdot t$$

o.t. = 1996; u.t.= 1 año

donde, **o.t. = 1996** y **u.t. = 1 año** nos indican que el origen del tiempo se ha especificado en el año 1996 y la unidad temporal con la que se ha trabajado es el año, respectivamente.

En cuanto al significado del parámetro **«b»,** es la pendiente de la recta estimada que representa a la tendencia anual de los asalariados en el Sector Público en España. En el caso que nos ocupa es, como en el estudiado antes, el crecimiento anual que, por término medio, experimentan los asalariados en el Sector Público en España; es decir, los asalariados en el Sector Público se incrementan cada año, por término medio, en 0,06 millones de personas. En cuanto al significado del parámetro **«a»,** ordenada en el origen, para este ejemplo concreto representa el valor estimado para el número de asalariados en el Sector Público en España en el año 1996. En nuestro caso, 2,2593 millones de personas.

t= 0 en 2002,5

En este caso, los datos y cálculos básicos para la estimación de la tendencia lineal son los que se muestran en la Tabla 5.7.

Tabla 5.7. Cálculos intermedios para la estimación de la tendencia

t	y_t	t^2	$t\,y_t$	y^2_t
-6,5	2,32	42,25	-15,08	5,3824
-5,5	2,36	30,25	-12,98	5,5696
-4,5	2,33	20,25	-10,485	5,4289
-3,5	2,36	12,25	-8,26	5,5696
-2,5	2,44	6,25	-6,1	5,9536
-1,5	2,51	2,25	-3,765	6,3001
-0,5	2,59	0,25	-1,295	6,7081
0,5	2,71	0,25	1,355	7,3441
1,5	2,8	2,25	4,2	7,84
2,5	2,86	6,25	7,15	8,1796
3,5	2,88	12,25	10,08	8,2944
4,5	2,91	20,25	13,095	8,4681
5,5	2,96	30,25	16,28	8,7616
6,5	3,06	42,25	19,89	9,3636
Totales **0**	**37,09**	**227,5**	**14,085**	**99,1637**

Teniendo en cuenta los datos de la tabla anterior, resulta

$$\bar{t} = \frac{\sum t}{T} = \frac{0}{14} = 0$$

$$\bar{y}_t = \frac{\sum y_t}{T} = \frac{37{,}09}{14} = 2{,}6493$$

$$S_{ty_t} = \frac{\sum t \cdot y_t}{T} - \bar{t} \cdot \bar{y}_t = \frac{14{,}085}{14} - 0 \cdot 2{,}6493 = 1{,}006$$

$$S_t^2 = \frac{\sum t^2}{T} - \bar{t}^2 = \frac{227{,}5}{14} - 0^2 = 16{,}25$$

Luego, en este caso, los valores estimados de los parámetros de la recta de tendencia son:

$$b = \frac{S_{tY_t}}{S_t^2} = \frac{1{,}006}{16{,}25} = 0{,}06$$

y, teniendo en cuenta que $a = \bar{y}_t - b \cdot \bar{t}$, determinamos el valor del parámetro **«a»**, que resulta ser:

$$a = 2{,}6493 - 0{,}06 \cdot 0 = 2{,}6493$$

Por lo tanto, la tendencia lineal estimada es

$$y_t^* = 2{,}6493 + 0{,}06 \cdot t$$
$$\text{o.t.} = 2002{,}5;\ \text{u.t.} = 1\ \text{año}$$

donde, **o.t. = 2002,5** y **u.t. = 1 año** indican que, en este caso, el origen del tiempo se ha tomado en julio de 2002 y la unidad temporal con la que se ha trabajado ha sido el año, respectivamente.

En cuanto al significado del parámetro **«b»**, pendiente de la recta estimada para representar a la tendencia anual de los asalariados en el Sector Público en España, representa, como en el caso estudiado antes, el crecimiento anual que, por término medio, experimentan los asalariados en dicho sector en el periodo analizado; es decir, en dicho periodo los asalariados en el Sector Público se han incrementado cada año, por término medio, en 0,06 millones de personas. En cuanto al significado del parámetro **«a»**, ordenada en el origen, para este ejemplo concreto representa el valor estimado, en millones de personas, para el número de asalariados en dicho sector en julio de 2002.

5.1.2. Estimación de la tendencia lineal cuando hay estacionalidad

Para estimar la tendencia lineal cuando en la serie se observan variaciones estacionales, es necesario el filtrado previo de la misma para eliminar dicha componente; después de ello, se puede estimar la tendencia de la nueva serie siguiendo el mismo procedimiento que en el caso precedente, ya que la serie filtrada solo contiene tendencia, ciclos y variaciones irregulares.

¿Cómo eliminamos la componente estacional de la serie original?

Ello dependerá del tipo y la cantidad de datos de que dispongamos. Así,

- **Si la variable a la que la serie se refiere es una variable flujo y disponemos de las observaciones relativas a un número suficientemente grande de años** (suponiendo que el periodo de las variaciones estacionales sea el año), la forma más cómoda de eliminar la componente estacional es mediante la agregación de los valores relativos a un mismo año. De este modo, pasamos a una serie de datos anuales que, por lo tanto, no contiene componente estacional.

- **Si la variable a la que la serie se refiere es una variable *stock* y disponemos de las observaciones relativas a un número suficientemente grande de años** (suponiendo que el periodo de las variaciones estacionales sea el año), la forma más cómoda de eliminar la componente estacional es mediante el promedio de los valores relativos a un mismo año, pasando de este modo a la serie de datos anuales en la que no está presente la componente estacional. Basándonos en la serie de datos anuales, estimamos la tendencia lineal como en el caso anterior.
- En otro caso **(pocos datos),** para eliminar la componente estacional con objeto de estimar una recta que represente el comportamiento de la variable a corto o medio plazo, (con pocos datos sería una tendencia a corto plazo, que puede ser útil con fines predictivos), es necesario filtrar la serie haciendo uso de una estimación de la componente estacional.

Ejemplo 5.8. Disponemos de los datos mensuales sobre ingresos por turismo (miles de euros) en España para los años del periodo 1990-2006. La representación gráfica de dicha serie temporal es la que figura a continuación.

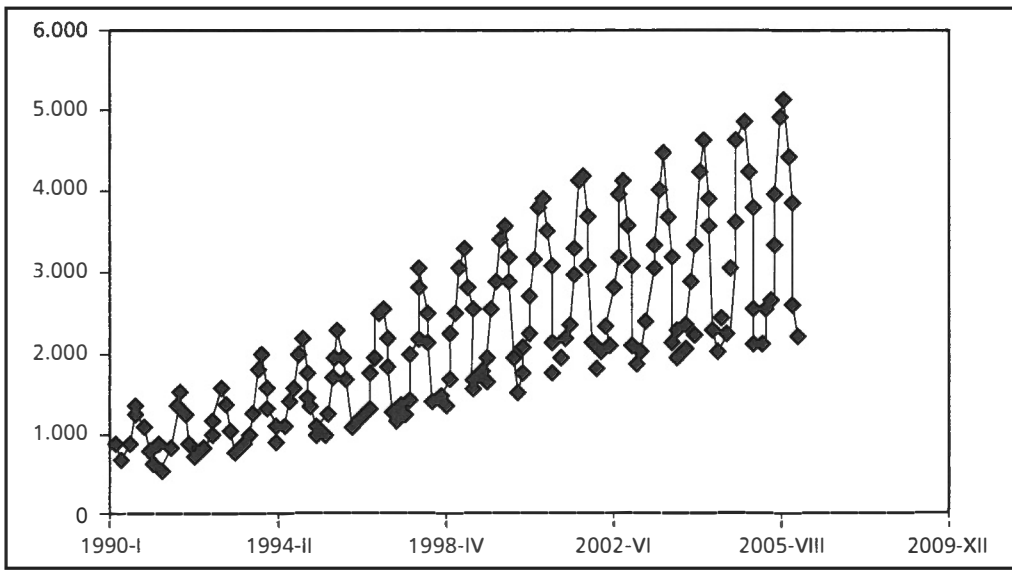


Gráfico 5.11. Evolución temporal de los ingresos por turismo

En el gráfico anterior se aprecia que la variable ingresos por turismo en España tiene un marcado carácter estacional. En este caso, disponemos de las observaciones mensuales relativas a diecisiete años. Número de años suficiente para estimar la tendencia a largo plazo. Para llevar a cabo dicha estimación, agregamos los datos mensuales relativos a un mismo año y transformamos la serie en una de datos anuales (en esta ya no está presente la componente estacional) y estimamos la recta que representa a la tendencia como en el caso ya estudiado.

La representación gráfica de la serie de datos anuales, en miles de millones de euros, junto con la tendencia lineal estimada es la que figura en el siguiente gráfico.

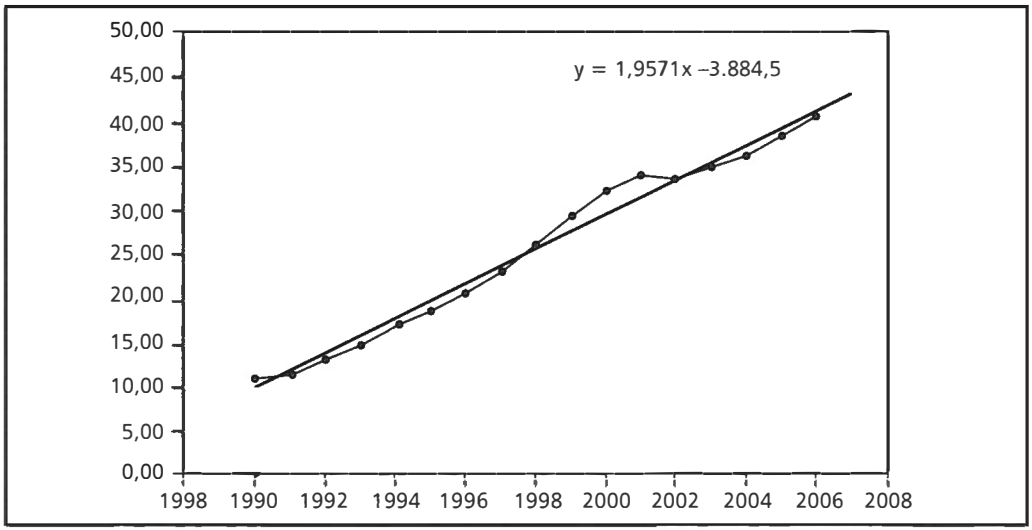


Gráfico 5.12. Evolución temporal de los ingresos por turismo

5.2. Estimación de la componente estacional

En el análisis clásico de series temporales es frecuente el interés por aislar la componente estacional de la serie con el objeto de realizar un análisis específico de dicha componente o de utilizarla con fines predictivos.

El método más simple para aislar la componente estacional es el de la **razón a la media móvil,** que está basado en la hipótesis de integración multiplicativa de las componentes y en la idea de que la serie de medias móviles **(MM_t)** que se obtiene asignando a cada observación el promedio de las que le anteceden y preceden hasta hacer el total de un año (si el periodo de las variaciones estacionales es ese) y de modo que ella sea la central (o esté entre las dos centrales), es una suavización de la serie original que solo contiene a la componente mixta **Tendencia - Ciclo**. Esto es, identificamos $T_t - C_t$ con MM_t.

Bajo la hipótesis multiplicativa, dividiendo la serie original **(y_t)** por la serie MM_t, obtenemos la componente estacional afectada de las variaciones irregulares; es decir,

$$\frac{y_t}{MM_t} = \frac{T_t \cdot C_t \cdot E_t \cdot I_t}{T_t \cdot C_t} = E_t \cdot I_t = IEVE_t$$

A los diferentes valores de $IEVE_t$ se les denomina **índices específicos de variación estacional** porque son proporciones (o porcentajes si se multiplican por cien) y comparan, para cada instante observado, el valor de la serie original con el valor medio que proporciona la tendencia-ciclo.

Por ejemplo, **si uno de estos índices vale 1,18, lo interpretamos como que, en el instante al que dicho índice corresponde, el valor de la serie original es un 18% superior al valor medio que proporciona la tendencia-ciclo.**

Cuando los $IEVE_t$ correspondientes a los distintos años para un mismo mes, trimestre, etc. son similares (lo que podemos visualizar a través de su representación gráfica o a partir de la tabla), se dice que la estacionalidad es estable; en caso contrario, diremos que la misma es evolutiva.

Si la estacionalidad es estable, los $\mathbf{IEVE}_t$ correspondientes a los distintos años para un **mismo** mes, trimestre, etc., se promedian, obteniéndose el índice general de variación estacional del mes, trimestre, etc., ($\mathbf{IGVE}_{mes}$, $\mathbf{IGVE}_{trimestre}$, etc.), que es el mismo para todos los años.

Para una serie de datos mensuales con componente estacional de periodicidad anual obtendremos doce IGVE, uno por cada mes del año, siendo su suma igual a 1.200 (12·100). Del mismo modo, para una serie de datos trimestrales con variaciones estacionales de periodicidad anual obtendremos cuatro IGVE, uno por cada trimestre del año, siendo su suma igual a 400 (4·100).

5.2.1. Utilización de los IGVE para filtrar la serie original

Los IGVE, cuando se conocen, se pueden utilizar para filtrar la serie original y eliminar de ella la componente estacional. La serie resultante, que se denomina **serie desestacionalizada,** permite el **análisis del comportamiento de la variable a corto plazo (análisis coyuntural).**

Desestacionalización

Para filtrar la serie basta dividir cada dato por el **IGVE** correspondiente al instante al que el mismo se refiere. Es decir, **para una serie de datos trimestrales, la serie desestacionalizada es la que resulta al dividir el dato de cada trimestre por el IGVE (expresado en tanto por uno) de dicho trimestre.**

Ejemplo 5.9. En la siguiente tabla figuran los datos mensuales sobre ingresos por turismo en España (en miles de €) para los años 2005 y 2006. En dicha tabla se muestran también los índices generales de variación estacional (IGVE) que describen el comportamiento estacional de los ingresos por turismo en España.

Tabla 5.8. Ingresos por turismo en miles de euros

	Ingresos por turismo (miles de €)	IGVE
2005-01	2.489,114	77,44
2005-02	2.118,604	68,84
2005-03	2.527,875	77,00
2005-04	2.297,491	80,80
2005-05	3.084,082	100,67
2005-06	3.627,712	113,31
2005-07	4.667,190	142,39
2005-08	4.904,089	150,41
2005-09	4.284,779	127,69
2005-10	3.838,741	113,13
2005-11	2.548,539	79,92
2005-12	2.170,062	68,29
2006-01	2.539,410	77,44
2006-02	2.175,142	68,84
2006-03	2.635,120	77,00
2006-04	2.663,491	80,80
2006-05	3.352,378	100,67
2006-06	3.988,373	113,31
2006-07	4.936,390	142,39
2006-08	5.148,982	150,41
2006-09	4.446,979	127,69
2006-10	3.924,371	113,13
2006-11	2.630,415	79,92
2006-12	2.268,864	68,29

Utilice la información proporcionada para resolver las siguientes cuestiones:

a) Compare el comportamiento estacional del ingreso por turismo en los meses de julio y noviembre.

b) Obtenga la serie del ingreso mensual por turismo desestacionalizada, represéntela gráficamente y compárela con la original.

Solución

a) El valor del IGVE correspondiente al mes de julio es 142,39, mientras que el correspondiente al mes de noviembre es 79,92. Lo anterior significa que, mientras que en el mes de julio los ingresos por turismo en España son superiores al ingreso medio mensual por este concepto en un 42,39%, en el mes de noviembre estos están un 20,08% por debajo del valor medio mensual.

b) Para obtener la serie desestacionalizada basta dividir cada dato mensual por el IGVE correspondiente al mes expresado en tanto por uno. Dicha serie se presenta en la siguiente tabla.

Tabla 5.9. Ingresos por turismo e ingresos por turismo desestacionalizados

	Ingresos por turismo (miles de €)	IGVE	Ingresos por turismo desestacionalizados (miles de €)
2005-01	2.489,114	77,44	3.214,248
2005-02	2.118,604	68,84	3.077,577
2005-03	2.527,875	77,00	3.282,955
2005-04	2.297,491	80,8	2.843,429
2005-05	3.084,082	100,67	3.063,556
2005-06	3.627,712	113,31	3.201,582
2005-07	4.667,190	142,39	3.277,751
2005-08	4.904,089	150,41	3.260,481
2005-09	4.284,779	127,69	3.355,610
2005-10	3.838,741	113,13	3.393,212
2005-11	2.548,539	79,92	3.188,863
2005-12	2.170,062	68,29	3.177,716
2006-01	2.539,410	77,44	3.279,197
2006-02	2.175,142	68,84	3.159,707
2006-03	2.635,120	77,00	3.422,234
2006-04	2.663,491	80,80	3.296,400
2006-05	3.352,378	100,67	3.330,067
2006-06	3.988,373	113,31	3.519,878
2006-07	4.936,390	142,39	3.466,809
2006-08	5.148,982	150,41	3.423,298
2006-09	4.446,979	127,69	3.482,637
2006-10	3.924,371	113,13	3.468,904
2006-11	2.630,415	79,92	3.291,310
2006-12	2.268,860	68,29	3.322,390

La representación gráfica de la serie original y de la serie desestacionalizada se muestra en el siguiente gráfico.

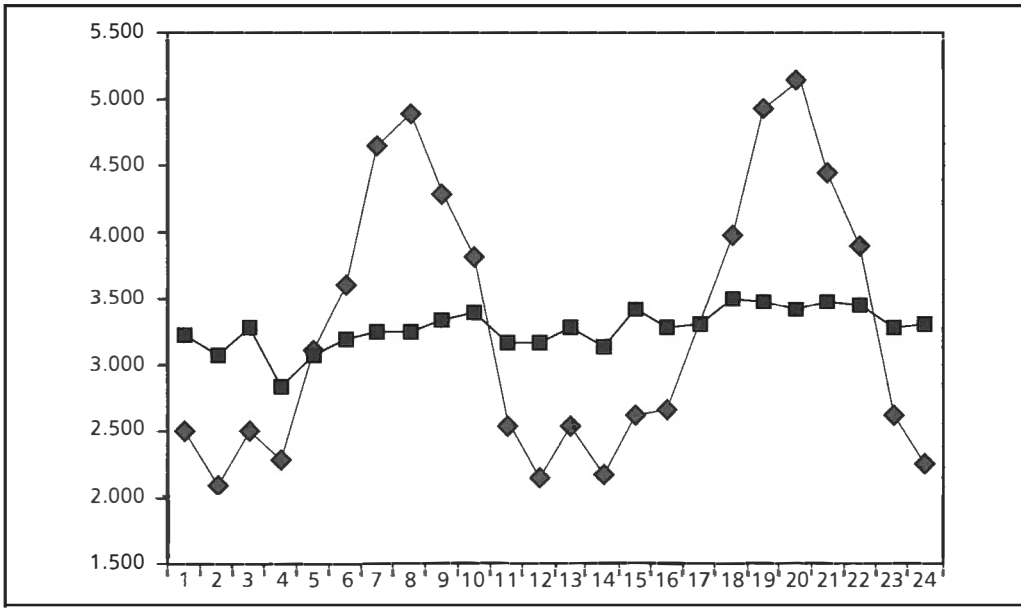


Gráfico 5.13. Ingresos por turismo e ingresos por turismo desestacionalizados

5.2.2. Predicciones a corto plazo

La serie desestacionalizada se puede utilizar para hacer predicciones a corto plazo. El procedimiento a seguir es el que se indica a continuación:

- Estimamos la línea de tendencia a corto plazo que, para nosotros será una recta.
- Utilizamos la recta estimada para hacer la predicción del valor medio correspondiente.
- Multiplicamos el valor medio por el IGVE correspondiente (expresado en tanto por uno) para afectarlo de la estacionalidad propia del periodo al que la predicción se refiere.

Ejemplo 5.10. Estime la tendencia a corto plazo del ingreso mensual por turismo en España. Dé una medida de su representatividad y haga, basándose en ella y en la información disponible, una predicción para el ingreso por turismo en España en el mes de enero de 2007.

Solución. Cuando estamos interesados en desestacionalizar la serie con objeto de obtener la tendencia a corto plazo con fines predictivos, trabajamos bajo la hipótesis de integración mixta de las componentes. Esto es

$$Y_t = (T_t + C_t + I_t) \cdot E_t \xrightarrow{\frac{IGVE_t}{100} = E_t} \frac{Y_t}{\frac{IGVE_t}{100}} = T_t + C_t + I_t$$

Estamos ahora en una situación análoga a la primera estudiada (estimación de la tendencia en ausencia de componente estacional)

$$\frac{Y_t}{E_t} = a + bt + e_t$$

Para estimar los parámetros **a** y **b** del modelo lineal, teniendo en cuenta que las observaciones son mensuales, cambiaremos primero el origen del tiempo. Por comodidad, ponemos el origen en el primer dato de la tabla (enero de 2005). Los cálculos necesarios para las estimaciones, que están basados en los presentados en la Tabla 5.9, son los que se muestran en la Tabla 5.10. Basándonos en la información contenida en dicha tabla, obtenemos que

$$\bar{t} = \frac{\sum t}{T} = \frac{276}{24} = 11{,}5$$

$$\bar{y}_{t;D} = \frac{\sum y_t}{T} = \frac{78.799{,}8085}{24} = 3.283{,}3253$$

$$S_{ty_{t;D}} = \frac{\sum t \cdot y_t}{T} - \bar{t} \cdot \bar{y}_t = \frac{922.824{,}5982}{24} - 11{,}5 \cdot 3.283{,}325354 = 692{,}7833476$$

$$S_t^2 = \frac{\sum t^2}{T} - \bar{t}^2 = \frac{4324}{24} - 11{,}5^2 = 47{,}91666667$$

Tabla 5.10. Tendencia lineal de los ingresos por turismo desestacionalizados. Cálculos auxiliares

	t	$y_{t;D}$	t^2	$t\ y_{t;D}$	$y^2_{t;D}$
2005-01	0	3.214,24845	0	0	10.331.393,1
2005-02	1	3.077,57699	1	3.077,57699	9.471.480,13
2005-03	2	3.282,954545	4	6.565,909091	10.777.790,55
2005-04	3	2.843,429455	9	8.530,288366	8.085.091,068
2005-05	4	3.063,556174	16	12.254,22469	9.385.376,429
2005-06	5	3.201,581502	25	16.007,90751	10.250.124,11
2005-07	6	3.277,751247	36	19.666,50748	10.743.653,23
2005-08	7	3.260,480686	49	22.823,3648	10.630.734,3
2005-09	8	3.355,610463	64	26.844,8837	11.260.121,58
2005-10	9	3.393,212234	81	30.538,9101	11.513.889,26
2005-11	10	3.188,862613	100	31.888,62613	10.168.844,76
2005-12	11	3.177,715625	121	34.954,87187	10.097.876,59
2006-01	12	3.279,196798	144	39.350,36157	10.753.131,64
2006-02	13	3.159,706566	169	41.076,18536	9.983.745,583
2006-03	14	3.422,233766	196	47.911,27273	11.711.683,95
2006-04	15	3.296,399752	225	49.445,99629	10.866.251,33
2006-05	16	3.330,066554	256	53.281,06487	11.089.343,25
2006-06	17	3.519,877328	289	59.837,91457	12.389.536,4
2006-07	18	3.466,809467	324	62.402,57041	12.018.767,88
2006-08	19	3.423,297653	361	65.042,65541	11.718.966,82
2006-09	20	3.482,636855	400	69.652,7371	12.128.759,46
2006-10	21	3.468,903916	441	72.846,98223	12.033.294,38
2006-11	22	3.291,31006	484	72.408,82132	10.832.721,91
2006-12	23	3.322,389808	529	76.414,96559	11.038.274,04
Total	**276**	**78.799,80851**	**4324**	**922.824,5982**	**259.280.851,8**

Teniendo en cuenta los cálculos previos, los valores estimados para los parámetros de la recta que representa la tendencia son

$$b=\frac{S_{ty_t}}{S_t^2}=\frac{692,7833476}{47,91666667}=14,458$$

y, teniendo en cuenta que $a=\bar{y}_t-b\cdot\bar{t}$, determinamos el valor del parámetro **«a»**, que resulta ser:

$$a = 3.283,325354 - 14,458 \cdot 11,5 = 3.117,058$$

Por lo tanto, la tendencia lineal estimada para la serie desestacionalizada es

$$y_t^* = 3.117,058 + 14,458 \cdot t$$

o.t.= enero de 2005; u.t. = 1 mes

En el siguiente gráfico se muestran tanto la serie desestacionalizada como la tendencia estimada para la misma.

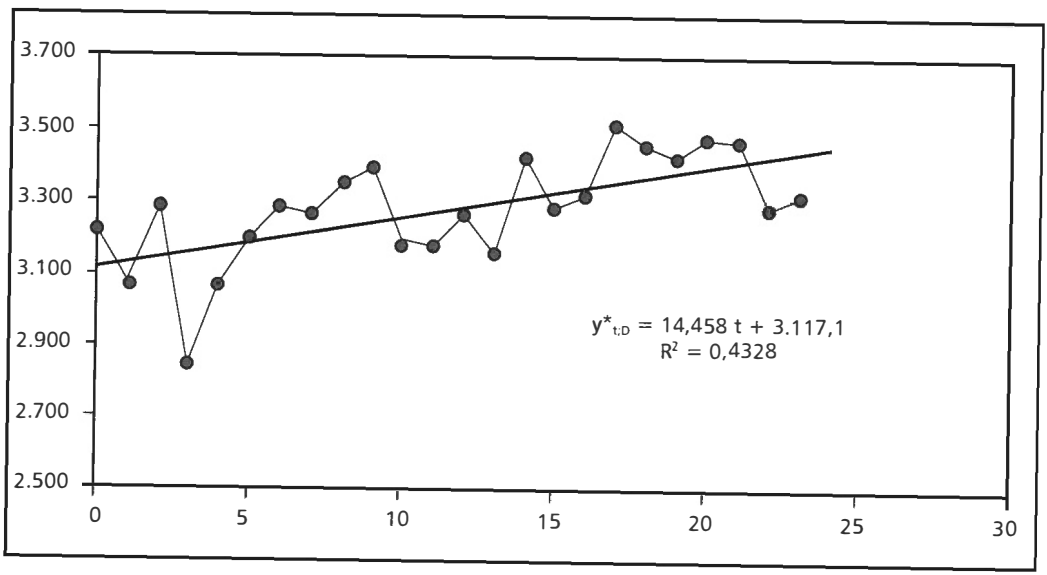


Gráfico 5.14. Ingresos por turismo: Serie desestacionalizada y tendencia lineal

Para obtener la predicción correspondiente a enero de 2007, hacemos primero la predicción del valor medio correspondiente a dicho mes, utilizando la tendencia estimada, después afectaremos a dicho valor de la estacionalidad propia del mes de enero, multiplicando el mismo por el IGVE correspondiente a enero, expresado en tanto por uno.

Paso 1. Hay que sustituir el valor de t que corresponde a enero de 2007 en la expresión de la tendencia lineal obtenida, sabiendo que t=0 en enero de 2005. El valor de t que corresponde es t=24, luego

$$y^*_{\text{enero 07;D}} = 3.117,058 + 14,458 \cdot 24 = 3.464,05 \text{ (en miles de euros)}$$

Paso 2. Teniendo en cuenta que IGVE (enero)=77,44, estimamos el valor del ingreso correspondiente a enero de 2007.

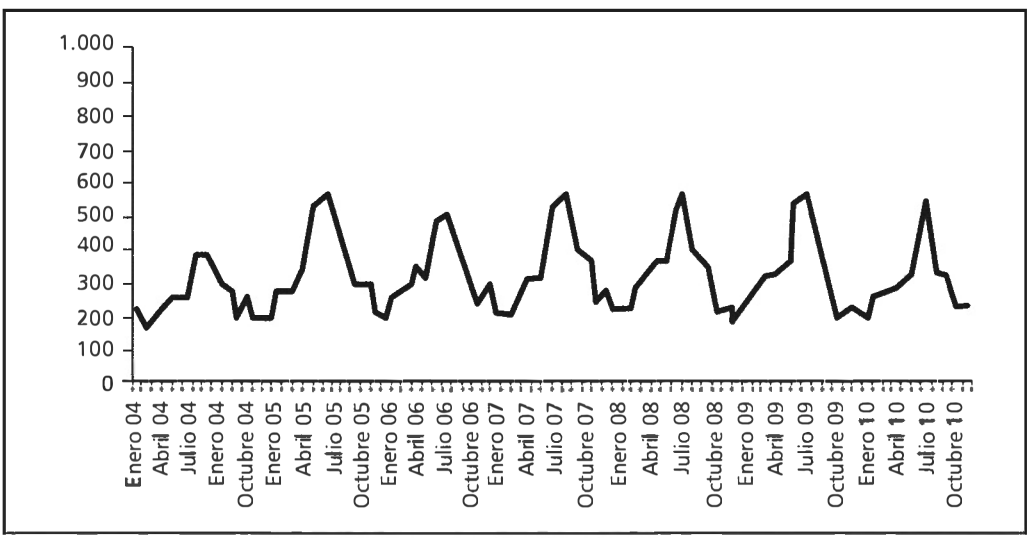
$$y^*_{enero07} = y^*_{enero07;D} \cdot \frac{IGVE(enero)}{100} = 3.464{,}05 \cdot 0{,}7744 = 2.682{,}56 \text{ miles de } €$$

Observación. Si, para una variable afectada de estacionalidad mensual, disponemos de la tendencia estimada con datos anuales y queremos hacer una predicción mensual basándonos en ella y en los IGVE relativos a la misma, se procede del siguiente modo:

- Estimamos con la tendencia el valor medio anual.
- Si la variable es flujo, el valor medio mensual será el resultado de dividir el anual por 12.
- Si la variable es *stock*, el valor medio mensual coincide con el valor medio anual.
- Por último, afectamos al valor medio mensual así obtenido de la estacionalidad propia del mes al que se refiere la predicción, multiplicándolo por el correspondiente IGVE.

5.3. Ejercicios

Ejercicio 5.1. En el Gráfico 5.15 se muestra la evolución del gasto medio de los turistas españoles que visitan Canarias y proceden de otras regiones españolas (a excepción de los turistas provenientes de Andalucía, Baleares, Cataluña, Valencia, Madrid). Dicha serie, expresada en euros, va desde enero de 2004 hasta diciembre 2010.

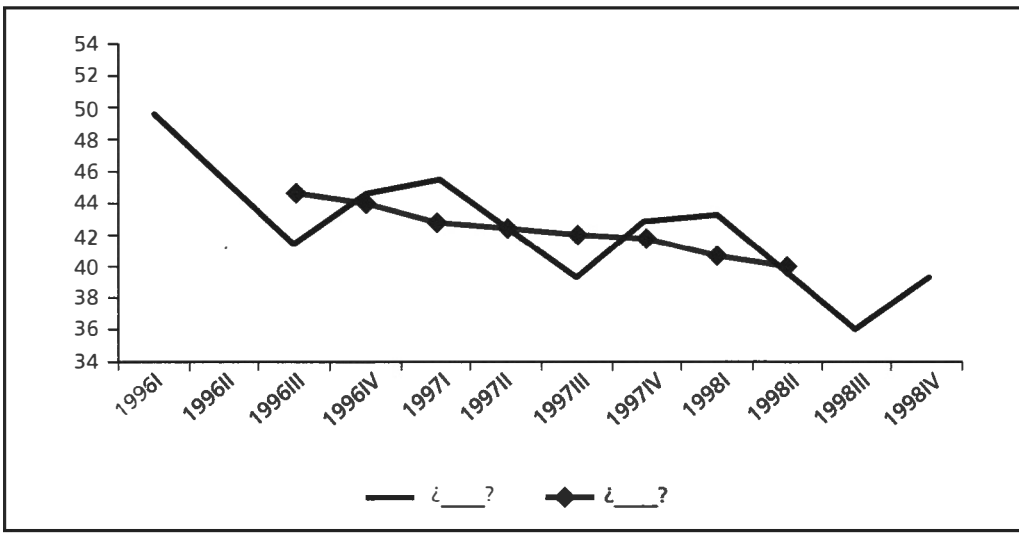


Fuente: *Encuesta de Gasto Turístico*, IET (2011)

Gráfico 5.15. Evolución del gasto medio de los turistas españoles que visitan Canarias

Atendiendo a su comportamiento a largo plazo, ¿es una serie estacionaria o evolutiva? Justifique su respuesta. ¿Qué le ocurre a la serie dentro de cada año?

Ejercicio 5.2. La siguiente representación (Gráfico 5.16) se ha realizado con los datos trimestrales de las personas paradas, expresados en miles, en la provincia de Málaga. Se recogen dos series temporales: la original y la que muestra la tendencia-ciclo de la original mediante la realización de una media móvil de amplitud cuatro observaciones.



Fuente: Elaboración propia a partir de la información del Servicio Público de Empleo (2010)

Gráfico 5.16. Datos trimestrales de las personas paradas en la provincia de Málaga (en miles)

Identifique la serie original y la transformada mediante la media móvil. ¿Qué evolución a medio plazo parece presentar la serie de personas paradas en la provincia de Málaga?

Ejercicio 5.3. En las dos tablas que siguen se presentan observaciones temporales relativas a España para un conjunto de variables.

Total expedientes de órdenes de pago tramitados por el FOGASA (2005)		Afiliados al Régimen Especial del Mar por cuenta propia (2005)		Demandas de primer empleo (2004)	
Enero	83	Enero	16.401	Enero	275.847
Febrero	1.569	Febrero	16.321	Febrero	269.450
Marzo	4.832	Marzo	16.126	Marzo	290.007
Abril	3.802	Abril	15.936	Abril	249.503
Mayo	4.485	Mayo	15.847	Mayo	258.162
Junio	3.678	Junio	15.779	Junio	320.279
Julio	3.504	Julio	15.756	Julio	383.022
Agosto	1.674	Agosto	15.748	Agosto	264.918
Septiembre	4.601	Septiembre	15.722	Septiembre	336.782
Octubre	2.390	Octubre	15.819	Octubre	318.611
Noviembre	3.209	Noviembre	15.976	Noviembre	301.598
Diciembre	2.952	Diciembre	15.966	Diciembre	251.341

Fuente: Fondo de Garantía Salarial del M.T.A.S.; Seguridad Social; y *Boletín Mensual de Estadística* del INE (2007)

Cooperativas constituidas (2005)		Índice General de Precios al Consumo (base 2001=100) (2005)		Permisos de trabajo concedidos a extranjeros (2005)	
Enero	164	Enero	110,752	Enero	30.926
Febrero	163	Febrero	111,0,39	Febrero	33.647
Marzo	191	Marzo	111,917	Marzo	39.211
Abril	157	Abril	113,529	Abril	73.180
Mayo	155	Mayo	113,746	Mayo	124.694
Junio	133	Junio	114,012	Junio	147.511
Julio	141	Julio	113,315	Julio	132.612
Agosto	67	Agosto	113,812	Agosto	107.172
Septiembre	124	Septiembre	114,51	Septiembre	105.748
Octubre	105	Octubre	115,442	Octubre	55.957
Noviembre	105	Noviembre	115,617	Noviembre	43.279
Diciembre	104	Diciembre	115,865	Diciembre	29.042

Fuente: Seguridad Social y *Boletín Anual de Estadísticas Laborales y Asuntos Sociales* del M.T.A.S (2006)

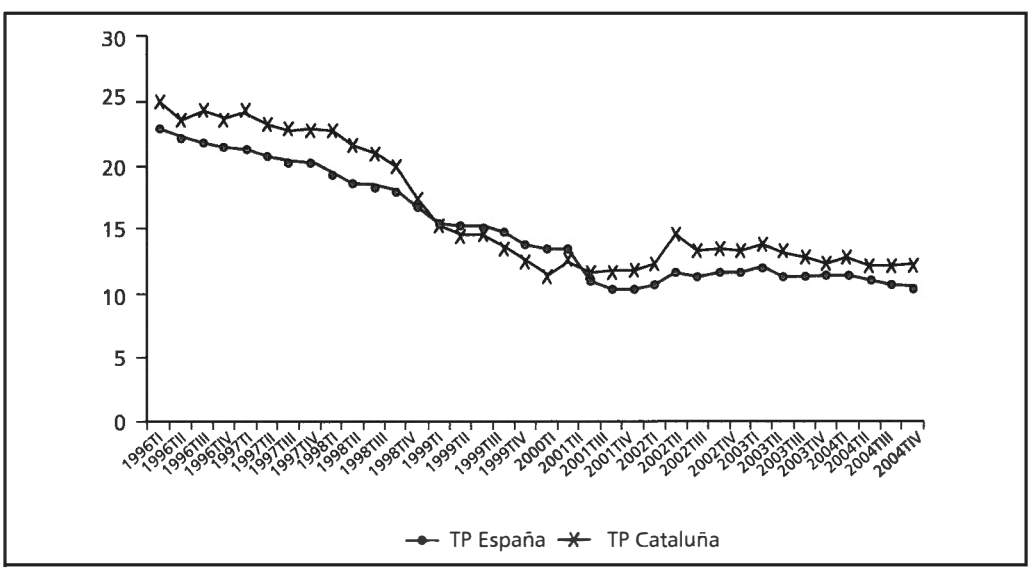
Se pide:

a) Indique, para cada una de las variables anteriores, si es flujo o *stock*.

b) Estime un valor para el IPC en el año 2005 y otro para el número de demandas de empleo en el año 2004.

c) ¿A cuánto ascendió el número de cooperativas constituidas en el tercer trimestre del año 2005? ¿Cuál fue el número de afiliados al Régimen Especial del Mar por cuenta ajena para dicho trimestre?

Ejercicio 5.4. En el Gráfico 5.17 se han representado las tasas de paro (TP) trimestrales en España y Cataluña durante el periodo 1996-I - 2004-IV.



Fuente: Elaboración propia a partir de la Encuesta de Población Activa (EPA), INE 2012

Gráfico 5.17. Tasas de paro trimestrales en España y Cataluña (1996-2004)

Se pide:

a) ¿De qué tipo son las observaciones que se han representado en el gráfico anterior: transversales, temporales o mixtas?

b) ¿Qué tipo de variable es la tasa de paro: flujo o *stock*? ¿Cómo obtendría la tasa de paro anual a partir de las tasas de paro trimestrales?

c) ¿Cuáles de las componentes teóricas de una serie temporal se aprecian en las series referidas a las tasas de paro de España y Cataluña que se han representado en el gráfico anterior?

d) A partir de la información proporcionada en dicho gráfico, describa la evolución a largo plazo de las tasas de paro en España y Cataluña.

e) En la Tabla 5.11 se muestran las tasas de paro trimestrales en Cataluña y las medias móviles de amplitud cuatro (MM) para dicha variable en el periodo 1996-I – 2001-IV. Complete la columna de medias móviles y represente la serie de medias móviles junto a la serie original. Comente el gráfico obtenido.

Tabla 5.11. Tasas de paro trimestrales en Cataluña y medias móviles (MM) para dicha variable (1996-2001)

Trimestres	TP Cataluña	MM (a = 4)
1996-I	24,95	
1996-II	23,63	
1996-III	24,32	24,08
1996-IV	23,7	23,97
1997-I	24,37	23,74
1997-II	23,32	23,42
1997-III	22,8	23,08
1997-IV	22,68	22,65
1998-I	22,69	22,21
1998-II	21,55	21,63
1998-III	21,01	20,61
1998-IV	19,87	
1999-I	17,29	
1999-II	15,33	
1999-III	14,45	
1999-IV	14,63	
2000-I	13,63	
2000-II	12,36	
2000-III	11,36	
2000-IV	12,7	
2001-I	11,42	
2001-II	11,36	
2001-III	11,73	
2001-IV	12,21	

Fuente: Elaboración propia a partir de la Encuesta de Población Activa del INE, 2012

f) Se ha procedido a estimar por mínimos cuadrados ordinarios la tendencia de la tasa de paro trimestral en Cataluña, haciendo uso de una tendencia lineal cuyo origen se sitúa en el primer trimestre del año 1996 y con unidad temporal el trimestre. Los resultados de la estimación de dicha tendencia son los siguientes:

$$\mathbf{Y_t = 23{,}4655 - 0{,}4062\ t + e_t} \qquad \mathbf{R^2 = 0{,}7611}$$

(o.t. = 1996-I = 0; u.t = 1 trimestre)

Comente el significado de la pendiente de la recta estimada y el valor del coeficiente de determinación del ajuste.

g) Sabiendo que se ha producido un empeoramiento económico desde el año 2007 hasta la actualidad y que ello se ha traducido en un aumento paulatino de las tasas de paro y suponiendo que esta situación ha tenido su reflejo en el contexto de la economía catalana, ¿qué ocurriría con la tendencia de la tasa trimestral de paro en Cataluña si considerásemos solamente los datos del periodo 2007-I – 2012-I? ¿Qué signo tendría la tendencia lineal a medio plazo estimada únicamente a partir de dichos datos?

Ejercicio 5.5. En la Tabla 5.12 se presentan los datos sobre las ventas en España de coches nuevos y las ventas de coches usados **(y_t),** expresadas en millones de euros, y algunos cálculos auxiliares.

Tabla 5.12. Datos sobre las ventas es España de coches nuevos y de coches usados (y_t) (millones de euros)

Años	Ventas coches nuevos (10^6)	t	t^2	y_t^2	$t\,y_t$
2001	1,51	0	0	1,17	0,00
2002	1,41	1	1	1,51	1,23
2003	1,46	2	4	1,85	2,72
2004	1,61	3	9	2,40	4,65
2005	1,65	4	16	2,59	6,44
2006	1,75	5	25	2,69	8,20
2007	1,61	6	36	3,31	10,92
2008	1,16	7	49	2,56	11,2
2009	0,95	8	64	2,62	12,96
2010	0,98	9	81	2,72	14,85
2011	0,82	10	100	2,89	17,00

Fuente: Elaboración propia a partir de GAUVAM y ANFAC (2011)

a) Estime la tendencia lineal de las ventas de coches nuevos en el periodo considerado y haga lo mismo para las ventas de coches usados.

b) Interprete las pendientes de las tendencias lineales estimadas y analice la bondad de ajuste de las mismas.

c) ¿Para qué año las ventas estimadas de coches nuevos y usados son iguales?

d) Realice las predicciones de las ventas de coches usados correspondientes a los años 2010, 2012 y 2020.

e) Indique cómo se denomina a cada una de las predicciones que ha realizado en el apartado anterior. ¿Son fiables? Justifique sus respuestas.

f) Obtenga el error en la predicción de las ventas de vehículos usados para el año 2010.

Ejercicio 5.6. Con los datos referidos a la estancia media por Comunidades Autónomas en campings durante el periodo correspondiente desde enero de 1999 hasta abril de 2012 se han calculado los Índices Generales de Variación Estacional (IGVE). Dichos índices junto con los datos mensuales observados durante el año 2011 se muestran en la Tabla 5.13.

Tabla 5.13. Índice Generales de Variación estacional (IGUE) (2011)

Meses	Índices Generales de Variación Estacional (IGVE)	Estancia media mensual (días) 2011
Enero	175,64	8,94
Febrero	148,85	6,99
Marzo	106,46	6,47
Abril	66,87	3,97
Mayo	63,60	3,97
Junio	57,61	3,46
Julio	65,79	3,32
Agosto	74,92	4,06
Septiembre	69,92	4,05
Octubre	80,70	4,41
Noviembre	137,23	7,91
Diciembre	152,42	9,23

Fuente: Elaboración propia y Encuesta de Ocupación en Alojamientos Turísticos, INE (2012)

Se pide:

a) Interprete el Índice General de Variación Estacional del mes de septiembre, bajo el supuesto de estacionalidad estable.

b) Desestacionalice las estancias medias en campings correspondientes al tercer trimestre de 2011.

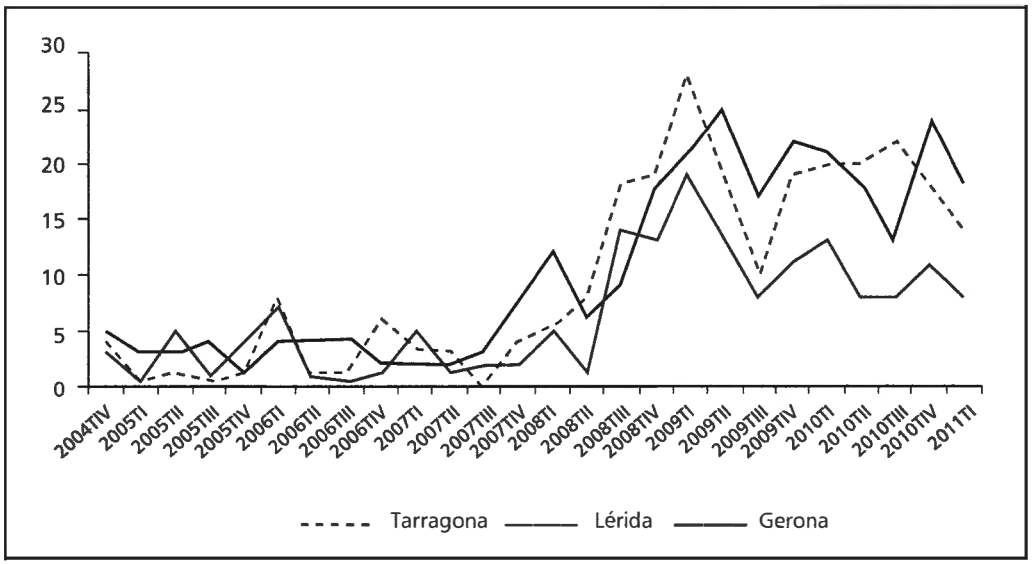
c) Sabiendo que la estancia media en campings para el mes de enero de 2012 fue de 9,3 días, estime las estancias medias previstas para los meses de febrero y marzo correspondientes a dicho año.

d) Con los datos anuales de las estancias medias en campings para los años 1999 a 2011, se ha estimado la tendencia lineal que se muestra a continuación

$$\mathbf{Y_t = 5{,}0031 + 0{,}0014\ t + e_t}$$
$$\mathbf{(o.t. = 1999;\ u.t. = año)}$$

Realice una predicción para el año 2012 y, a partir de ella, estime la estancia media correspondiente al mes de febrero de dicho año y compare el valor obtenido con el que ha calculado en el apartado c). Comente los resultados obtenidos.

Ejercicio 5.7. En el Gráfico 5.18 se presenta la evolución del número trimestral de empresas que entraron en concurso en el periodo 2004-IV – 2011-I. Los datos representados corresponden solo a tres de las cuatro provincias de Cataluña (Gerona, Lérida y Tarragona).



Fuente: Elaboración propia a partir de la Estadística de Procedimiento Concursal, INE (2012)

Gráfico 5.18. Evolución del número trimestral de empresas que entran en proceso concursal (2004-2011)

Las estimaciones de las tendencias lineales para cada una de las provincias que forman la comunidad son las que aparecen en el siguiente cuadro

Provincia	Tendencia lineal	Bondad del ajuste
Barcelona	$Y_t = -17{,}5299 + 12{,}35623\ t + e_t$ (o.t. = 2004-IV; ut. 1 trimestre)	$R^2 = 0{,}7645$
Gerona	$Y_t = 0{,}91429 + 0{,}89777\ t + e_t$ (o.t. = 2004-IV; ut. 1 trimestre)	$R^2 = 0{,}7044$
Lérida	$Y_t = 0{,}726495 + 0{,}044957\ t + e_t$ (o.t. = 2004-IV; ut. 1 trimestre)	$R^2 = 0{,}4199$
Tarragona	$Y_t = -1{,}829 + 0{,}9186\ t + e_t$ (o.t. = 2004-IV; ut. 1 trimestre)	$R^2 = 0{,}6572$

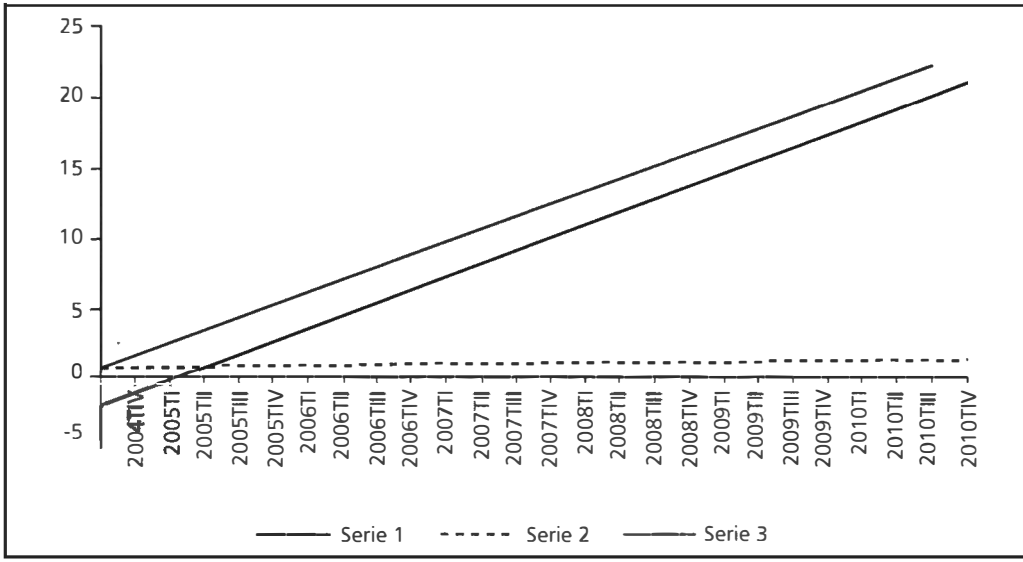
Fuente: Elaboración propia a partir del INE

Se pide:

a) Analizando el gráfico de las series temporales referido a las provincias de Gerona, Lérida y Tarragona, ¿cómo ha evolucionado en dichas provincias el número de empresas que entran en concurso de acreedores durante el periodo considerado?

b) Indique en qué provincia catalana se ha producido un mayor incremento medio trimestral de empresas concursadas. Justifique su respuesta numéricamente.

c) En el Gráfico 5.19 se han representado las tendencias lineales, correspondientes a las provincias de Gerona, Lérida y Tarragona, que reflejan la evolución del número de empresas que entraron en concurso de acreedores en el periodo analizado. Indique cuál de las rectas corresponde a cada una de las tres provincias. Justifique su respuesta.



Fuente: Elaboración propia a partir de la Estadística de Procedimiento Concursal, INE (2012)

Gráfico 5.19. Evolución del número trimestral de empresas que entran en concurso de aceedores. Tendencia lineal (2004-2010)

d) Obtenga el número previsto de empresas concursadas en Cataluña para el primer trimestre del año 2012.

Ejercicio 5.8. Con los datos trimestrales de la Encuesta de Población Activa para el periodo 1996-I - 2004-III se ha procedido a la estimación de las tendencias lineales que representan la evolución de las tasas de actividad masculina y femenina, respectivamente, para la Comunidad Autónoma de Andalucía. A continuación, se muestran algunos resultados de la estimación

$$\textbf{Actividad masculina } Y_t = 65{,}34 + 0{,}07t + e_t \qquad R^2 = 0{,}6609$$

$$\textbf{(o.t.= 2000-II, u.t = 1 trimestre)}$$

$$\textbf{Actividad femenina } Y_t = 37{,}66 + 0{,}13t + e_t \qquad S_y^2 = 2{,}6544 \qquad S_e^2 = 0{,}8323$$

$$\textbf{(o.t.= 2000-II, u.t = 1 trimestre)}$$

Se pide:

a) Interprete los parámetros de la tendencia estimada para la tasa de actividad masculina.

b) ¿Cabe la posibilidad de que la tasa de actividad femenina se equipare a la masculina? ¿Qué condición debería darse para que esto ocurriese? Determine el instante de tiempo y el nivel de tasa de actividad al que se producirá la igualación.

Estadística Descriptiva

Coordinadora:
M.ª Dolores Sarrión Gavilán

Autores:
M.ª Dolores Benítez Márquez
José Luis Iranzo Acosta
Fernando Isla Castillo
M.ª Dolores Sarrión Gavilán

El libro Estadística Descriptiva está pensado para ser utilizado como texto de referencia en cursos introductorios de Estadística en titulaciones de grado del área de Ciencias Sociales, estando especialmente orientado a los ámbitos económico-financiero, empresarial, turístico, de las relaciones laborales y de la administración pública. Teniendo en cuenta el tipo de alumno al que el texto va dirigido, la exposición de los contenidos teóricos se acompaña de numerosos ejemplos resueltos y ejercicios propuestos, que ayudan al alumno a asimilar los diferentes conceptos y sus aplicaciones, facilitándole el estudio de la materia. Los enunciados de la mayor parte de los ejercicios y los ejemplos resueltos están basados en datos reales relativos a algún aspecto de los ámbitos anteriormente mencionados.

Los autores son profesores del Departamento de Estadística y Econometría (Economía Aplicada) de la Universidad de Málaga.

ISBN: 978-84-481-8331-8

www.mcgraw-hill.es

The McGraw Hill Companies