Introducción a la Estadística Descriptiva

2ª Edición

Carla Rey Graña
María Ramil Díaz

netbiblo
www.netbiblo.com

INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA. 2ª Edición

netbiblo
www.netbiblo.com

NETBIBLO, S. L.
C/. Rafael Alberti, 6 bajo izq.
Sta. Cristina 15172 Oleiros (La Coruña) – Spain
Tlf: +34 981 91 55 00 • Fax: +34 981 91 55 11
editorial@netbiblo.com

ISBN: 978-84-9745-167-3
Depósito Legal: C-907-2007

Directora Editorial: Cristina Seco
Editora: Lorena Bello
Producción Editorial: Gesbiblo, S. L.

Impreso en España – Printed in Spain

PRESENTACIÓN

Este manual pretende ser sólo una primera aproximación al análisis cuantitativo de las series de datos de las variables.

Este tipo de análisis ha sido muy criticado, porque la información que contienen los números está sujeta a errores de medición que condicionan los resultados obtenidos. Además, la interpretación que se hace de ellos es subjetiva, de manera que los mismos números pueden ser leídos de distintas formas por personas distintas. Pero, aunque es cierto que cuando los resultados que proporciona se interpretan de forma sesgada o poco objetiva pueden obtenerse conclusiones absurdas, si somos cuidadosos al valorarlos, todas las técnicas cuantitativas, incluso las más sencillas, pueden proporcionar información muy interesante.

El análisis estadístico de las variables puede abordarse, básicamente, desde la perspectiva de una sola o de varias variables.

Con el enfoque univariante, al que se refiere la primera parte de este manual, no se pretende explicar el comportamiento de la variable ni establecer ninguna relación de causalidad. Por tanto, no es necesario disponer de ninguna información relativa al comportamiento de otras variables. Para efectuarlo, basta conocer una serie de datos de la variable en estudio.

Por supuesto, en primer lugar, es necesario tener correctamente especificado el contenido de la serie. Por ejemplo, por lo que se refiere a las utilizadas para el análisis del turismo, debe conocerse el distinto significado que tienen las series de visitantes y turistas, o las diferencias entre turismo emisor, receptor e interno. Y debemos saber también si la serie proporciona información global o parcial, como sucede con la que se refiere sólo al turismo hotelero, si los datos están en cantidades físicas o en valores y, si se trata de valores, si éstos son nominales o reales; si la serie se ha visto afectada por algún cambio metodológico, si la información ha sido obtenida de forma exhaustiva o procede de una encuesta sujeta a errores de muestreo, etcétera. La variable que queremos estudiar debe estar perfectamente definida desde el punto de vista estadístico; debe ser conocida en todos sus aspectos.

Una vez cumplido este requisito, pueden aplicarse algunas de las técnicas univariantes más simples, tales como el análisis gráfico, el cálculo de las principales medidas de posición, dispersión, concentración y forma, el análisis de series temporales, etcétera.

Respecto al análisis gráfico, la elección de la escala que se toma en los ejes puede hacer que la impresión visual que recibimos de la serie sea muy distinta. Según que escalas se tomen puede parecer que la serie varía rápidamente y es muy irregular o que varía lentamente y es muy regular. Así que es necesario ser razonable con el tipo de gráfico que se elige y muy cuidadosos con su interpretación.

La media y las demás medidas de posición o medidas centrales, como la mediana y la moda, tratan de representar o de condensar la información contenida en la serie de datos. Pero la media no es más que una medida resumen de la serie y, en ocasiones, proporciona sólo una información parcial, que puede ser muy poco representativa de ella. Hay algunos ejemplos muy populares al respecto. Según la información que proporciona la media, si un individuo come dos pollos y otro ninguno, ninguno de los dos tiene hambre, porque por término medio, se han comido un pollo cada uno. Si un individuo mete la cabeza en el horno y los pies en el congelador, tiene la temperatura media ideal.

Cualquiera de estos ejemplos, pone de manifiesto la necesidad de complementar la información que proporciona la media con algún indicador de la dispersión de la variable, como la varianza o la desviación típica. Cuando estas medidas toman valores elevados, como sucede en los casos señalados, la media es un resumen de la serie muy pobre, muy poco fiable. Es conveniente valorar conjuntamente la medida de posición central y de dispersión.

El análisis de series temporales permite alisarlas para destacar la tendencia que siguen. Pero, a veces, precisamente las irregularidades eliminadas son importantes para detectar los puntos extraños, que pueden contener mucha información. Por ejemplo, el año 1992 es especial para las series de turismo en España, porque concurren varios fenómenos importantes: se celebró la Olimpiada en Barcelona, Madrid fue Capital Cultural de Europa y en Sevilla se celebró la Exposición Universal. Si suavizamos la serie y eliminamos el pico correspondiente a ese año, esta información se pierde. Por tanto, el alisado también es una técnica que debe ser utilizada con precauciones.

En el enfoque multivariante, que se aborda en la segunda parte del manual, se procede al estudio conjunto de las variables para analizar la relación entre ellas, lo cual permite reducir la incertidumbre con la que se toman las decisiones de planificación y gestión.

Las técnicas estadísticas más sencillas para cuantificar el grado de relación que existe entre distintas variables son el análisis de la covarianza, la correlación y la regresión lineal, en el caso más simple, que es el caso bivariante. Respecto a este tipo

de análisis, también es necesario establecer algunas precisiones, porque la relación entre dos variables, x e y, puede adoptar formas muy diversas.

Se dice que una relación es de causalidad; es decir, que la variable x es la causa de la variable y, si y sólo si las variaciones de x provocan variaciones en y. Una relación es de interdependencia si la relación se da en los dos sentidos; es decir, los valores de x influyen en los valores de y y simultáneamente los valores de y influyen en los de x. En una relación indirecta, la asociación que se observa entre x e y se debe a que ambas dependen de una tercera variable z. Y finalmente, también puede suceder que la asociación entre los valores de x e y sea fruto de la casualidad, siendo ésta la relación que conocemos como espuria.

Las técnicas que se utilizan para analizar la existencia de relación entre las variables permiten cuantificar hasta que punto están relacionadas, pero no permiten, sin embargo, deducir cuál es tipo de relación que existe entre ellas. Son los conocimientos a priori sobre el comportamiento del fenómeno estudiado los que sugieren la existencia ciertas relaciones entre las variables, que los datos y las técnicas estadísticas, únicamente, confirman o refutan. En consecuencia, su utilización requiere un conocimiento previo del fenómeno de covariación que se está estudiando para aplicarlos correctamente.

En definitiva, estos que hemos comentado son sólo algunos ejemplos que ponen de manifiesto el hecho de que al abordar cualquier análisis estadístico, siempre debe tenerse en cuenta que si se efectúa ignorando el problema concreto al que se refiere será, precisamente, un análisis estadístico ignorante.

Pero a pesar de que los métodos cuantitativos tienen limitaciones, también ofrecen un amplio abanico de posibilidades para extraer la información que contienen los números. Desde esta perspectiva, abordamos la elaboración de este libro con la intención de facilitar una primera visión, muy simple, respecto a las técnicas más elementales para proceder al análisis de las series de datos.

Carla Rey Graña
María Ramil Díaz

Contenido

TEMA 1

ELABORACIÓN DE TABLAS DE DATOS ESTADÍSTICOS

1.1. Datos estadísticos

La primera utilidad de la estadística es la de proporcionar un conjunto de normas que permite elaborar las tablas numéricas adecuadas para cuantificar un determinado fenómeno. Dicho fenómeno está formado por un conjunto de personas o cosas que llamamos **población**. Las personas o cosas que integran una población se denominan **elementos** o **unidades estadísticas**.

Cada uno de los elementos de una población puede describirse según uno o varios caracteres. Por ejemplo, si el objetivo de un determinado análisis son los viajeros que llegan a un país, puede centrarse la atención en su edad, su estado civil, su país de procedencia, etcétera.

Si estamos interesados en el estudio de la edad de los viajeros, se trata de analizar una característica que toma valores numéricos, a la cual denominamos **variable**. En cambio, si nos interesa, por ejemplo, su país de procedencia, se trata de analizar una característica cualitativa, denominada **atributo**. Los atributos no son, pues, susceptibles de cuantificarse en la forma convencional mediante una escala numérica.

Al observar las diferentes variables o atributos se obtiene un conjunto de resultados, numérico o no, denominado **conjunto de datos**. Los obtenidos al observar un atributo se denominan **modalidades**, mientras que los correspondientes a una variable se denominan **valores**.

En función del número de valores que pueden tomar las variables, se distinguen las discretas de las continuas. Las discretas son aquellas que pueden tomar un número finito (o infinito numerable) de valores; por ejemplo, el número de visitantes que recibe un país, el número de plazas de un hotel, etcétera. Las continuas son aquellas que pueden tomar infinitos valores dentro de un intervalo; el gasto medio por persona de los turistas que visitan un país es, por ejemplo, una variable continua.

Según el tipo de valores observados, se distinguen las variables temporales de las atemporales. Las primeras hacen referencia la misma unidad estadística en diferentes períodos de tiempo (meses, trimestres, años o cualquier otra unidad temporal); por ejemplo, la tasa de ocupación bruta de un hotel determinado cada año de los comprendidos entre 1975 y 2000. Las segundas se refieren a diferentes unidades estadísticas en un mismo período de tiempo; por ejemplo, la tasa de ocupación bruta de distintos hoteles en el año 2000.

1.2. Elaboración de tablas estadísticas unidimensionales

1.2.1. Tablas de distribución de frecuencias con datos sin agrupar

Nos ocupamos ahora de la elaboración de una tabla estadística referida a una sola variable; es decir, que sólo recoge la información correspondiente a una característica de cada uno de los elementos de la población.

Supongamos, por ejemplo, que hemos obtenido información numérica sobre los precios de los menús servidos durante un día en un restaurante determinado. Dicha información es la siguiente:

Precio = {6, 8, 6, 8, 6, 8, 12, 6, 8, 8, 6, 8, 8, 8, 12, 12, 8, 8, 12, 6, 8, 6, 6, 8, 12, 6, 6, 6, 6, 6} euros

Podemos observar que sólo hay tres precios diferentes, 6, 8 y 12 euros, que se repiten, y además sus valores están desordenados. La información así presentada resulta muy poco manejable a efectos de su análisis estadístico. En primer lugar debemos ordenar, de menor a mayor precio, por ejemplo, y agrupar los valores comunes en una tabla. A este proceso se le denomina **tabulación**.

En la columna de la izquierda de la tabla se presentan los valores x_i de los tres precios correspondientes a los menús que se han servido el día elegido, mientras que en la columna de la derecha figura el número de veces que se repite cada uno de ellos. En estadística, el número de veces que se repite cada valor o dato de la variable se denomina **frecuencia absoluta** (o simplemente frecuencia) y, en general, se representa por n_i.

En este ejemplo $n_1 = 13$, lo que significa que el primer valor de la variable se repite 13 veces; es decir, que se han servido 13 menús de 6 euros; $n_2 = 12$, lo que significa que el segundo valor de la variable se repite 12 veces; es decir, que se han servido 12 menús de 8 euros y $n_3 = 5$, lo que significa que el tercer valor de la variable se repite 5 veces; es decir, que se han servido 5 menús de 12 euros.

Tabla 1.1. Precio de los menús

Precio (x_i)	Número de menús servidos (n_i)
$x_1 = 6$	$n_1 = 13$
$x_2 = 8$	$n_2 = 12$
$x_3 = 12$	$n_3 = 5$
Total	30

Sumando las frecuencias absolutas se obtiene el número total de valores observados de la variable, que representamos por N.

En este ejemplo: $N = n_1 + n_2 + n_3 = 13 + 12 + 5 = 30$.

En general: $N = n_1 + n_2 + ... + n_n = \sum_{i=1}^{n} n_i$

Ahora bien, la frecuencia absoluta no da una idea respecto a si es o no elevada. Para saberlo, debemos referirla al conjunto de los datos.

Tabla 1.2. Frecuencia absoluta y relativa

x_i	n_i	f_i
6	13	$f_1 = n_1 / N = 0,43$
8	12	$f_2 = n_2 / N = 0,40$
12	5	$f_3 = n_3 / N = 0,17$
Total	30	1,00

Definimos así la **frecuencia relativa**, f_i, que se obtiene por cociente entre la frecuencia absoluta (n_i) y el número total de datos (N).

Dada su definición, es obvio que el valor mínimo de la frecuencia relativa es cero y su valor máximo es la unidad. Por tanto, la frecuencia relativa es tanto más elevada cuanto más próximo está su valor a uno.

La Tabla 1.2 recoge los valores de las frecuencias relativas para la distribución de los precios. En ella, las frecuencias relativas están expresadas en tanto por uno, pero también se pueden expresar en tanto por ciento, multiplicando por cien cada uno de sus valores. Tendremos, así, que el 43 por ciento de los menús servidos ha sido de 6 euros, el 40 por ciento ha sido de 8 euros y el 17 por ciento restante ha sido de 12 euros.

La suma de las frecuencias relativas es igual a la unidad:

En este ejemplo: $f_1 + f_2 + f_3 = 0{,}43 + 0{,}40 + 0{,}17 = 1$.

En general: $\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} \frac{n_i}{N} = \frac{n_1}{N} + \frac{n_2}{N} + ... + \frac{n_n}{N} = \frac{n_1 + n_2 + ... + n_n}{N} = \frac{N}{N} = 1$

También puede tener interés calcular las **frecuencias acumuladas**, tanto absoluta como relativa.

La **frecuencia acumulada absoluta**, que representamos por N_i, indica el número de valores de la variable iguales al considerado o inferiores a él, y se obtiene sumando, para cada valor, su frecuencia absoluta más las correspondientes a los valores anteriores de la variable.

Tabla 1.3. Frecuencia acumulada absoluta

x_i	n_i	f_i	N_i
6	13	0,43	$N_1 = n_1 = 13$
8	12	0,40	$N_2 = n_1 + n_2 = 25$
12	5	0,17	$N_3 = n_1 + n_2 + n_3 = 30$
Total	30	1,00	

N_2 representa la frecuencia absoluta acumulada correspondiente al segundo valor de la variable (8 euros) que se cifra en 25, lo que nos indica que se han servido 25 menús con un precio igual o inferior a 8 euros. Podemos observar, además, que la

primera frecuencia absoluta acumulada es igual a la primera frecuencia absoluta, y que la última frecuencia absoluta acumulada coincide con el número de datos disponibles, que en nuestro ejemplo son 30.

En general: $N_n = n_1 + n_2 + \ldots + n_n = \sum_{i=1}^{n} n_i = N$

La **frecuencia acumulada relativa**, representada por F_i, se obtiene al dividir cada frecuencia acumulada absoluta (N_i) entre el número total de datos (N), o bien sumando, para cada valor, su frecuencia relativa más las correspondientes a los valores anteriores de la variable.

Tabla 1.4. Frecuencia acumulada relativa

x_i	n_i	f_i	N_i	F_i
6	13	0,43	13	$F_1 = N_1 / N = f_1$ = 0,43
8	12	0,40	25	$F_2 = N_2 / N = f_1 + f_2$ = 0,83
12	5	0,17	30	$F_3 = N_3 / N = f_1 + f_2 + f_3$ = 1,00
Total	30	1,00		

La frecuencia acumulada relativa correspondiente al segundo valor de la variable, F_2, indica, pues, que el 83 por ciento de los menús servidos en el restaurante tiene un precio igual o inferior a 8 euros. Como puede observarse, la primera frecuencia relativa acumulada es igual a la primera frecuencia relativa, y la última frecuencia relativa acumulada es igual a la unidad.

En general: $F_n = f_1 + f_2 + \ldots + f_n = \sum_{i=1}^{n} f_i = 1$

Por último, los datos pueden presentarse en una tabla que resume todo lo expuesto anteriormente.

Esta tabla recoge los valores de la variable y sus frecuencias, absolutas y relativas, simples y acumuladas. Dado su contenido, se la conoce con el nombre de **tabla de distribución de frecuencias**. Considerando que todas las demás pueden obtenerse a partir de la frecuencia absoluta, se representa como los diferentes valores que en cada caso toma el par (x_i, n_i).

Tabla 1.5. Distribución de frecuencias

x_i	n_i	f_i	N_i	F_i
6	13	0,43	13	0,43
8	12	0,40	25	0,83
12	5	0,17	30	1,00
Total	30	1,00		

Hasta ahora, como hemos visto, la información se ha dispuesto asociando a cada valor de la variable su frecuencia. Esta forma de presentar los datos se utiliza cuando la variable toma un pequeño número de valores distintos. Ahora bien, si la variable toma un número grande o muy grande de valores distintos y disponemos los datos de esta manera, se obtienen unas columnas muy largas, que no proporcionan la visión de conjunto deseada.

1.2.2. Tablas de distribución de frecuencias con datos agrupados

Si el número de valores que toma la variable es suficientemente grande resulta aconsejable, para una mayor comodidad en el tratamiento de la información, agrupar estos valores en un número reducido de clases o **intervalos**. La agrupación de los datos facilita su manejo, pero debe tenerse en cuenta que, mientras que en las distribuciones no agrupadas disponemos de toda la información correspondiente a una variable, en las distribuciones agrupadas se pierde parte de la información.

Por ejemplo, supongamos que se dispone de los datos correspondientes a los precios pagados por las consumiciones realizadas en una cafetería a lo largo de un determinado día. Dado que existen muchos valores diferentes de la variable comprendidos entre 0 y 15 euros, para tener una visión de conjunto, las cantidades pagadas se han agrupado en clases o intervalos. Los valores de la variable y sus correspondientes frecuencias absolutas se presentan en la Tabla 1.6, que recoge en la primera columna los cinco intervalos considerados, de los cuales L_{i-1} es el extremo inferior y L_i es el extremo superior.

Como puede observarse, la tabla proporciona información respecto a cuántas consumiciones de entre 3 y 6 euros se han servido en el establecimiento, pero no indica si ha habido consumiciones de 4 ó 5 euros, o cuántas se han servido de cada

uno de los importes comprendidos en el intervalo. Obviamente, para la elaboración de la tabla es necesario recoger el máximo de información, pero como se ha señalado, al hacer la agrupación, aunque se facilita el tratamiento de los datos, parte de la información se pierde.

Tabla 1.6. Precio de las consumiciones

Precio ($L_{i-1} - L_i$)	Número de consumiciones (n_i)
0 – 3	40
3 – 6	30
6 – 9	10
9 – 12	5
12 – 15	5

Para agrupar los datos de una variable en intervalos o clases es necesario, en primer lugar, conocer el **recorrido** o **rango** de la variable, que es la diferencia entre su mayor y menor valor.

En el ejemplo anterior el recorrido de la variable es de 15 – 0 = 15 euros.

A continuación, hemos de decidir la **amplitud** de los intervalos, que se representa como c_i y es la diferencia entre su extremo superior y su extremo inferior; es decir, $c_i = L_i - L_{i-1}$.

En general es recomendable que, en la medida de lo posible, los intervalos sean de amplitud constante, ya que en algunos aspectos es más sencillo el tratamiento estadístico de la información. No obstante, por encima de este requisito, debe estar el de que la tabla estadística resultante exprese lo más fielmente posible las características de la variable, por lo que en algunos casos será preferible que los intervalos sean de amplitud variable.

En el ejemplo anterior, los intervalos considerados tienen una amplitud constante, igual a 3 euros.

Cuando la amplitud es constante, el recorrido es igual al número de intervalos por su amplitud. Luego, si se conoce la amplitud puede deducirse el número de

intervalos; o bien, si se fija el número de intervalos puede determinarse la amplitud. Para establecer el número de intervalos que deben considerarse no hay reglas fijas, aunque en la práctica suele oscilar entre 5 y 15.

Una vez establecido el número de intervalos a considerar y su amplitud, se plantea el problema de determinar a qué intervalo corresponde un valor de la variable cuando coincide con uno de sus extremos. Habitualmente se considera que los intervalos son semiabiertos por la izquierda, lo cual se denota como $(L_{i-1} - L_i]$, y significa que se componen de todos aquellos valores comprendidos entre L_{i-1} y L_i, incluido el extremo superior y excluido el inferior.

En el ejemplo que hemos planteado, una consumición de 6 euros puede estar, en principio, incluida en el intervalo comprendido entre 3 y 6 euros o en el siguiente, de 6 a 9 euros. Si los intervalos considerados son semiabiertos por la izquierda, como es habitual, está incluida en el intervalo de 3 a 6 euros.

Sin embargo, esta no es la única posibilidad, puesto que también pueden definirse intervalos cerrados $[L_{i-1} - L_i]$ que incluyen los dos extremos, abiertos $(L_{i-1} - L_i)$ que no incluyen ninguno de los dos extremos, o semiabiertos por la derecha $[L_{i-1} - L_i)$ que incluyen el extremo inferior pero no el superior.

Tabla 1.7. Precio de las consumiciones

$L_{i-1} - L_i$	x_i	n_i
0 – 3	1,5	40
3 – 6	4,5	30
6 – 9	7,5	10
9 – 12	10,5	5
12 – 15	13,5	5

Por último, como valor representativo de cada intervalo se toma su punto medio $x_i = (L_{i-1} + L_i) / 2$, al que se denomina **marca de clase**.

Si en el ejemplo considerado se calculan las marcas de clase, la tabla de frecuencias absolutas de la variable es tal como la Tabla 1.7. Para completar la

distribución de frecuencias con datos agrupados, debe seguirse el procedimiento que se ha descrito en el caso de los datos sin agrupar.

En cualquiera de los dos casos, el proceso de obtención de esta tabla es mucho más sencillo utilizando una hoja de cálculo.

1.2.3. Elaboración de la tabla de distribución de frecuencias con EXCEL

El **programa** se ejecuta haciendo doble clic con el ratón en el icono de acceso directo o bien pinchando el botón **Inicio–Programas–EXCEL**.

De esta forma, se abre una hoja de cálculo en blanco. La primera operación que se debe realizar para elaborar la tabla de distribución de frecuencias es la introducción de los datos.

Por ejemplo, supongamos que se dispone de la información correspondiente a la edad de los clientes de un determinado establecimiento hotelero:

Edad = {19, 25, 32, 44, 51, 28, 32, 20, 60, 45, 54, 23, 25, 36, 40, 45, 33, 22, 55, 48, 25, 40, 23, 45, 33, 19, 36, 22, 48, 32, 25, 36, 28, 19, 25, 22, 32, 44, 45, 55, 32, 25, 60, 40, 45, 33, 22, 28, 44, 45} años

Se sitúa el cursor con el ratón en la celda A1, en la que se introduce x_i como nombre de la variable. A continuación, en la celda A2, se teclea el primer dato (19), en la celda A3 el segundo (20), y así sucesivamente, hasta que se hayan introducido todos los datos en celdas consecutivas de la columna A. Para pasar de una celda a otra, puede utilizarse el ratón, las teclas de dirección o la tecla Intro.

El siguiente paso consiste en ordenar los datos de menor a mayor. Para ello se pincha el menú **Datos**, en el que se elige la opción **Ordenar**. Al efectuar esta operación, se abre un cuadro de diálogo en el que se debe indicar que queremos ordenar los datos en orden ascendente según los valores de la variable x_i. Finalmente, se pincha **Aceptar** para ejecutar la operación.

Una vez ordenados los datos, para obtener los valores de las frecuencias absolutas, es necesario contar cuántas veces se repite cada uno de los valores de la variable.

Para efectuar esta operación, nos situamos en la celda B1 y le damos el nombre n_i. En la celda B2 se pincha en el menú **Insertar**, el submenú **Función**. Elegimos las funciones **Estadísticas** y dentro de ellas **Contar**. A continuación se abre un cuadro de diálogo en el que se indica el rango de celdas que queremos que cuente. Como, en primer lugar, queremos saber cuántas celdas contienen el número 19, debemos indicar el rango **A2:A4**, pinchando con el ratón la celda A2 y, sin soltar, arrastrándolo hasta la celda A4. Al pinchar **Aceptar** aparecerá un número 3, que indica que hay 3 celdas con un valor igual a 19.

Tabla 1.8. Frecuencias absolutas

	A	B	C
1	x_i	n_i	
2	19	3	
3	20	1	
4	22	4	
5	23	2	
6	25	6	
7	28	3	
8	32	5	
9	33	3	
10	36	3	
11	40	3	
12	44	3	
13	45	6	
14	48	2	
15	51	1	
16	54	1	
17	55	2	
18	60	2	
19			

Nos situamos ahora en la celda que está a la derecha del siguiente valor, es decir en la celda B5, y efectuamos el mismo proceso para contar cuántas celdas contienen el valor 20. La operación se repite hasta finalizar. Luego se eliminan las filas en las que la

celda correspondiente a la frecuencia está vacía, sombreándolas con el ratón y la instrucción **Eliminar** del menú **Edición**.

Se obtiene así la Tabla 1.8, que contiene los valores de la variable y sus correspondientes frecuencias absolutas.

Para completar la tabla de distribución, deben obtenerse el resto de las frecuencias. Comenzamos por obtener la columna correspondiente a las frecuencias relativas. Como se ha indicado, dichas frecuencias se obtienen dividiendo cada frecuencia absoluta entre el número total de datos, que es la suma de las frecuencias absolutas.

Para obtener el número de datos, se sitúa el cursor en la celda B19, y se ejecuta la instrucción **Insertar–Función–Matemáticas–Suma**, que abre un cuadro de diálogo en el que debemos indicar el rango de celdas que se desea sumar. En este caso, dicho rango es **B2:B18**. Al **Aceptar**, en dicha celda aparece el resultado de la suma; en este caso, 50.

Tabla 1.9. Obtención de la distribución de frecuencias

	A	B	C	D	E	F
1	x_i	n_i	f_i	N_i	F_i	
2	19	3	=B2/B$19	=B2	=D2/B$19	
3	20	1		=B3+D2		
4	22	4				
...	...	...	...	...	...	
18	2	2				
19		Σ(B2:B18)=50				
20						

Una vez efectuada esta operación, se sitúa el cursor en la celda C2, para indicar la fórmula para el cálculo de las frecuencias relativas. La correspondiente al primer valor de la variable se obtiene dividiendo su frecuencia absoluta, que es el valor que ocupa la celda B2, entre el número total de datos, que es el valor que ocupa la celda B19. Puede observarse que entre la letra B y el número 19 se inserta el símbolo $, para indicar que al copiar esta fórmula en las demás casillas siempre se ha de tomar como denominador el valor fijo contenido en la celda B19.

Para obtener el valor de la segunda frecuencia relativa se selecciona con el ratón la celda C2, que contiene la fórmula para el cálculo de f_1, y en el menú **Edición** elegimos **Copiar**. Se coloca ahora el ratón en la celda C3, se pincha y sin soltar el botón, se arrastra hasta la celda C18, y en el menú **Edición** se selecciona **Pegar** (que, en las tablas hemos representado con el símbolo ). Con este procedimiento se obtienen las frecuencias relativas para los demás valores de la variable.

A continuación, se formulan las columnas D y E para obtener las frecuencias acumuladas, absolutas y relativas, de la forma que se indica en la Tabla 1.9. Una vez efectuado el proceso descrito, debe obtenerse un resultado tal como el que se muestra a continuación:

Tabla 1.10. Distribución de frecuencias

	A	B	C	D	E	F
1	x_i	n_i	f_i	N_i	F_i	
2	19	3	0,06	3	0,06	
3	20	1	0,02	4	0,08	
4	22	4	0,08	8	0,16	
5	23	2	0,04	10	0,20	
6	25	6	0,12	16	0,32	
7	28	3	0,06	19	0,38	
8	32	5	0,10	24	0,48	
9	33	3	0,06	27	0,54	
10	36	3	0,06	30	0,60	
11	40	3	0,06	33	0,66	
12	44	3	0,06	36	0,72	
13	45	6	0,12	42	0,84	
14	48	2	0,04	44	0,88	
15	51	1	0,02	45	0,90	
16	54	1	0,02	46	0,92	
17	55	2	0,04	48	0,96	
18	60	2	0,04	**50**	**1,00**	
19		**50**	**1,00**			
20						

Los resultados obtenidos en esta tabla deben ser tales que:

1. La suma de la columna de frecuencias absolutas es el número de datos

2. La suma de la columna de frecuencias relativas es igual a la unidad

3. El valor de la última frecuencia absoluta acumulada es el número de datos

4. El valor de la última frecuencia relativa acumulada es igual a la unidad

Si fuese necesario corregir la entrada de alguno de los datos iniciales, la hoja de cálculo opera de tal manera que automáticamente corrige todas las operaciones efectuadas con él.

Supongamos ahora que el hotel al que se ha hecho referencia pretende ofrecer un determinado servicio dirigido a sus clientes más jóvenes y que, para analizar si resultará rentable, desea conocer el porcentaje de clientes que tienen una edad igual o inferior a 35 años y cuál es el número de clientes que tienen una edad superior a 35 años.

Esta información puede obtenerse a partir de la tabla de distribución de frecuencias.

1. Porcentaje de clientes con edad igual o inferior a 35 años:

Dado que no hay clientes de 34 ni de 35 años, debemos determinar el porcentaje de clientes con edad igual o inferior a 33 años. La frecuencia relativa acumulada correspondiente a 33 años (0,54) indica que el 54 por ciento de los clientes tienen una edad igual o inferior a 33 años.

2. Número de clientes con edad superior a 35 años:

Observando la columna de las frecuencias absolutas acumuladas, sabemos que hay 27 personas que tienen una edad igual o inferior a 33 años. Si el número total de clientes es de 50, 23 (= 50 – 27) tienen una edad superior a 33 y, por tanto, a 35 años, ya que no hay clientes de 34 ni de 35 años.

Aunque la tabla de distribución de frecuencias recoge toda la información disponible respecto a una variable, puede resultar útil traducirla en un gráfico que permita asimilar rápidamente y sin esfuerzo sus principales características.

1.3. Representaciones gráficas

Las representaciones gráficas son un medio complementario para describir el fenómeno que se trata de analizar. Entre ellas, puede hacerse referencia, básicamente, a los diagramas de líneas, de barras, a los histogramas de frecuencias y a los diagramas de sectores.

Los diagramas de líneas y de barras se utilizan cuando la variable toma un número reducido de valores diferentes; es decir, con distribuciones no agrupadas en intervalos. Estas representaciones gráficas se realizan mediante un sistema de ejes de coordenadas cartesianas, tomando, generalmente, en el eje de abscisas, la escala para los valores de la variable y en el de ordenadas, la escala para los valores de las frecuencias. Si se representa un punto para cada par formado por un valor de la variable y su correspondiente frecuencia, (x_i, n_i), se obtiene un conjunto de puntos. Si los puntos unen con una línea, se obtiene el **diagrama de líneas**. Si para cada valor de la variable se traza una barra vertical con altura igual a la frecuencia, se obtiene el **diagrama de barras**.

El **histograma de frecuencias** se utiliza cuando los valores de la variable están agrupados en intervalos. Se utiliza, también, un sistema de ejes de coordenadas cartesianas. En el eje de abscisas se sitúan los intervalos de la variable y, sobre ellos, tomándolos como base, se construyen rectángulos de tal forma que su área sea igual a la frecuencia absoluta de cada intervalo.

Si los intervalos son de amplitud constante, la altura de los rectángulos coincidirá con las correspondientes frecuencias absolutas, ya que al ser las bases iguales, las áreas dependerán sólo de la altura. En cambio, si los intervalos son de amplitud variable, la altura de los rectángulos debe coincidir con la **densidad de frecuencia** (d_i), que se define como el cociente entre la frecuencia absoluta y la amplitud de cada intervalo. Así, el área del rectángulo (= base x altura) coincide con la frecuencia absoluta del intervalo.

En efecto:

Área = base x altura = $c_i \times d_i = c_i \times (n_i / c_i) = n_i$

En general, para representar fenómenos cualitativos, suelen utilizarse los **diagramas de sectores**, aunque éste no es un uso exclusivo, puesto que también se utilizan de forma generalizada cuando se trabaja con variables. El diagrama consiste en dividir un círculo en sectores cuyo ángulo central sea proporcional a la frecuencia absoluta correspondiente y, por consiguiente, su área resulte también proporcional a dicha frecuencia.

Veamos algunos **ejemplos** de cada uno de estos tipos de representaciones gráficas, utilizando la hoja de cálculo EXCEL.

En primer lugar, representamos gráficamente los 50 valores de la variable x_i = edad de los clientes de un establecimiento hotelero, contenidos en la Tabla 1.8, a la que se ha hecho referencia en el epígrafe 1.2.3. Para ello, en la hoja que contiene los datos de la variable, se sitúa el cursor en una celda en blanco y en el menú **Insertar** se elige la opción **Gráfico**. Esta instrucción inicia el asistente para gráficos.

Se abre entonces un cuadro de diálogo en el que se debe determinar el tipo de gráfico. Supongamos que se elige el tipo **Líneas**. Pinchando con el ratón el botón **Siguiente**, el asistente pide información respecto a si los datos de la variable que se desea representar se han introducido en filas o en columnas. Marcamos la casilla de verificación Columnas. A continuación debe indicarse el rango de datos que se desea representar, que está formado por las celdas en las que figuran los valores de las frecuencias absolutas, ya que éstas son las que determinan las alturas.

Se pincha, entonces, la pestaña Series, para indicar el nombre de la variable, que figura en la celda A1, y el rango de los Rótulos del eje de categorías (x), formado por las celdas correspondientes a los valores de la variable, A2:A18.

El tercer paso del asistente para gráficos permite añadir, si se desea, el título, alguna leyenda, la ubicación, etcétera.

Una vez finalizado el gráfico, el cuadro de diálogo ofrece dos posibilidades: situar el gráfico en una hoja nueva, o bien insertarlo como un objeto en la misma hoja en la que se han introducido los datos. El resultado obtenido será similar al que se presenta en el Gráfico 1.1.

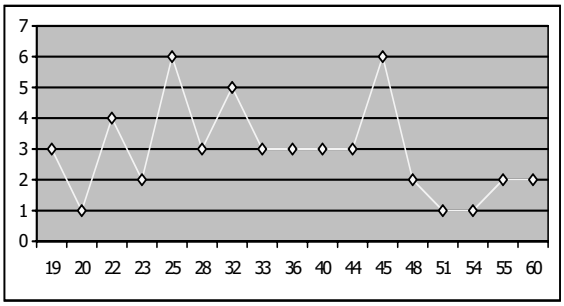


Gráfico 1.1. Edad de los clientes

Siguiendo el procedimiento descrito, pero eligiendo como tipo de gráfico **Columnas**, se obtiene el diagrama de barras, que será análogo al siguiente:

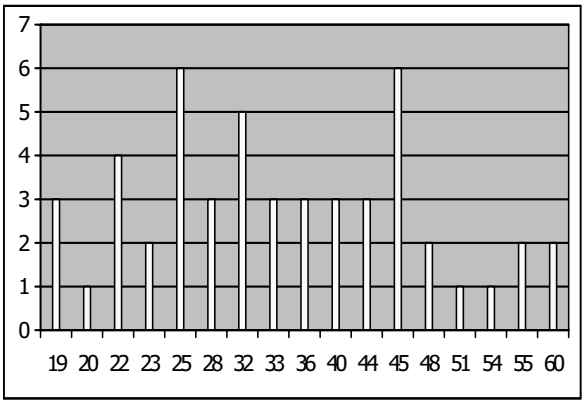


Gráfico 1.2. Edad de los clientes

Si como tipo de gráfico se elige **Barras**, se toma la escala para los valores de la variable en el eje de ordenadas y para las frecuencias en el de abscisas, al revés de lo que hemos hecho hasta ahora. Veamos, a continuación, la representación de esta misma tabla de datos seleccionando el tipo de gráfico Barras.

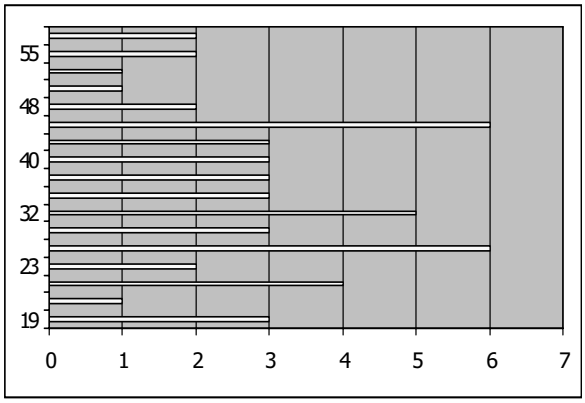


Gráfico 1.3. Edad de los clientes

Supongamos ahora que se desea efectuar la representación gráfica de los datos que contiene la Tabla 1.6, relativos a la variable x_i = precio de las consumiciones realizadas durante un día en una cafetería, a la que se ha hecho referencia en el epígrafe 1.2.2.

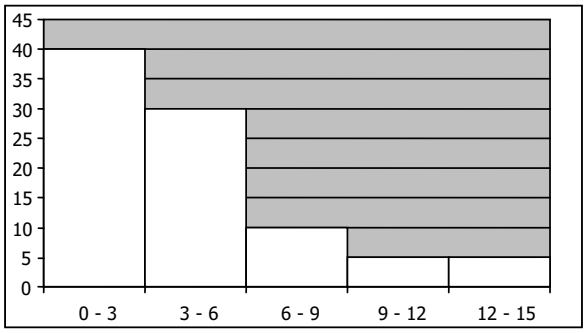


Gráfico 1.4. Precio de las consumiciones

Para hacer este gráfico con EXCEL, el procedimiento es el mismo que en el caso del diagrama de barras. Para indicar que el ancho de las barras coincide con el tamaño del intervalo, ha de seleccionarse Opciones dentro del Formato de Series de Datos, que aparece al pulsar el botón derecho del ratón cuando están seleccionadas las barras, e introducir el valor cero como ancho de rango.

Para ver como se hace el gráfico cuando los intervalos son de amplitud variable, vamos a utilizar la información contenida en la Tabla 1.11, que se refiere a la variable x_i = número de días de estancia de los clientes de un hotel.

Tabla 1.11. Días de estancia

L_{i-1}- L_i	n_i	$d_i = n_i / c_i$
2 – 4	6	3,00
4 – 7	5	1,66
7 – 10	7	2,33
10 – 20	6	0,60

En este caso, el procedimiento es el mismo que en los anteriores, excepto porque en el rango de datos deben indicarse las celdas en las que figuran los valores de las densidades de frecuencia, obtenidas por cociente entre la frecuencia absoluta y la amplitud del intervalo. El resultado obtenido será similar al Gráfico 1.5, donde puede observarse que todas las columnas tienen el mismo ancho, como si todos los intervalos tuviesen la misma amplitud, pero la altura de las barras no coincide con la frecuencia, sino con la densidad de frecuencia de cada intervalo.

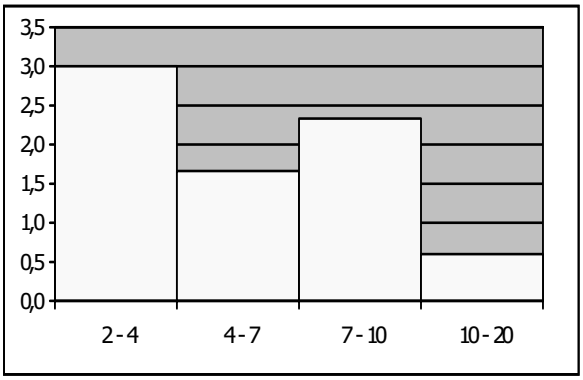


Gráfico 1.5. Días de estancia

También podemos representar la variable x_i = precio de las consumiciones, a la que ya nos hemos referido, empleando el diagrama de sectores. Para ello, se elige el

tipo de gráfico **Circular** y el resultado obtenido será como el que se presenta en el Gráfico 1.6.

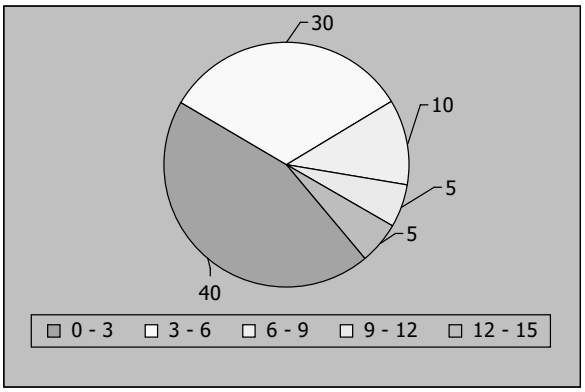


Gráfico 1.6. Precio de las consumiciones

El gráfico de sectores puede utilizarse también para representar un atributo. Por ejemplo, supongamos que se dispone de la información correspondiente a la distribución por motivo del viaje de los visitantes de un determinado punto turístico que se presenta en la Tabla 1.12.

Tabla 1.12. Distribución por motivo del viaje

Motivo del viaje	Número visitantes
Vacacional	100
Religioso	5
Negocios	30
Visitas familiares y amigos	20

Con estos datos, y seleccionando el tipo **Circular**, pero esta vez con la opción **Efectos en 3 dimensiones**, obtendremos un resultado similar al que se recoge en el Gráfico 1.7.

Aunque los que hemos visto hasta ahora son los tipos de gráfico más utilizados, la hoja de cálculo EXCEL ofrece muchas otras posibilidades. Como ejemplo, hemos recogido en los Gráficos 1.8 y 1.9 la información correspondiente a la distribución por motivo del viaje, con algunos otros. En particular, el Gráfico 1.8 corresponde al tipo **Cilíndrico** y el Gráfico 1.9 al tipo **Barras rústico**.

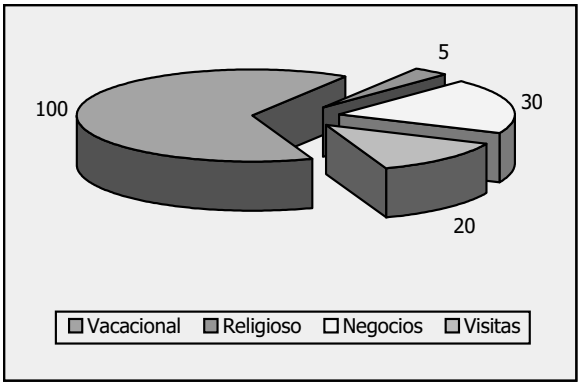


Gráfico 1.7. Motivo del viaje

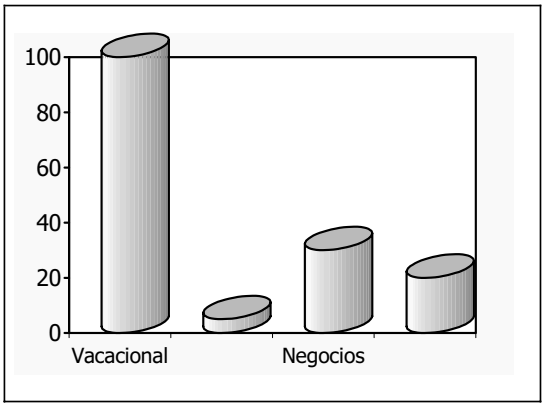


Gráfico 1.8. Motivo del viaje

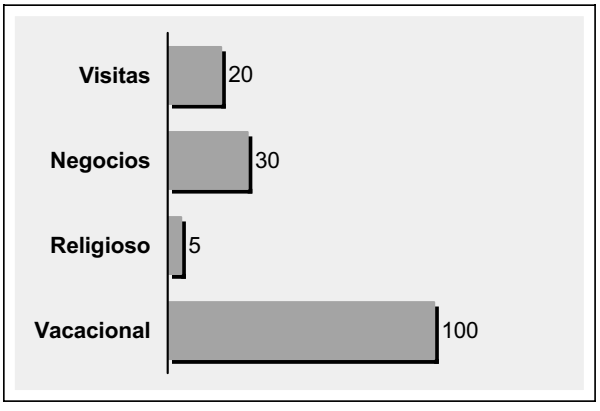


Gráfico 1.9. Motivo del viaje

TEMA 2
ANÁLISIS DE TABLAS DE DATOS ESTADÍSTICOS

2.1. Introducción

La tabla de distribución de frecuencias proporciona una información detallada del comportamiento de una variable, cuyas características principales pueden ser rápidamente visualizadas mediante un gráfico. Pero, en muchos casos, es interesante disponer también de algunas medidas que resuman adecuadamente la información contenida en los datos, principalmente por lo que se refiere a los valores centrales, la dispersión, la concentración y la forma. El estudio conjunto de todas estas características permite tener una visión condensada y completa del fenómeno a analizar.

Así, por ejemplo, en el análisis de la edad de los clientes de un determinado establecimiento hotelero puede ser interesante conocer qué edad tienen por término medio, cuál es la edad que tienen la mayoría de los clientes, qué edad deja por encima y por debajo de ella el mismo número de clientes, etcétera. Las medidas estadísticas que se utilizan para cuantificar estas características que se refieren a los valores centrales de las variables se denominan **medidas de posición**.

Estas medidas de tendencia central sintetizan todos los valores de la distribución, pero proporcionan sólo una información parcial, especialmente si la variable toma valores muy diferentes entre sí, puesto que, en tal caso, los valores centrales de la serie son poco representativos de la serie en su conjunto. Por eso debe valorarse también la disparidad o la dispersión de los datos.

En el ejemplo propuesto, la edad de los clientes en promedio condensa la información disponible respecto a los valores de la variable edad, pero no proporciona ninguna información respecto a su variabilidad. Sería interesante valorar también la disparidad de las edades, lo cual puede hacerse mediante la utilización de algunas **medidas de dispersión**.

Los valores de las medidas de dispersión son, como hemos dicho, indicativos de la mayor o menor disparidad entre los valores de la variable, pero no proporcionan detalles respecto a la amplitud del intervalo en el cual se concentran la mayoría de dichos valores.

En el ejemplo que venimos analizando, es interesante conocer la edad de los clientes en promedio y algún indicador respecto al grado de dispersión de las edades. Pero dados los valores de estas medidas, la serie es muy distinta si la mayoría de las edades están muy concentradas en torno al promedio y unos pocos valores son muy distintos que si la distribución se reparte aproximadamente por igual entre todos los valores de la variable. Para proceder al análisis de esta cuestión se definen las **medidas de concentración**.

Por último, también proporcionan información relevante las medidas relativas a la forma del histograma de frecuencias y, en especial, a su simetría o asimetría y a su mayor o menor apuntamiento.

La distribución de las edades es más o menos simétrica si el número de valores de la edad que están por encima y por debajo de uno dado es aproximadamente el mismo. Si la distribución de las edades es apuntada, la mayoría de los valores de la edad se concentran en torno a uno dado, mientras que si es plana, los valores están menos concentrados. Para cuantificar estas características de la distribución, se utilizan las **medidas de forma**.

2.2. Medidas de posición de tendencia central y no centrales

2.2.1. Medidas de posición de tendencia central

La **media aritmética**, o más abreviadamente el **promedio** o la **media**, es la medida de posición de tendencia central más utilizada. Su concepto es muy sencillo, puesto que se trata de repartir por igual el valor global o conjunto de las observaciones de la variable entre el número de datos disponibles.

Se define, por tanto, por cociente entre la suma de los valores de la variable y el número total de datos. Al calcularla debe tenerse en cuenta que para obtener la suma de los valores de la variable es necesario sumar los productos de cada valor por el número de veces que se repite; es decir, por su frecuencia absoluta.

Para dar la definición de la media en términos analíticos, se la representa como $\bar{x}$, mientras que, como sabemos, x_i hace referencia a cada valor de la variable x y n_i a la frecuencia correspondiente a cada valor.

De esta forma:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_n n_n}{N} = \frac{\sum_{i=1}^{n} x_i n_i}{N}$$

La media tiene en cuenta, por tanto, todos los valores observados de la variable. Es uno de sus valores, probablemente no observado, y viene dado en las mismas unidades de medida.

Veamos ahora, con un ejemplo, cómo se calcula la media aritmética en el caso más sencillo de **distribuciones no agrupadas** en intervalos.

En la Tabla 2.1 se presentan los datos relativos a la edad de los participantes en una actividad recreativa ofrecida en un paquete turístico.

Tabla 2.1. Edad de los participantes

Edad (x_i)	N° de participantes (n_i)
18	2
20	14
24	16
26	11
38	2
46	2
50	3

Para obtener el promedio de edad de los participantes en la actividad, se construye la Tabla 2.2, en la que se añaden a la Tabla 2.1, la columna que recoge los productos de los valores de la variable por sus correspondientes frecuencias y la fila que recoge las sumas de las frecuencias absolutas y de los productos de los valores de la variable por sus frecuencias absolutas.

Tabla 2.2. Cálculos intermedios para la obtención de la media

x_i	n_i	$x_i n_i$
18	2	36
20	14	280
24	16	384
26	11	286
38	2	76
46	2	92
50	3	150
Total	50	1.304

De la información que proporciona la tabla, se deduce el valor medio de la edad:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_7 n_7}{N} = \frac{\sum_{i=1}^{7} x_i n_i}{N} = \frac{1.304}{50} = 26{,}08$$

Luego, los participantes en la actividad tienen una media de edad de 26 años, aproximadamente.

En el caso de las **distribuciones agrupadas** en intervalos, para el cálculo de la media aritmética se asume la concentración en el punto medio de todos los valores incluidos en cada intervalo y, por tanto, se utiliza la marca de clase como valor representativo.

Tabla 2.3. Establecimientos hoteleros según el número de habitaciones

Nº de habitaciones ($L_{i-1} - L_i$)	Nº de establecimientos (n_i)
0 – 15	3
15 – 30	1
30 – 50	3
50 – 60	1
60 – 115	1

Por ejemplo, supongamos que se desea conocer cuántas habitaciones tienen, por término medio, los hoteles ubicados en un determinado municipio, y que se dispone de la información recogida en la Tabla 2.3.

El procedimiento de cálculo es exactamente el mismo que en el caso de las distribuciones sin agrupar, excepto porque en este caso, ha de calcularse previamente la marca de clase de cada intervalo, tal como se muestra en la Tabla 2.4.

Tabla 2.4. Cálculos intermedios para la obtención de la media

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$
0 – 15	7,50	3	22,50
15 – 30	22,50	1	22,50
30 – 50	40,00	3	120,00
50 – 60	55,00	1	55,00
60 – 115	87,50	1	87,50
Total		9	307,50

De la información que proporciona esta tabla, se deduce la media de habitaciones por hotel:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_5 n_5}{N} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{307,5}{9} = 34,17$$

Luego, por término medio, los hoteles del municipio tienen unas 34 habitaciones.

El **cálculo de la media** puede realizarse rápidamente **con EXCEL**.

Si los valores de la variable se repiten una sola vez, bien porque no existen valores iguales, bien porque no se han agrupado los valores comunes, basta con introducir los datos y situar el cursor en una celda vacía, en la cual se ejecuta la instrucción **Insertar–Función–Estadísticas–Promedio**, especificando el rango en el que se han introducido los datos. Si las frecuencias absolutas no son unitarias, la media se obtiene siguiendo las instrucciones que se presentan en la Tabla 2.5, que proporcionarán un resultado similar al mostrado en la Tabla 2.6.

En dichas tablas, x_i representa los distintos valores de la variable x en el caso de distribuciones sin agrupar y de las marcas de clase de los correspondientes intervalos en el caso de distribuciones agrupadas.

El ejemplo numérico elegido para ilustrar el procedimiento corresponde a los datos relativos a la edad de los participantes en la actividad recreativa ofrecida en el paquete turístico al que se refiere la Tabla 2.1.

Tabla 2.5. Procedimiento para la obtención de la media con EXCEL

	A	B	C	D
1	x_i	n_i	$x_i n_i$	
2	18	2	=A2*B2	
3	20	14	(1)	
4	…	…	…	
8	50	3	(1)	
9	Total	Σ (B2:B8) (2)	Σ (C2:C8) (2)	
10	Media	=C9/B9		
11				

(1) Previamente debe copiarse la celda C2
(2) Insertar–Función–Matemáticas–Suma, para el correspondiente rango de valores

Tabla 2.6. Obtención de la media con EXCEL

	A	B	C	D
1	x_i	n_i	$x_i n_i$	
2	18	2	36	
3	20	14	280	
4	24	16	384	
5	26	11	286	
6	38	2	76	
7	46	2	92	
8	50	3	150	
9	Total	50	1.304	
10	Media	26,08		
11				

La media aritmética tiene algunas **propiedades** interesantes, que son aplicables en distintos contextos. Entre otras, están las siguientes:

1. La suma de las desviaciones de los valores de la variable con respecto a su media aritmética es cero.

Esta propiedad puede expresarse empleando el lenguaje matemático de la siguiente forma:

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

En efecto:

Suma de frecuencias absolutas = N° de datos (N)

$$\sum_{i=1}^{n}(x_i - \bar{x})n_i = \sum_{i=1}^{n}(x_i n_i - \bar{x}n_i) = \sum_{i=1}^{n} x_i n_i - \bar{x}\sum_{i=1}^{n} n_i = N\frac{\sum_{i=1}^{n} x_i n_i}{N} - \bar{x}N = N\bar{x} - \bar{x}N = 0$$

Se multiplica y divide por N

Media aritmética ($\bar{x}$)

2. Si a todos los valores de una variable se les suma una constante k, la media aritmética queda aumentada en esa constante; es decir, el valor de la media se ve afectado por los cambios de origen.

Para demostrarlo, partimos de una distribución (x_i, n_i), cuya media aritmética es $\bar{x}$. Consideramos ahora una nueva distribución, resultado de efectuar un cambio de origen a la variable x, (x_i^o, n_i). Según esta propiedad, la media aritmética de la nueva distribución será $\bar{x}^o = \bar{x} + k$.

En efecto:

$$\bar{x}^o = \frac{\sum_{i=1}^{n} x_i^o n_i}{N} = \frac{\sum_{i=1}^{n}(x_i + k)n_i}{N} = \frac{\sum_{i=1}^{n}(x_i n_i + k n_i)}{N} = \frac{\sum_{i=1}^{n} x_i n_i}{N} + \frac{k\sum_{i=1}^{n} n_i}{N} = \bar{x} + k$$

3. Si todos los valores de una variable se multiplican por una constante k, su media aritmética también queda multiplicada por esa constante; es decir, a la media aritmética le afectan los cambios de escala.

Para demostrarlo, partimos de una distribución (x_i, n_i), cuya media aritmética es $\bar{x}$. Consideramos ahora una nueva distribución, resultado de efectuar un cambio de escala a la variable x, (x_i^e, n_i). Según esta propiedad, la media aritmética de la nueva distribución será $\bar{x}^e = k\ \bar{x}$.

En efecto:

$$\bar{x}^e = \frac{\sum_{i=1}^{n} x_i^e n_i}{N} = \frac{\sum_{i=1}^{n} (kx_i) n_i}{N} = \frac{k\sum_{i=1}^{n} x_i n_i}{N} = k\bar{x}$$

4. La suma de los cuadrados de las desviaciones de los valores observados de la variable respecto a una constante cualquiera k es mínima cuando dicha constante es la media aritmética. Esta propiedad puede expresarse en términos matemáticos de la siguiente forma:

$$S = \sum_{i=1}^{n} (x_i - k)^2 n_i \text{ es mínima si } k = \bar{x}$$

Para demostrar esta propiedad, recordamos que la primera condición de mínimo de una función es que sea nula su derivada. En este caso, la derivada de S con respecto a k debe ser nula. Por tanto, se obtiene la derivada de S con respecto a k, se iguala a cero y se resuelve, para deducir el valor que ha de tener k para que la condición de mínimo se cumpla.

$$\frac{\delta S}{\delta k} = -2\sum_{i=1}^{n} (x_i - k) n_i = 0$$

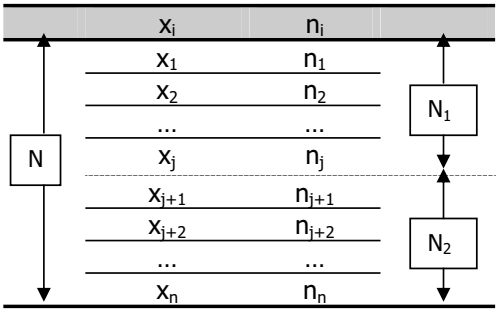
$$\sum_{i=1}^{n} (x_i - k) n_i = \sum_{i=1}^{n} (x_i n_i - k n_i) = \sum_{i=1}^{n} x_i n_i - k\sum_{i=1}^{n} n_i = \sum_{i=1}^{n} x_i n_i - kN = 0$$

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

5. Si de un conjunto de datos se obtienen p subconjuntos disjuntos; es decir, que no tienen ningún elemento en común, la media aritmética del conjunto se relaciona con las medias aritméticas de los subconjuntos a través de la expresión:

$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \ldots + \bar{x}_p N_p}{N} = \frac{\sum_{i=1}^{p} \bar{x}_i N_i}{N}$$

En efecto, consideremos una distribución formada por los N valores de una variable, con la que se forman dos subconjuntos que no tienen valores en común: el primero de ellos, contiene los j primeros valores de la variable y el segundo contiene los N – j elementos restantes, de tal manera que el primer subconjunto tiene N_1 datos y el segundo $N_2 = N - N_1$ datos y $N_1 + N_2 = N$.



Entonces:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{\sum_{i=1}^{j} x_i n_i + \sum_{i=j+1}^{n} x_i n_i}{N} = \frac{N_1 \dfrac{\sum_{i=1}^{j} x_i n_i}{N_1} + N_2 \dfrac{\sum_{i=j+1}^{n} x_i n_i}{N_2}}{N} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N}$$

Analizamos, a continuación, algunos **ejemplos** prácticos en los que resulta útil aplicar estas propiedades:

1. Cinco empresarios distintos han elaborado el presupuesto para un determinado complejo turístico. Las valoraciones de cada uno de ellos, en miles de euros, han sido las siguientes:

Tabla 2.7. Presupuestos

Empresario	Presupuesto
A	782
B	824
C	752
D	788
E	842

Como cálculo definitivo se optó por considerar la media de estos cinco valores, de manera que el presupuesto se cifró en 797,6 miles de euros.

Al revisar las cifras, sin embargo, se observó que ninguno de los empresarios había tenido en cuenta los gastos de primer establecimiento, que ascienden a 60 mil euros, y que deberían haber sido sumados en todos los presupuestos propuestos antes de calcular la media. Teniendo en cuenta esta información, ¿cuál sería el presupuesto correcto?

Si a todos los valores de la variable se le deben sumar k = 60 mil euros (es decir, 60 unidades, puesto que los presupuestos están expresados en miles de euros) de gastos de primer establecimiento, $x_i^o = x_i + k = x_i + 60$, lo que equivale a efectuar un cambio de origen de la variable. Como $\bar{x}^o = \bar{x} + k = 797,6 + 60 = 857,6$, luego el presupuesto alcanzaría la cifra de 857,6 miles de euros.

2. Sabiendo que un euro equivale a 166,386 pesetas, ¿cuál sería el valor del presupuesto presentado como definitivo en el apartado anterior expresado en miles de pesetas?

Se trata de un cambio en las unidades de medida de la variable; es decir, de un cambio de escala, puesto que cada valor de la variable en miles de pesetas sería $x_i^e = 166,386x_i = kx_i$. Como $\bar{x}^e = k\,\bar{x} = 166,386 \times 857,6 = 142.692,634$. Luego el presupuesto medio, una vez efectuado el cambio de escala, sería de 142.692,634 miles de pesetas.

3. De una determinada distribución de frecuencias se sabe que la suma de los cuadrados de las desviaciones de los valores observados de la variable respecto a una constante cualquiera k es mínima cuando k = 57. ¿Qué información proporciona este dato?

Sabemos que $\sum_{i=1}^{n}(x_i - k)^2 n_i$ es mínima cuando $k = \bar{x}$.

Luego si esta expresión es mínima cuando la constante k es igual a 57, 57 es el valor de la media aritmética.

4. En un parque de atracciones la media de gastos por visitante es de 24 euros. Sabiendo que los adultos gastan una media de 26 euros y los niños gastan una media de 21 euros, calcule el porcentaje de adultos y de niños que visitan el parque.

Sabemos que la media de gasto del conjunto de los visitantes del parque es de 24 euros. Conocemos también los valores de las medias correspondientes a dos subconjuntos disjuntos formados, el primero de ellos, por los adultos, cuya media de gasto es de 26 euros y el segundo, por los niños, cuya media de gasto es de 21 euros. Finalmente, también sabemos que el total de visitantes del parque es la suma de los adultos y los niños, $N = N_1 + N_2$, luego si al parque entran N_1 adultos, los niños son $N_2 = N - N_1$.

La media del conjunto de los datos se relaciona con la media de los subconjuntos a través de la expresión:

$$\overline{x} = \frac{N_1\overline{x}_1 + N_2\overline{x}_2}{N}$$

Luego:

$$24 = \frac{N_1 26 + (N - N_1)21}{N}$$

de donde:

$$24N = 26N_1 + 21N - 21N_1$$

$$N_1 = \frac{3}{5}\ N$$

Y el porcentaje de adultos sobre el total es:

$$P_1 = \frac{N_1}{N} \times 100 = \frac{3/5N}{N} \times 100 = 60$$

Es decir, el 60 por ciento de los visitantes del parque son adultos y el 40 por ciento restante son niños.

En el análisis cuantitativo del Sector Turístico también es de uso frecuente la **media aritmética ponderada**, que se denota como $\bar{x}_w$. La peculiaridad es que, en el cálculo de esta media, se le asigna a cada valor de la variable una ponderación o peso (w_i) distinto de su correspondiente frecuencia absoluta.

La expresión general para su cálculo es la siguiente:

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_n w_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

Veamos, con un ejemplo, en qué casos puede tener interés, y cuál es el procedimiento de cálculo para la obtención de esta medida.

Supongamos que se dispone de los datos relativos al gasto diario estimado en euros de los turistas de una determinada zona, según la categoría a la que corresponde el establecimiento hotelero en el que se alojan.

Tabla 2.8. Gasto diario

Categoría	Gasto (x_i)	n_i
Tres estrellas	120	1
Dos estrellas	60	1
Una estrella	30	1

Si para calcular el gasto medio diario se utiliza una media aritmética simple, se deduce que por término medio los turistas gastan al día 70 euros, ya que:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3}{N} = \frac{\sum_{i=1}^{3} x_i n_i}{N} = \frac{210}{3} = 70$$

Ahora bien, supongamos que se dispone, además, de la información recogida en la Tabla 2.9 respecto al número de hoteles de cada categoría que se ubican en dicha zona. Sabemos, por tanto, que en la zona se ubican 80 hoteles, de los cuales casi el 60 por ciento son de una estrella, algo más del 30 por ciento son de dos estrellas y sólo un 10 por ciento son de tres estrellas. Para obtener un valor del gasto medio más

realista, por tanto, el de los turistas alojados en hoteles de una estrella debe pesar en el cálculo de la media el 60 por ciento del total, el de los alojados en hoteles de dos estrellas el 30 por ciento del total y el de los alojados en hoteles de tres estrellas el 10 por ciento del total.

Tabla 2.9. Gasto diario

Categoría	Nº de hoteles	Gasto (x_i)	n_i
Tres estrellas	8	120	1
Dos estrellas	25	60	1
Una estrella	47	30	1

Por eso, en un caso como éste, sería más apropiado calcular una media aritmética ponderada, utilizando como ponderaciones el número de establecimientos de cada categoría.

De este modo, se tiene:

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + x_3 w_3}{w_1 + w_2 + w_3} = \frac{\sum_{i=1}^{3} x_i w_i}{\sum_{i=1}^{3} w_i} = \frac{8 \times 120 + 25 \times 60 + 47 \times 30}{8 + 25 + 47} = 48{,}38$$

Luego el gasto medio diario sería de 48,38 euros, frente a los 70 euros que se obtienen como media aritmética simple.

Para la **obtención de la media ponderada con EXCEL**, el procedimiento es el mismo que el que se utiliza para obtener la media simple, sustituyendo en las correspondientes instrucciones el valor de la frecuencia absoluta por el de los factores de ponderación para cada valor.

Otra de las medidas de posición de tendencia central es la **mediana**, que se denota como M_e, y es el valor de la variable que sitúa el mismo número de datos por encima que por debajo de él, siempre y cuando los valores de la variable estén ordenados de forma creciente o decreciente, como es habitual.

Si las **frecuencias absolutas** son **unitarias**; es decir, cada valor de la variable se repite una sola vez, bien porque no existen valores iguales, bien porque no se han agrupado los valores comunes, para obtener el valor de la mediana han de considerarse dos situaciones:

1. El **número de datos** es **impar**: en este caso la mediana es el valor de la variable que ocupa el lugar central de la serie.

Así, si los datos de la variable son $\{2, 5, 6, 9, 15\}$ la mediana es $M_e = 6$, ya que este valor deja dos datos por encima y dos por debajo de él.

2. El **número de datos** es **par**: en este caso, en la distribución hay dos valores centrales, y ninguno de ellos es exactamente la mediana, sino que, generalmente, se toma como mediana la media aritmética de los dos valores.

Por ejemplo, si los datos de la variable son $\{3, 4, 7, 9, 15, 20, 23, 24\}$, los valores centrales son 9 y 15, que dejan tres datos por encima y tres por debajo de ellos. La mediana será, pues, $M_e = (9 + 15) / 2 = 12$.

Cuando se trata de obtener la mediana para **distribuciones no agrupadas** en intervalos y con frecuencias absolutas distintas de uno, en primer lugar, se calculan las frecuencias absolutas acumuladas correspondientes a cada valor de la variable. A continuación, separamos la distribución en dos mitades, calculando el valor de ½ N. Finalmente, se busca, en la columna correspondiente a las frecuencias absolutas acumuladas, el primer valor que es superior o igual a ½ N.

Si la frecuencia absoluta acumulada supera a ½ N, la mediana es el valor x_i de la variable correspondiente a dicha frecuencia. En el caso de que la frecuencia absoluta acumulada coincida exactamente con ½ N, convencionalmente se toma como mediana la media aritmética del valor x_i de la variable al que corresponde dicha frecuencia y el siguiente, x_{i+1}.

Veamos un par de ejemplos, con los datos que figuran en las Tablas 2.10 y 2.11, que corresponden al número de personas que practican una determinada actividad de tiempo libre según las horas semanales que le dedican, en dos establecimientos distintos.

En el primer caso, el número total de datos es N = 25. Luego, ½ N = 12,5. La primera frecuencia absoluta acumulada mayor que 12,5 es $N_3 = 14$. Por tanto, la mediana es el valor de la variable al que le corresponde $N_3 = 14$; es decir, $x_3 = 5$. Por tanto, en el primer establecimiento, el número de personas que dedican 5 horas o menos y 5 horas o más a la semana a dicha actividad es el mismo.

Tabla 2.10. Obtención de la mediana en el Caso 1

x_i	n_i	N_i	
1	2	2	
3	4	6	
5	8	14	$N_3 > ½ N$
7	6	20	
9	5	25	

$M_e = 5$

Tabla 2.11. Obtención de la mediana en el Caso 2

x_i	n_i	N_i	
1	2	2	
3	4	6	
5	6	12	$N_3 = ½ N$
7	7	19	
9	5	24	

$M_e = ½ (5+7)$

En el segundo caso, el número total de datos es N = 24. Luego, ½ N = 12. Al observar la columna de las frecuencias absolutas acumuladas resulta que N_3 vale 12; es decir, que coincide exactamente con ½ N. El valor de la variable al que le corresponde $N_3 = 12$ es $x_3 = 5$, y la mediana es:

$$M_e = \frac{x_i + x_{i+1}}{2} = \frac{5+7}{2} = 6$$

Es decir, en el segundo establecimiento, el número de personas que dedican 6 horas o menos y 6 horas o más a la semana a dicha actividad es el mismo.

Para **distribuciones agrupadas** en intervalos, independientemente de que su amplitud sea constante o variable, este procedimiento permite determinar el intervalo en el que se encuentra la mediana, pero no su valor, que se aproxima mediante la expresión:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i$$

donde L_{i-1} es el límite inferior del intervalo correspondiente a la primera frecuencia acumulada mayor o igual que ½ N, c_i es la amplitud del intervalo, n_i es la frecuencia absoluta correspondiente a dicho intervalo y N_{i-1} es la frecuencia acumulada correspondiente al intervalo inmediatamente anterior.

Para ilustrarla, utilizamos la distribución de frecuencias que figura en la Tabla 2.12, que se refiere al número de viajes diarios de los autobuses de una determinada compañía, clasificados en función de la distancia recorrida en kilómetros.

Tabla 2.12. Obtención de la mediana

$L_{i-1} - L_i$	n_i	N_i	
0 – 2	14	14	
2 – 4	16	30	
4 – 6	28	58	$N_3 > ½ N$
6 – 8	24	82	
8 – 10	18	100	

I. M_e = 4–6

El número total de datos es N = 100. Luego, ½ N = 50. La primera frecuencia absoluta acumulada mayor que 50 es N_3 = 58; por tanto, la mediana estará en el intervalo 4 – 6. Para concretar su valor:

$$M_e = 4 + \frac{50 - 30}{28} \times 2 = 5{,}43$$

Es decir, 5,43 kilómetros es la distancia que deja igual número de viajes por debajo que por encima de ella, o lo que es lo mismo, los autobuses de esta compañía

efectúan el mismo número de viajes a una distancia inferior o igual a 5,43 kilómetros que a una distancia igual o superior a 5,43 kilómetros.

También en las distribuciones agrupadas en intervalos puede suceder, aunque es poco frecuente, que la frecuencia absoluta acumulada coincida exactamente con ½ N. En tal caso, la mediana coincide exactamente con el límite superior del intervalo al que corresponda dicha frecuencia acumulada.

Para el **cálculo de la mediana con EXCEL** en el caso de frecuencias unitarias, se introducen los valores de la variable y se sitúa el cursor en una celda vacía. Se selecciona en el menú la opción Insertar–Función–Estadísticas–Mediana, y en el cuadro de diálogo que se abre, se indica el rango en el que se han introducido los datos. Al Aceptar, se obtiene el valor de la mediana.

Cuando las frecuencias no son unitarias, una opción es desagrupar los datos para que tengan frecuencias unitarias, utilizando adecuadamente las instrucciones Copiar y Pegar, y utilizar el procedimiento automático descrito en el párrafo anterior. Otra opción es calcular las frecuencias absolutas acumuladas en la forma descrita en el epígrafe 1.2.3, y efectuar las operaciones matemáticas que definen la mediana, según que se trate de una distribución agrupada o no agrupada en intervalos.

La **moda** es el valor de la variable que más veces se repite; es decir, el que tiene una mayor frecuencia absoluta. En general, no suele utilizarse como única medida de posición, sino que se presenta acompañando a la media o a la mediana. La excepción es el caso de los atributos, ya que es la única de las medidas de tendencia central que puede obtenerse.

En el caso de **distribuciones no agrupadas** en intervalos, la determinación de la moda es inmediata. Simplemente, es el valor x_i de la variable al que le corresponde un valor mayor de la frecuencia absoluta n_i.

Veamos un ejemplo, con los datos del número de personas que practican una determinada actividad de tiempo libre según las horas semanales que le dedican, a los que ya nos hemos referido.

En el primer establecimiento, la moda es $x_3 = 5$ horas, puesto que la mayor frecuencia absoluta registrada es $n_3 = 8$.

Tabla 2.13. Obtención de la moda en el Caso 1.

	x_i	n_i	
	1	2	
	3	4	
$M_o = 5$ ←	5	8	→ $>n_i = 8$
	7	6	
	9	5	

Tabla 2.14. Obtención de la moda en el Caso 2.

	x_i	n_i	
	1	2	
	3	4	
	5	6	
$M_o = 7$ ←	7	7	→ $>n_i = 7$
	9	5	

En el segundo establecimiento, sin embargo, la moda es $x_4 = 7$ horas, puesto que la mayor frecuencia absoluta registrada es $n_4 = 7$.

Para las **distribuciones agrupadas en intervalos de igual amplitud**, se utiliza este mismo procedimiento para determinar el intervalo modal. Pero, de forma similar a lo que sucede en el caso de la mediana, hemos de concretar el valor aproximado de la moda a través de la siguiente expresión:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \times c_i$$

donde L_{i-1} es el extremo inferior del intervalo al que le corresponde una mayor frecuencia absoluta, n_{i-1} es la frecuencia absoluta correspondiente al intervalo inmediatamente anterior al modal, n_{i+1} es la frecuencia absoluta del intervalo siguiente al modal y c_i es la amplitud del intervalo modal, que, en este caso de amplitud constante, coincidirá con la de los demás.

Así, por ejemplo, la Tabla 2.15 recoge la información correspondiente al número de servicios prestados por las empresas de restauración de una determinada zona, agrupados en intervalos de igual amplitud.

Tabla 2.15. Obtención de la moda

L_{i-1} - L_i	n_i	
0 – 50	3	
50 – 100	5	
100 – 150	12	$>n_i = 12$
150 – 200	8	
200 – 250	2	

I. M_o

La frecuencia absoluta más elevada es $n_3 = 12$. Sabemos, pues, que la moda está en el intervalo 100 – 150. Para concretar aproximadamente su valor:

$$M_o = 100 + \frac{8}{5+8} \times 50 = 130{,}77$$

El valor de la distribución que más veces se repite es, por tanto, 130,77.

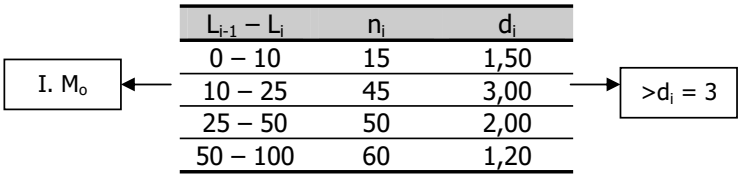
En el caso de **distribuciones agrupadas en intervalos de amplitud variable**, para obtener la moda, en lugar de las frecuencias absolutas se utilizan las densidades de frecuencia, y el intervalo modal es aquél al que le corresponde la mayor. Para concretar su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i$$

que es idéntica a expresión utilizada en el caso de distribuciones agrupadas en intervalos de igual amplitud, excepto porque, como hemos dicho, las densidades de frecuencia sustituyen a las frecuencias absolutas.

Para ilustrar el procedimiento para su obtención, se utilizan los datos que corresponden a la distribución de frecuencias de la variable número de empleados en actividades del sector hostelería en un determinado municipio, que se recogen en la Tabla 2.16.

Tabla 2.16. Obtención de la moda



$L_{i-1} - L_i$	n_i	d_i
0 – 10	15	1,50
10 – 25	45	3,00
25 – 50	50	2,00
50 – 100	60	1,20

La densidad de frecuencia más elevada es $d_2 = 3$, luego la moda estará en el intervalo 10 – 25. Para determinar su valor aproximado:

$$M_o = 10 + \frac{2}{1,5+2} \times 15 = 18,57$$

Luego, 18,57 es la aproximación del valor de esta distribución que más veces se repite.

En todas las distribuciones que hemos visto hasta ahora la moda es única. Sin embargo, dada su definición, es obvio que esto no es necesariamente así. Con frecuencia, podemos encontrar distribuciones bimodales, trimodales, etcétera, que tienen más de una moda.

Para la **obtención de la moda con EXCEL** sólo es necesario obtener los valores de las frecuencias absolutas siguiendo el proceso descrito en el epígrafe 1.2.3. y, en el caso de distribuciones agrupadas, efectuar las operaciones matemáticas que definen la moda según que los intervalos sean de amplitud constante o variable.

En definitiva, tanto la moda como la mediana y la media son indicadores de la tendencia central de las variables, y su cálculo es un simple proceso mecánico que no tiene ninguna dificultad.

La medida más utilizada es, sin duda, la media aritmética, en la que se basan multitud de técnicas de análisis estadístico. Sin embargo, cuando en la distribución hay valores extremos, la media, al tener en cuenta todos y cada uno de los valores de la variable, puede proporcionar una representación distorsionada de la tendencia central

de los datos. En este caso, la mediana o la moda, a las que no les afectan los valores extremos, proporcionan un valor más realista de dicha tendencia.

Por otra parte, si los datos son de naturaleza cualitativa, ni la media ni la mediana pueden obtenerse, de tal manera que la moda es la única medida de tendencia central disponible.

2.2.2. Medidas de posición no centrales

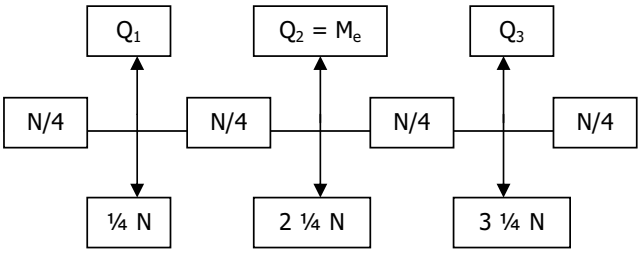
Los indicadores analizados en el epígrafe anterior proporcionan un valor central que puede considerarse que sintetiza o condensa el conjunto de valores de una variable. De las tres medidas que se han considerado, sólo la mediana hace referencia a una determinada posición dentro del conjunto de valores de la variable, puesto que deja la mitad de los valores de la variable por debajo y por encima de ella.

Esta idea respecto a la posición se puede generalizar definiendo un conjunto de indicadores, que se denominan cuantiles, que permiten determinar que posición ocupa un valor concreto de la variable respecto al conjunto de sus valores, puesto que son los puntos de la distribución que la dividen en intervalos que comprenden el mismo número de valores.

Entre los cuantiles de uso más frecuente están los cuartiles, que son los tres valores de la distribución que la dividen en cuatro partes iguales; es decir, en cuatro intervalos dentro de cada cual están incluidos el 25 por ciento de los valores de la variable. De forma similar pueden definirse los quintiles, que dividen la distribución en cinco partes iguales, los deciles, que la dividen en diez partes iguales, y los percentiles, que la dividen en 100 partes iguales. Nosotros nos centraremos solamente en los cuartiles.

Tal como se muestra en el diagrama de la página siguiente, el primer cuartil Q_1 es el valor de la distribución que ocupa el lugar ¼ N, y deja a su izquierda el 25 por ciento de los datos y el 75 por ciento restante a su derecha. El segundo cuartil Q_2 es valor de la distribución que ocupa el lugar 2 ¼ N = ½ N, que deja a su izquierda el 50 por ciento de los datos y el 50 por ciento restante a su derecha, de tal manera que coincide con la mediana. El tercer y último cuartil Q_3 ocupa el lugar 3 ¼ N, y deja a su

izquierda el 75 por ciento de los datos y a su derecha el 25 por ciento restante. El método de cálculo es similar al que hemos visto para la mediana.



En el caso de **distribuciones no agrupadas** en intervalos, en primer lugar, deben obtenerse las frecuencias absolutas acumuladas correspondientes a cada valor de la variable. Para obtener el primer cuartil, se divide el número total de datos entre cuatro, se selecciona el primer valor de la columna de frecuencias absolutas acumuladas mayor o igual que ¼ N y el valor de la variable al que le corresponde esta frecuencia acumulada es el primer cuartil. Para obtener el segundo o el tercer cuartil hemos de calcular 2 ¼ N = ½ N o 3 ¼ N, respectivamente, y repetir el procedimiento descrito.

Así por ejemplo, supongamos que se dispone de los datos recogidos en la Tabla 2.17, respecto a los precios, expresados en euros, de las bebidas servidas en una cafetería y al número de consumiciones realizadas a cada precio.

El primer cuartil Q_1 es el valor de la distribución que ocupa el lugar ¼ N = 18,75. El primer valor de la columna de frecuencias absolutas acumuladas que lo supera es N_1 = 20. Por tanto, Q_1 = 1. Luego en el grupo formado por el 25 por ciento de las bebidas más baratas servidas, el precio máximo es de un euro.

Podemos también calcular el precio más bajo que tendrá una bebida en el grupo formado por el 25 por ciento de las bebidas más caras. Esta cifra es el valor que toma el tercer cuartil, Q_3 que es el que corresponde a la primera frecuencia absoluta acumulada mayor o igual a 3 ¼ N = 56,25. Por tanto, Q_3 es igual a 3 ya que 65 es el primer valor de la columna correspondiente a las frecuencias absolutas acumuladas mayor que 56,25.

Tabla 2.17. Obtención de los cuartiles

	x_i	n_i	N_i	
$Q_1 = 1$ ←	1,00	20	20	→ $N_1 > ¼ N$
	2,00	15	35	
	2,50	12	47	
$Q_3 = 3$ ←	3,00	18	65	→ $N_4 > 3 ¼ N$
	3,50	10	75	

En el caso de **distribuciones agrupadas** en intervalos, al igual que sucede con la mediana, el hecho de que los intervalos sean de amplitud constante o variable no afecta al cálculo de los cuartiles.

Siguiendo el método que acabamos de exponer, se obtiene el intervalo en el que se encuentran los cuartiles. Su valor aproximado puede concretarse por medio de la expresión:

$$Q_r = L_{i-1} + \frac{\frac{rN}{4} - N_{i-1}}{n_i} \times c_i$$

donde r toma el valor 1, 2 ó 3 según que se trate del primer, segundo o tercer cuartil, L_{i-1} es el límite inferior del intervalo correspondiente a la primera frecuencia absoluta acumulada mayor o igual que ¼ N, c_i es la amplitud del intervalo, n_i su frecuencia absoluta y N_{i-1} la frecuencia absoluta acumulada correspondiente al intervalo inmediatamente anterior.

Para ilustrar el caso de las distribuciones agrupadas, se dispone de la información que proporciona la Tabla 2.18, en la que figuran los sueldos mensuales de los empleados de un establecimiento hotelero en un año determinado.

Los sueldos varían entre 361 y 3.006 euros. Utilizando los cuartiles, se puede dividir la distribución en cuatro partes iguales, de tal forma que, por ejemplo, el tercer cuartil indica cuánto gana como mínimo un empleado que se sitúa en el grupo formado por el 25 por ciento de los empleados que mayor sueldo tienen.

Tabla 2.18. Obtención de los cuartiles

L_{i-1} - L_i	n_i	N_i
361 – 541	10	10
541 – 722	20	30
722 – 1.082	30	60
1.082 – 1.322	20	80
1.322 – 1.803	10	90
1.803 – 3.006	5	95

I. Q_3 ← 1.082 – 1.322 → $N_4 > 3\ ¼\ N$

Para conocer el valor de la distribución que ocupa el lugar 3 ¼ N = 71,25, en primer lugar, se determina en qué intervalo está situado. La primera frecuencia absoluta acumulada mayor que 71,25 es N_4 = 80, luego el tercer cuartil tomará un valor comprendido entre 1.082 y 1.322 euros. Para concretarlo:

$$Q_3 = L_{i-1} + \frac{\frac{3N}{4} - N_{i-1}}{n_i} \times C_i = 1.082 + \frac{71,25 - 60}{20} \times 240 = 1.217$$

Por lo tanto, como mínimo, un empleado que se sitúa en el grupo formado por el 25 por ciento de los que mayor sueldo tienen, gana 1.217 euros.

Al igual que ocurría con la mediana, la **obtención de los cuartiles con EXCEL** es muy sencilla, siempre y cuando los valores de la variable se repitan una sola vez, bien porque no existen valores iguales, bien porque no se han agrupado los valores comunes. Para ello, después de introducir los datos, se sitúa el cursor en una celda en blanco y se selecciona en el menú Insertar, la opción Función–Estadísticas–Cuartiles. En el cuadro de diálogo que se abre se indica, en Matriz, el rango en el que hemos introducido los valores de la variable, y en Cuartil, 1, 2 ó 3 según queramos calcular el primer, segundo o tercer cuartil.

Cuando las frecuencias no son unitarias, una opción es desagrupar los datos para que tengan frecuencias unitarias, utilizando adecuadamente las instrucciones Copiar y Pegar, y utilizar el procedimiento automático descrito en el párrafo anterior. Otra opción es calcular las frecuencias absolutas acumuladas en la forma descrita en el epígrafe 1.2.3, y efectuar las operaciones matemáticas que definen los cuartiles, según que se trate de una distribución agrupada o no agrupada en intervalos.

2.3. Medidas de dispersión absolutas y relativas

Las medidas de posición de tendencia central permiten condensar la información disponible respecto a una variable, pero su utilidad depende de hasta qué punto estas medidas de síntesis son representativas de la distribución.

Si, por ejemplo, determinamos la media aritmética de la edad de dos individuos que tienen 10 y 50 años, el resultado que se obtiene es el mismo que si calculamos la media de edad de otros dos individuos que tienen 25 y 35 años. En cualquiera de los dos casos, la edad promedio es de 30 años.

Sin embargo, en el primer caso, 30 es un valor escasamente representativo de los valores 10 y 50, que están considerablemente alejados de él, puesto que la desviación de cada valor con respecto a la media es de 20 unidades. En el segundo caso, sin embargo, 30 es un valor razonablemente representativo de 25 y 35, que no son valores excesivamente alejados de él, puesto que la desviación de cada valor respecto a la media es de 5 unidades.

Si queremos conocer en qué grado la media aritmética de una distribución resume adecuadamente el conjunto de los datos, debe analizarse la desviación de cada valor respecto a la media, ya que si todos los valores están próximos a ella, es representativa, pero en caso contrario, es poco representativa de la serie de datos y sólo proporciona una información parcial respecto a su comportamiento.

En consecuencia, para evitar conclusiones erróneas, y completar la información que proporcionan las medidas de posición, hemos de presentarlas acompañadas de, al menos, una medida de dispersión que indique el grado de proximidad que existe entre los valores de la variable.

Esta **variabilidad** o **dispersión** puede valorarse a partir de la mayor o menor desviación de los valores de la variable respecto a cualquiera de las medidas de posición de tendencia central. Dado que la más utilizada es la media aritmética, en este epígrafe nos referimos a las medidas de dispersión respecto a ella, absolutas, como la varianza y la desviación típica, y relativas, como el coeficiente de variación de Pearson.

2.3.1. Medidas de dispersión absolutas

La **varianza**, s^2, es un indicador de la dispersión de los datos con respecto al valor medio de la variable. Se define como el promedio de los cuadrados de las desviaciones de los valores de la variable respecto a su media aritmética, ponderados por sus respectivas frecuencias:

$$s^2 = \frac{(x_1 - \overline{x})^2 n_1 + (x_2 - \overline{x})^2 n_2 + ... + (x_n - \overline{x})^2 n_n}{N} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i}{N}$$

Para calcular la varianza no es preciso hallar las diferencias entre los valores de la variable y la media para elevarlas al cuadrado, ya que, operando en el numerador de la fracción que la define:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i = \sum_{i=1}^{n} (x_i^2 - 2\overline{x}x_i + \overline{x}^2) n_i = \sum_{i=1}^{n} (x_i^2 n_i - 2\overline{x}x_i n_i + \overline{x}^2 n_i) =$$

$$= \sum_{i=1}^{n} x_i^2 n_i - 2\overline{x}\sum_{i=1}^{n} x_i n_i + \overline{x}^2 \sum_{i=1}^{n} n_i = \sum_{i=1}^{n} x_i^2 n_i - 2\overline{x}(N\overline{x}) + \overline{x}^2 N = \sum_{i=1}^{n} x_i^2 n_i - N\overline{x}^2$$

Entonces:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 n_i}{N} = \frac{\sum_{i=1}^{n} x_1^2 n_i - N\overline{x}^2}{N} = \frac{\sum_{i=1}^{n} x_1^2 n_i}{N} - \overline{x}^2$$

El valor numérico de la varianza es mayor cuanto más dispersos están los valores de la variable y, por tanto, cuanto menos representativa es la media.

Dada su definición, la varianza está expresada en las unidades de medida de la variable elevadas al cuadrado. Para expresar la variabilidad en las mismas unidades que la variable, se define la **desviación típica** o **desviación estándar**, s, como la raíz cuadrada positiva de la varianza.

Veamos, ahora, con un ejemplo, cómo se calculan la varianza y la desviación típica en el caso más sencillo de **distribuciones no agrupadas** en intervalos. Para ello, vamos a utilizar de nuevo los datos correspondientes a la edad de las 50 personas que participaron en una actividad recreativa, que han servido ya en el epígrafe 2.2.1 para ilustrar el cálculo de la media aritmética.

Tabla 2.19. Cálculos intermedios para la obtención de la varianza

x_i	n_i	$x^2_i n_i$
18	2	648
20	14	5.600
24	16	9.216
26	11	7.436
38	2	2.888
46	2	4.232
50	3	7.500
Total	50	37.520

Sabemos que la edad promedio de los participantes es de 26,08 años. Para obtener la varianza, se construye la columna de productos $x_i^2 n_i$ y la fila de totales de esta columna y de la de frecuencias absolutas.

Con esta información, se tiene:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{37.520}{50} - 26{,}08^2 = 70{,}23 \Rightarrow s = \sqrt{70{,}23} = 8{,}38$$

La desviación típica de la edad de los participantes en la actividad es, por tanto, 8,38 años.

Al igual que para la media aritmética, en el caso de **distribuciones agrupadas** en intervalos, para el cálculo de la varianza se utiliza la marca de clase como representante de cada intervalo.

Como ejemplo, utilizamos de nuevo los datos relativos a los establecimientos hoteleros ubicados en un determinado municipio, clasificados por el número de

habitaciones, que se han empleado ya en el epígrafe 2.2.1, para el cálculo de la media aritmética con este tipo de distribuciones.

Tabla 2.20. Cálculos intermedios para la obtención de la varianza

$L_{i-1} - L_i$	x_i	n_i	$x_i^2 n_i$
0 – 15	7,50	3	168,75
15 – 30	22,50	1	506,25
30 – 50	40,00	3	4.800,00
50 – 60	55,00	1	3.025,00
60 – 115	87,50	1	7.656,25
Total		9	16.156,25

Sabemos que el valor medio de la variable es de 34,17 habitaciones. Entonces:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{16.156,25}{9} - 34,17^2 = 627,55 \Rightarrow s = \sqrt{627,55} = 25,05$$

Y la desviación típica resulta igual a 25,05 habitaciones.

Para la **obtención de la varianza** y la desviación típica **con EXCEL** cuando las frecuencias son unitarias, una vez introducidos los datos, se sitúa el cursor en una celda vacía y se ejecuta la instrucción Insertar–Función–Estadísticas–Varianza (o Desviación Típica). Se abre un cuadro de diálogo en el que se indica el rango en el que están los valores de la variable y al Aceptar, se obtiene el resultado.

Cuando las frecuencias no son unitarias, las instrucciones que deben seguirse figuran en la Tabla 2.21. Al ejecutarlas, el resultado obtenido será similar al mostrado en la Tabla 2.22.

En estas tablas, x_i representa los distintos valores de la variable para distribuciones sin agrupar y las correspondientes marcas de clase para las distribuciones agrupadas.

El ejemplo numérico elegido para ilustrar el procedimiento corresponde a los datos relativos a la edad de los participantes en una actividad recreativa.

Tabla 2.21. Procedimiento para la obtención de la varianza con EXCEL

	A	B	C	D
1	x_i	n_i	$x^2_i n_i$	
2	18	2	=A2^2*B2	
3	20	14	(1)	
4	…	…	…	
8	50	3	(1)	
9	Total	Σ (B2:B8) (2)	Σ (C2:C8) (2)	
10	Varianza	$=(C9/B9) - 26{,}08^2$		
11	Desviación típica	=Raíz(B10) (3)		

(1) Previamente debe copiarse la celda C2
(2) Insertar – Función – Matemáticas – Suma, para el correspondiente rango de valores
(3) Insertar – Función – Matemáticas – Raíz, para el correspondiente rango de valores

Tabla 2.22. Obtención de la varianza con EXCEL

	A	B	C	D
1	x_i	n_i	$x^2_i n_i$	
2	18	2	648	
3	20	14	5.600	
4	24	16	9.216	
5	26	11	7.436	
6	38	2	2.888	
7	46	2	4.232	
8	50	3	7.500	
9	Total	50	37.520	
10	Varianza	70,23		
11	Desviación típica	8,38		

Algunas de las **propiedades** más interesantes **de la varianza y la desviación típica** son las siguientes**:**

1. Como se deduce de su definición, la varianza y la desviación típica son siempre positivas, puesto que la varianza es el cociente de una suma de cuadrados (es decir, es una suma en la que todos los sumandos son positivos) y el número de datos ($N > 0$) y la desviación típica es su raíz cuadrada positiva.

2. A la varianza y a la desviación típica no les afectan los cambios de origen; es decir, si a todos los valores de una variable se les suma una constante k, la varianza y la desviación típica no varían.

Para demostrarlo, partimos de una distribución de frecuencias (x_i, n_i), con media aritmética $\bar{x}$ y varianza s^2. Si se efectúa un cambio de origen, la media aritmética de la nueva distribución (x_i^o, n_i) es $\bar{x}^o = \bar{x} + k$.

En cuanto a la varianza:

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

Si a la varianza no le afecta el cambio de origen, tampoco le afecta a la desviación típica:

$$s^o = \sqrt{s^{2o}} = \sqrt{s^2} = s$$

3. A la varianza y a la desviación típica les afectan los cambios de escala, puesto que al multiplicar todos los valores de una variable por una constante k, la desviación típica queda multiplicada por dicha constante, y la varianza por su cuadrado.

Para demostrarlo, partimos de una distribución de frecuencias (x_i, n_i), con media aritmética $\bar{x}$ y varianza s^2. Si se efectúa un cambio de escala, la media aritmética de la nueva distribución (x_i^e, n_i) es $\bar{x}^e = k\ \bar{x}$.

En cuanto a la varianza:

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

Por consiguiente, la desviación típica de la nueva distribución quedará multiplicada por la constante k, ya que:

$$s^e = \sqrt{s^{2e}} = \sqrt{k^2 s^2} = ks$$

Un problema que se plantea con frecuencia es el de **comparar la dispersión** de dos o más distribuciones. Dado que las desviaciones típicas miden la dispersión en torno a la media y vienen dadas en las mismas unidades de medida que la correspondiente variable, sólo son directamente comparables cuando son iguales las medias aritméticas y las dos variables que se pretende comparar están dadas en las mismas unidades.

En otro caso, es incorrecto efectuar la comparación a través de las desviaciones típicas, y hemos de recurrir a alguna de las medidas de dispersión relativas, que se concretan en forma de cociente. Entre ellas están el coeficiente de apertura, el recorrido relativo y otras, pero una de las más utilizadas es el coeficiente de variación de Pearson.

2.3.2. Medidas de dispersión relativas

El **coeficiente de variación**, CV, se define por cociente entre la desviación típica y la media aritmética de la variable e indica, por tanto, el número de veces que la desviación típica contiene a la media. Dado que la desviación típica y la media están expresadas en las mismas unidades, al efectuar el cociente, el valor que se obtiene es adimensional.

El valor mínimo del coeficiente de variación es cero, que es el valor que toma cuando es igual a cero el numerador de la fracción; es decir, la desviación típica. En tal caso, todos los valores de la variable son iguales a la media, de manera que la dispersión de los valores en torno a la media es nula y la media es una representación perfecta de la serie de datos. La media es tanto más representativa cuanto más próximo a cero está el coeficiente de variación, y cuanto más elevado es el coeficiente de variación, menos representativa es la media. El inconveniente de esta medida es que si es igual a cero el denominador de la fracción; es decir, la media de la variable, carece de significado.

Veamos un ejemplo de su aplicación. Supongamos que el salario medio mensual de los trabajadores del Hotel A es de 589 euros al mes y su desviación típica de 74,5 euros; y que el salario medio mensual de los trabajadores del Hotel B es de 691 euros

al mes y su desviación típica de 492 euros. ¿Cuál de los dos salarios medios es más representativo de su distribución?

Se observa, en primer lugar, que aunque las dos variables que se desea comparar están medidas en las mismas unidades, no tienen medias iguales, por lo que no sería correcto comparar directamente los valores de las desviaciones típicas, que son medidas de dispersión respecto a la media aritmética. Calculamos, por tanto, los correspondientes coeficientes de variación.

$$CV_A = \frac{s_A}{\bar{x}_A} = \frac{74,5}{589} = 0,1265 \text{ (12,65 por ciento)}$$

$$CV_B = \frac{s_B}{\bar{x}_B} = \frac{492}{691} = 0,7120 \text{ (71,20 por ciento)}$$

El salario medio mensual es más representativo en el caso del Hotel A, ya que la dispersión es mucho menor. De hecho, en el caso del Hotel B, la media del salario no es representativa de la distribución, puesto que la dispersión es muy elevada: la desviación típica es más del 70 por ciento del valor medio de la variable.

2.4. Medidas de forma

Para analizar la forma de las distribuciones, en particular, vamos a referirnos a dos de sus características principales: la asimetría y el apuntamiento.

Con respecto a la **asimetría**, se dice que una distribución de frecuencias es simétrica cuando los valores de la variable equidistantes de un valor central (por ejemplo, la media aritmética) tienen la misma frecuencia absoluta. En tal caso, su representación gráfica es tal que la mitad de la misma es igual a la otra mitad. Si se doblase por la mitad el histograma de frecuencias de la variable, las dos mitades serían semejantes. En otro caso, la distribución es asimétrica.

El histograma representado en el Gráfico 2.1 se ha obtenido con los valores observados de una variable en una determinada muestra, que no es más que una representación aproximada de la población a la que pertenece, tanto mejor, cuanto más grande es su tamaño. Cuando la muestra es pequeña, el histograma tiene un

aspecto quebrado. Pero, si la variable es continua, el tamaño de la muestra puede aumentar indefinidamente, aumentando así el número de valores observados de la variable.

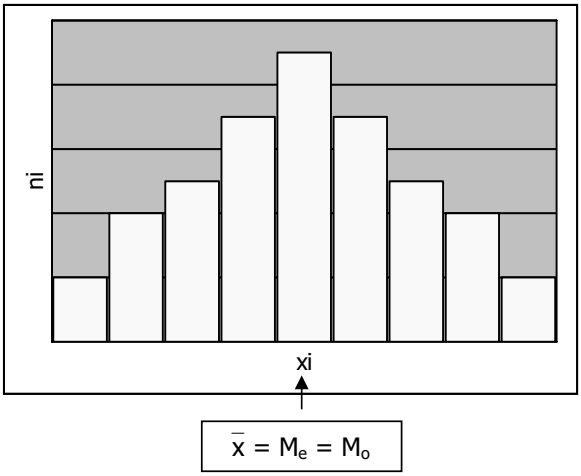


Gráfico 2.1. Histograma de frecuencias

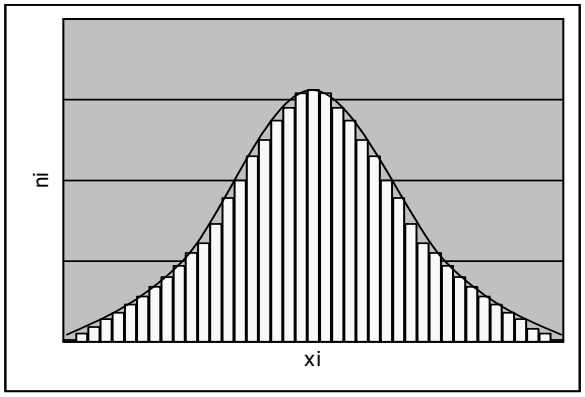


Gráfico 2.2. Línea suavizada del histograma de frecuencias

En términos gráficos, en el histograma aumenta el número de barras, que se van haciendo cada vez más estrechas. El polígono de frecuencias va suavizando su forma, hasta que en el límite, cuando el número de intervalos tiende a infinito y, por tanto, su amplitud tiende a cero, presenta el aspecto ideal de una curva completamente suavizada, como la que se muestra en el Gráfico 2.2.

Aunque existen diversos métodos estadísticos que permiten la obtención de esta línea ideal, de momento, basta con considerar que podría obtenerse suavizando la línea escalonada del histograma.

Como puede observarse, la curva obtenida en este caso es simétrica, de tal manera que la media aritmética, la mediana y la moda coinciden. Tiene una cota máxima y decrece hacia los extremos, sin cortar nunca al eje de abscisas. Tiene, además, dos puntos de inflexión o de cambio de curvatura, que equidistan de los valores centrales. Finalmente, el perfil de la curva es similar al de una campana, por lo que se dice que la distribución es **campaniforme**.

Se ha observado que los datos numéricos de una gran cantidad de fenómenos, tanto naturales como sociales, se rigen por una distribución de frecuencias que presenta estas características, por lo que habitualmente se denomina **distribución normal**.

La distribución normal es, como hemos visto, simétrica. Pero existen otras distribuciones, tanto campaniformes como no campaniformes, que presentan asimetrías a la derecha o positivas o a la izquierda o negativas. Se dice que existe asimetría a la derecha o positiva si la curva tiene cola a la derecha; es decir, si las frecuencias descienden más lentamente por la derecha que por la izquierda. Por el contrario, si la curva presenta cola a la izquierda; es decir, si las frecuencias descienden más lentamente por la izquierda que por la derecha, se dice que la distribución presenta asimetría a la izquierda o negativa.

Para medir la asimetría de una distribución pueden emplearse diferentes coeficientes. Dos de los más utilizados, son los de Pearson y Fisher.

El **coeficiente de asimetría de Pearson** se define como:

$$A_P = \frac{\bar{x} - M_o}{s}$$

En una distribución campaniforme y simétrica, su valor es igual a cero, puesto que la media y la moda son iguales.

Si el coeficiente es mayor que cero, la media está a la derecha de la moda, y la distribución es asimétrica a la derecha o positiva. Si el coeficiente es menor que cero, la media está a la izquierda de la moda, y la distribución es asimétrica a la izquierda o negativa.

A modo de ejemplo, se analiza la asimetría de la distribución de los precios (en euros) por habitación en los hoteles ubicados en una determinada zona turística, recogida en la Tabla 2.23.

Tabla 2.23. Precios por habitación

Precio por habitación ($L_{i-1} - L_i$)	Nº de hoteles (n_i)
0 – 45	8
45 – 60	10
60 – 90	4
90 – 120	3
120 – 210	2
210 – 571	1

Para obtener el coeficiente de asimetría de Pearson es necesario calcular previamente la media, la moda y la desviación típica de la distribución, teniendo en cuenta que está agrupada en intervalos de amplitud variable. Los cálculos intermedios se presentan en la Tabla 2.24.

Tabla 2.24. Obtención del coeficiente de asimetría de Pearson

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	c_i	d_i
0 – 45	22,5	8	180,0	4.050,00	45	0,1777
45 – 60	52,5	10	525,0	27.562,50	15	0,6666
60 – 90	75,0	4	300,0	22.500,00	30	0,1333
90 – 120	105,0	3	315,0	33.075,00	30	0,1000
120 – 210	165,0	2	330,0	54.450,00	90	0,0222
210 – 571	390,5	1	390,5	152.490,25	361	0,0027
Total		28	2.040,5	294.127,75		

Con la información que proporciona dicha tabla, se obtienen los siguientes resultados:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{2.040,5}{28} = 72,875$$

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 45 + \frac{0,1333}{0,1777 + 0,1333} \times 15 = 51,428$$

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{294.127,75}{28} - 72,87^2 = 5.193,796$$

$$s = \sqrt{5.193,796} = 72,068$$

$$A_P = \frac{\bar{x} - M_o}{s} = \frac{72,875 - 51,428}{72,068} = 0,297$$

Por tanto, la distribución es asimétrica a la derecha o positiva, tal como puede observarse también en la correspondiente representación gráfica, que muestra cola a la derecha, puesto que las densidades de frecuencia descienden más lentamente por la derecha que por la izquierda.

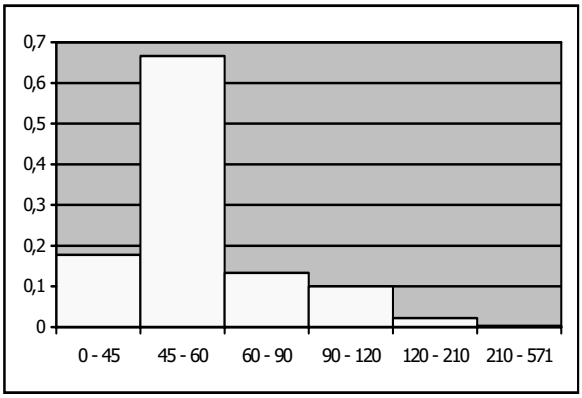


Gráfico 2.3. Histograma de frecuencias

Obsérvese que en este gráfico, las alturas de las barras no coinciden con las frecuencias absolutas, sino que, como la distribución de los precios está agrupada en

intervalos de amplitud variable, dichas alturas coinciden con las densidades de frecuencia.

Para el **cálculo del coeficiente de asimetría de Pearson con EXCEL**, deben ejecutarse las instrucciones recogidas en la Tabla 2.25. Al efectuarlas, se obtendrá un resultado similar al que muestra la Tabla 2.26.

Tabla 2.25. Procedimiento para el cálculo de A_P con EXCEL

	A	B	C	D	E	F	G
1	$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	c_i	d_i
2	0 – 45	=(0+45)/2	8	=B2*C2	=B2^2*C2	=45-0	=C2/F2
3	45 – 60	=(45+60)/2	10			=60-45	
4	60 – 90	=(60+90)/2	4			=90-60	
…	…	…	…	…	…	…	…
7	210 – 571	=(210+571)/2	1			=571-210	
8	Total		Σ(C2:C7)	Σ(D2:D7)	Σ(E2:E7)		
9	Media	=D8/C8					
10	Varianza	=(E8/C8)-(B9^2)					
11	D. Típica	=Raíz(B10)					
12	Moda	=45+(G4/(G2+G4))*E3					
13	C. Pearson	=(B9-B12)/B11					

Tabla 2.26. Obtención de A_p con EXCEL

	A	B	C	D	E	F	G
1	$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	c_i	d_i
2	0 – 45	22,5	8	180,0	4.050,00	45	0,177
3	45 – 60	52,5	10	525,0	27.562,50	15	0,666
4	60 – 90	75,0	4	300,0	22.500,00	30	0,133
5	90 – 120	105,0	3	315,0	33.075,00	30	0,100
6	120 – 210	165,0	2	330,0	54.450,00	90	0,022
7	210 – 571	390,5	1	390,5	152.490,25	361	0,002
8	Total		28	2.040,5	294.127,75		
9	Media	72,875					
10	Varianza	5.193,796					
11	D. Típica	72,068					
12	Moda	51,428					
13	C. de Pearson	0,297					

El coeficiente de asimetría de Pearson es muy sencillo, pero tiene el inconveniente de que no resulta demasiado fiable para medir asimetrías relativamente

pequeñas. Además, dada su definición, sólo es válido para distribuciones unimodales y campaniformes.

Una alternativa más general es el **coeficiente de asimetría de Fisher**, que se define como:

$$A_F = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3 n_i \ / N}{s^3}$$

En el caso de simetría, este coeficiente es igual a cero. Si la distribución es asimétrica a la derecha, toma un valor positivo y toma un valor negativo si se trata de una distribución asimétrica a la izquierda.

Para obtener el valor del coeficiente con los datos de los precios por habitación, tenemos ya, porque los hemos calculado previamente, los valores de la media y la desviación típica. Sólo es necesario obtener la suma de las desviaciones de cada valor respecto a la media elevadas al cubo por la frecuencia absoluta, recogida en la Tabla 2.27.

Tabla 2.27. Obtención del coeficiente de asimetría de Fisher

$L_{i-1} - L_i$	x_i	n_i	$x_i - \overline{x}$	$(x_i - \overline{x})^3$	$(x_i - \overline{x})^3 n_i$
0 – 45	22,5	8	-50,375	-127.833,64	-1.022.669,17
45 – 60	52,5	10	-20,375	-8.458,49	-84.584,90
60 – 90	75,0	4	2,125	9,59	38,38
90 – 120	105,0	3	32,125	33.153,50	99.460,50
120 – 210	165,0	2	92,125	781.866,31	1.563.732,63
210 – 571	390,5	1	317,625	32.043.801,60	32.043.801,60
Total		28			32.599.779,00

Entonces:

$$A_F = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3 n_i \ / N}{s^3} = \frac{32.599.779 \ / \ 28}{72{,}068^3} = 3{,}110$$

El valor que se ha obtenido para el coeficiente de asimetría de Fisher es mayor que cero, por lo que podemos concluir que se trata de una distribución asimétrica a la derecha o positiva tal como muestra su gráfico y tal como indica también el coeficiente de asimetría de Pearson.

Para el **cálculo del coeficiente de asimetría de Fisher con EXCEL**, deben ejecutarse las instrucciones recogidas en la Tabla 2.28. Al efectuarlas, se obtendrá un resultado similar al que muestra la Tabla 2.29.

Tabla 2.28. Procedimiento para el cálculo de A_F con EXCEL

	A	B	C	D	E	F
1	$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^3 n_i$
2	0 – 45	=(0+45)/2	8	=B2*C2	=B2^2*C2	=(B2-$9)^3*C2
3	45 – 60	=(45+60)/2	10			
4	60 – 90	=(60+90)/2	4			
…	…	…	…	…	…	…
7	210 – 571	=(210+571)/2	1			
8	Total		Σ(C2:C7)	Σ(D2:D7)	Σ(E2:E7)	Σ(F2:F7)
9	Media	= D8/C8				
10	Varianza	= (E8/C8)-(B9^2)				
11	D. Típica	= Raíz(B10)				
12	C. Fisher	=(F8/C8)/B11^3				

Tabla 2.29. Obtención de A_F con EXCEL

	A	B	C	D	E	F
1	$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^3 n_i$
2	0 – 45	22,5	8	180,0	4.050,00	-1.022.669,17
3	45 – 60	52,5	10	525,0	27.562,50	-84.584,90
4	60 – 90	75,0	4	300,0	22.500,00	38,38
5	90 – 120	105,0	3	315,0	33.075,00	99.460,50
6	120 – 210	165,0	2	330,0	54.450,00	1.563.732,63
7	210 – 571	390,5	1	390,5	152.490,25	32.043.801,60
8	Total		28	2.040,5	294.127,75	32.599.779,00
9	Media	72,875				
10	Varianza	5.193,796				
11	D. Típica	72,068				
12	C. Fisher	3,110				

La otra característica que vamos a estudiar en relación con la forma de una distribución es su **apuntamiento** o **curtosis**. Las medidas de curtosis se aplican a distribuciones campaniformes; es decir, unimodales y simétricas, o con una ligera asimetría.

Tomando como referencia la curva normal, que se ha descrito en este epígrafe, la curtosis es el mayor o menor apuntamiento con respecto a ella. Se dice que una curva es muy apuntada, si es más alta y estrecha que la normal y si es más plana y ancha que la normal, se dice que es poco apuntada.

La curva normal de referencia es mesocúrtica, mientras que se llaman leptocúrticas las curvas más apuntadas que la normal y platicúrticas las menos apuntadas que la normal.

Como criterio de apuntamiento se elige el de las distancias de cada valor de la variable respecto a su media elevadas a la cuarta potencia, de tal manera que, para medir el grado de apuntamiento de una distribución, se utiliza el denominado **coeficiente de exceso**, al que se denota como g.

Dicho coeficiente se define como:

$$g = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4 n_i / N}{s^4} - 3$$

El primer sumando de esta expresión, es el grado de apuntamiento correspondiente a la distribución de frecuencias analizada. El segundo, es el grado de apuntamiento para la distribución normal. Luego el coeficiente de exceso es la diferencia entre el apuntamiento de la distribución correspondiente y el de una normal que tenga su misma media y varianza.

Si su valor es igual a cero, la distribución es igual de apuntada que una normal de igual media y varianza; es decir, que es mesocúrtica. Si el coeficiente de exceso es mayor que cero, la distribución es más apuntada que una normal de igual media y varianza; es decir, es leptocúrtica. Si el coeficiente de exceso es menor que cero, la

distribución es menos apuntada que una normal de igual media y varianza; es decir, es platicúrtica.

Partiendo de los datos del ejemplo anterior, que se refieren a los precios (en euros) por habitación de los hoteles ubicados en una determinada zona turística, se procede a la obtención del coeficiente de exceso, para lo cual se elabora la Tabla 2.30, que contiene los correspondientes cálculos intermedios.

Tabla 2.30. Obtención del coeficiente de exceso

$L_{i-1} - L_i$	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4$	$(x_i - \bar{x})^4 n_i$
0 – 45	22,5	8	-50,375	6.439.619,94	51.516.959,53
45 – 60	52,5	10	-20,375	172.341,73	1.723.417,38
60 – 90	75,0	4	2,125	20,39	81,56
90 – 120	105,0	3	32,125	1.065.056,25	3.195.168,75
120 – 210	165,0	2	92,125	72.029.434,2	144.058.868,4
210 – 571	390,5	1	317,625	1,0178E+10	10.177.912.484
Total		28			10.378.406.980

Entonces:

$$g = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 n_i / N}{s^4} - 3 = \frac{10.378.406.980 / 28}{72{,}068^4} - 3 = 13{,}740 - 3 = 10{,}740$$

Por tanto, la distribución es leptocúrtica; es decir, es más apuntada que una normal de igual media y varianza.

En cuanto a la **obtención del coeficiente de exceso con EXCEL**, se describe el proceso en la Tabla 2.31, y se muestra el resultado en la Tabla 2.32.

Tabla 2.31. Procedimiento para la obtención de g con EXCEL

	A	B	C	D	E	F
1	$L_{i-1} - L_i$	x_i	n_i	$X_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^4 n_i$
2	0 – 45	=(0+45)/2	8	=B2*C2	=B2^2*C2	=(B2-B$9)^4*C2
3	45 – 60	=(45+60)/2	10			
4	60 – 90	=(60+90)/2	4			
…	…	…	…	…	…	…
7	210 – 571	=(210+571)/2	1			
8	Total		Σ(C2:C7)	Σ(D2:D7)	Σ(E2:E7)	Σ(F2:F7)
9	Media	=D8/C8				
10	Varianza	=(E8/C8)-(B9^2)				
11	D.Típica	=Raíz(B10)				
12	Moda	=G3+(G4/(G2+G4))*F3				
13	C. Apunt.	=((F8/C8)/B11^4)-3				

Tabla 2.32. Obtención de g con EXCEL

	A	B	C	D	E	F
1	$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	$(x_i - \bar{x})^4 n_i$
2	0 – 45	22,5	8	180,0	4.050,00	51.516.959,53
3	45 – 60	52,5	10	525,0	27.562,50	1.723.417,38
4	60 – 90	75,0	4	300,0	22.500,00	81,56
5	…	105,0	3	315,0	33.075,00	3.195.168,75
6	210 – 571	165,0	2	330,0	54.450,00	144.058.868,4
7		390,5	1	390,5	152.490,25	10.177.912.484
8	Total		28	2.040,5	294.127,75	10.378.406.980
9	Media	72,875				
10	Varianza	5.193,796				
11	D. Típica	72,068				
12	C. Apunt.	10,740				

2.5. Medidas de concentración

Finalmente, se analizan algunas de las medidas de concentración, que permiten conocer el grado de equidistribución de la variable; es decir, que indican si los valores de la variable están distribuidos o no de forma equitativa. Las medidas de concentración hacen referencia, por tanto, al mayor o menor grado de igualdad en el reparto del total de los valores de la variable.

Por ejemplo, si se analiza la distribución de las pernoctaciones registradas en establecimientos hoteleros a lo largo de los 12 meses del año, como extremos, pueden presentarse dos casos:

1. Concentración mínima o equidistribución: si todos los meses del año se registra el mismo número de pernoctaciones.

2. Concentración máxima: si el total de pernoctaciones se registra en un solo mes del año.

Para analizar el grado de igualdad en el reparto de los valores de una variable, generalmente se utilizan la curva de Lorenz y el índice de Gini.

La curva de concentración o **curva de Lorenz** permite efectuar el análisis de la concentración desde el punto de vista gráfico.

Veamos un ejemplo de su aplicación, con los datos correspondientes a los salarios de los trabajadores (en miles de euros) y al número de trabajadores que perciben cada salario en una determinada empresa hostelera, información que se presenta en la Tabla 2.33.

Tabla 2.33. Salarios de los trabajadores

Salarios (x_i)	Nº Trabajadores (n_i)
6	3
12	7
18	8
24	4
30	2
36	1

Si la distribución de los salarios fuese totalmente igualitaria, el 10 por ciento de los trabajadores debería percibir el 10 por ciento del total de los salarios, al 25 por ciento de los trabajadores debería corresponderle el 25 por ciento del total de salarios, la mitad de los trabajadores debería recibir la mitad de los salarios, etcétera. En tal caso, el reparto de los salarios sería equitativo.

Por el contrario, si la distribución de los salarios está muy concentrada, un porcentaje pequeño de trabajadores percibe un porcentaje muy elevado de la masa salarial.

Por tanto, para analizar el grado de concentración; es decir, el grado de desigualdad en la retribución de los diferentes trabajadores, deben calcularse, en primer lugar, los porcentajes acumulados de valores de la variable y de las frecuencias relativas, tal como se muestra en la Tabla 2.34.

La columna de valores de p_i recoge los porcentajes de frecuencia relativa acumulada, que indican el peso que tiene cada frecuencia absoluta acumulada sobre el total de los datos, en porcentaje. La columna de valores de q_i indica el valor acumulado de la variable, obtenido como suma, para cada valor, del producto $x_i n_i$ más los anteriores, expresado en porcentaje.

Tabla 2.34. Porcentajes acumulados

x_i	n_i	N_i	F_i	$p_i = 100F_i$	$x_i n_i$	$U_i = Ac\ (x_i n_i)$	$q_i = 100(U_i / U_n)$
6	3	3	0,12	12	18	18	4,1
12	7	10	0,40	40	84	102	23,3
18	8	18	0,72	72	144	246	56,2
24	4	22	0,88	88	96	342	78,1
30	2	24	0,95	95	60	402	91,8
36	1	25	1,00	100	36	438	100,0

En este caso, a una frecuencia relativa acumulada del 12 por ciento, le corresponde un valor acumulado de la variable del 4,1 por ciento, lo cual significa que el 12 por ciento de los valores observados del salario corresponden al primer valor que toma la variable, que representa un 4,1 por ciento sobre su valor total; es decir, que el 12 por ciento de los trabajadores recibe un 4,1 por ciento de los salarios. Los demás resultados de la tabla, se interpretan en los mismos términos: el 40 por ciento de los trabajadores percibe el 23,3 por ciento de los salarios, el 72 por ciento de los trabajadores percibe el 56,2 por ciento de los salarios, etcétera.

Una vez efectuados estos cálculos, para representar gráficamente la concentración, se utiliza un sistema de coordenadas cartesianas, representando en el eje de abscisas los valores que toma p_i y en el de ordenadas los correspondientes a q_i.

Al estar p_i y q_i expresados en porcentaje, la escala y el campo de variación de cada uno de los ejes coinciden, de tal manera que la curva de concentración encaja perfectamente en un cuadrado.

En el cuadrado se traza la diagonal que une el vértice inferior izquierdo con el superior derecho. Esta diagonal, obviamente, pasa por el origen y por todos los puntos que tienen coordenadas iguales y, por tanto, representa la mínima concentración, ya que en todos sus puntos se cumple que $p_i = q_i$. Por el contrario, cuanto más elevada es la concentración, más elevado es el valor de q_i que le corresponde a un valor pequeño de p_i, de tal manera que la curva tiende a confundirse con los lados inferior e izquierdo del cuadrado. Luego, la curva de Lorenz indica una concentración débil si está próxima a la diagonal del cuadrado y una concentración mayor a medida que se aleja de la diagonal.

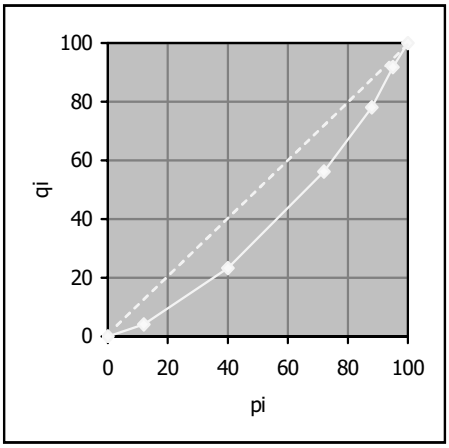


Gráfico 2.4. Curva de Lorenz de la distribución de los salarios

La curva de Lorenz de la distribución de los salarios no está excesivamente alejada de la diagonal del cuadrado, luego no se observa en dicha distribución una elevada concentración.

El **índice de Gini** resume en un solo valor numérico la información gráfica proporcionada por la curva de Lorenz, puesto que es el doble del valor del área

comprendida entre dicha curva y la diagonal del cuadrado, bajo el supuesto convencional de que el área del cuadrado es igual a la unidad.

El valor del índice está, por tanto, comprendido entre 0 y 1, puesto que si la curva coincide con la diagonal el área vale cero, y si coincide con los lados del cuadrado es igual a ½. Dado que la concentración es mínima cuando la curva se superpone a la diagonal y máxima cuando la curva se superpone a los lados del cuadrado, el índice de Gini indica un grado mayor de igualdad en el reparto cuanto más próximo a cero está su valor y un grado mayor de concentración cuanto más próximo a la unidad está su valor.

También se basa en los valores que toman, en porcentaje, la frecuencia relativa acumulada y el valor acumulado de la variable, p_i y q_i.

Se define de la siguiente forma:

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

Cuando la concentración es máxima, $q_i = 0$, por tanto, el índice de Gini es igual a la unidad. Cuando la concentración es mínima, $p_i = q_i$, por tanto, el índice de Gini es igual a cero.

En el caso de la distribución de los salarios, para obtener el índice de Gini se efectúan los cálculos intermedios recogidos en la Tabla 2.35.

Tabla 2.35. Obtención del índice de Gini

$p_i = 100F_i$	$q_i = 100(U_i / U_n)$	$p_i - q_i$
12	4,1	7,9
40	23,3	16,7
72	56,2	15,8
88	78,1	9,9
95	91,8	3,2
100	100	

Y se tiene:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{53,5}{307} = 0,174$$

El índice no está alejado de cero, luego la distribución está escasamente concentrada.

Veamos otro ejemplo del análisis de la concentración de la distribución. A un grupo de turistas que visitó Galicia durante quince días, al final de su estancia, se le preguntó acerca del gasto (en euros) efectuado en dicho período. Los resultados de la encuesta se presentan en la Tabla 2.36.

Tabla 2.36. Gasto turístico

Gasto ($L_{i-1} - L_i$)	Nº de turistas (n_i)
300 – 450	5
450 – 751	7
751 – 902	20
902 – 1.352	12
1.352 – 1.803	4
1.803 – 3.005	2

Tabla 2.37 Cálculos intermedios para representar la curva de Lorenz

x_i	n_i	N_i	F_i	$p_i = 100F_i$	x_in_i	$U_i = Ac\ (x_in_i)$	$q_i = 100(U_i / U_n)$
375	5	5	0,1	10	1.875	1.875	3,97
600,5	7	12	0,24	24	4.203,5	6.078,5	12,86
826,5	20	32	0,64	64	16.530	22.608,5	47,85
1.127	12	44	0,88	88	13.524	36.132,5	76,47
1.577,5	4	48	0,96	96	6.310	42.442,5	89,82
2.404	2	50	1	100	4.808	47.250,5	100,00

De la comparación de las columnas de valores de p_i y q_i se deduce que el 10 por ciento de los turistas gastan aproximadamente el 4 por ciento del total del gasto de los

50 visitantes, el 24 por ciento gasta aproximadamente el 13 por ciento del total, …, y el 96 por ciento gasta aproximadamente el 90 por ciento del total. Con estos valores de p_i y q_i se obtiene la curva de concentración que se muestra en el Gráfico 2.5.

Tabla 2.38. Cálculos intermedios para la obtención del índice de Gini

$p_i = 100F_i$	$q_i = 100\ (U_i / U_n)$	$p_i - q_i$
10	3,97	6,03
24	12,86	11,14
64	47,85	16,15
88	76,47	11,53
96	89,82	6.18
100	100,00	

Para cuantificar la concentración a través del índice de Gini, es necesario obtener la columna de valores de las diferencias $p_i - q_i$, cuyos resultados figuran en la Tabla 2.38.

Con la información que contiene se deduce que:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{51,03}{282} = 0,180$$

El índice de Gini está próximo a cero, lo cual indica que la distribución del gasto no está excesivamente concentrada.

Este resultado es coherente con el obtenido para la curva de Lorenz que, como puede verse en el Gráfico 2.5, no está excesivamente alejada de la diagonal del cuadrado, de tal manera que es indicativa, igualmente, de que la concentración del gasto es reducida.

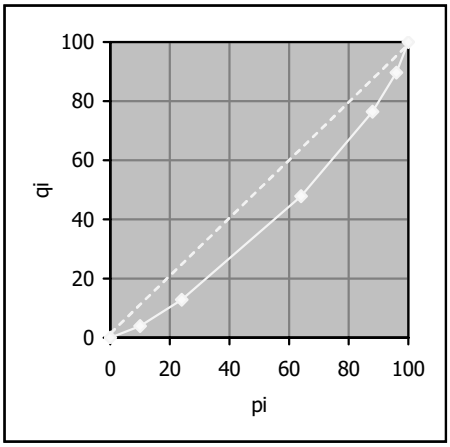


Gráfico 2.5. Curva de Lorenz de la distribución del gasto

El procedimiento para la **obtención del índice de Gini con EXCEL**, se describe en la Tabla 2.39. Al ejecutar las instrucciones indicadas, el resultado obtenido será similar al mostrado en la Tabla 2.40.

El ejemplo numérico elegido para efectuar la aplicación es el correspondiente a la distribución del gasto de los turistas.

Tabla 2.39. Procedimiento para la obtención del índice de Gini con EXCEL

	A	B	C	D	E	F	G	H	I
1	x_i	n_i	N_i	F_i	$p_i = 100F_i$	x_in_i	$U_i = Ac(x_in_i)$	$q_i = 100(U_i/U_n)$	$p_i - q_i$
2	375	5	=B2	=C2/C$7	=D2*100	=A2*B2	=F2	=(G2/G$7)*100	=F2-H2
3	600,5	7	=C2+B3				=G2+F3		
4	826,5	20							
...	...	...	...	...	...	...	...	...	...
7	2.404	2							
8					Σ (E2:E6)			Σ (H2:H6)	Σ (I2:I6)
9	IG		=I8/E8						

Tabla 2.40. Obtención del índice de Gini con EXCEL

	A	B	C	D	E	F	G	H	I
1	x_i	n_i	N_i	F_i	p_i	$x_i n_i$	U_i	q_i	$p_i - q_i$
2	375	5	5	0,1	10	1.875	1.875	3,97	6,03
3	600,5	7	12	0,24	24	4.203,5	6.078,5	12,86	11,14
4	826,5	20	32	0,64	64	16.530	22.608,5	47,85	16,15
5	1.127	12	44	0,88	88	13.524	36.132,5	76,47	11,53
6	1.577,5	4	48	0,96	96	6.310	42.442,5	89,82	6,18
7	2.404	2	50	1	100	4.808	47.250,5	100,00	0,00
8					282	47.250,5			51,03
9	IG	0,180							

TEMA 3
NÚMEROS ÍNDICES

3.1. Introducción

En los capítulos anteriores se han presentado las medidas más utilizadas para resumir la información que respecto a una variable contiene su tabla de distribución de frecuencias, principalmente por lo que se refiere a los valores centrales, la dispersión, la concentración y la forma. Nos ocupamos ahora de analizar las técnicas estadísticas diseñadas para analizar sus variaciones.

Entre ellas están los números índices, que permiten establecer comparaciones entre el valor que una variable toma en una observación concreta y el correspondiente a otra que se toma como referencia. La técnica es adecuada para efectuar dichas comparaciones tanto en el tiempo como en el espacio, aunque lo más frecuente es que se aplique para el análisis de la evolución temporal; es decir, para cuantificar los cambios a lo largo del tiempo.

3.2. Números índices simples y complejos

Un **número índice** es, por tanto, una medida que permite expresar en términos porcentuales cuánto ha aumentado o disminuido una variable en relación a un determinado **dato base** o **de referencia** que, generalmente, corresponde a un período de tiempo.

Para fijar el período base ha de tenerse en cuenta que los índices se expresan como porcentaje de variación de la variable respecto a él, de manera que no debe corresponderle un dato anómalo o atípico que se deba a algún acontecimiento extra-ordinario. **El valor de la variable en el período base ha de ser representativo de la serie**. Es frecuente, por ejemplo, elegir como referencia el período al que corresponde la moda o (aproximadamente) el valor medio.

Habitualmente se utiliza la letra x para designar a la variable objeto de estudio, su valor en el período base se representa por x_0 y el correspondiente al momento a comparar, denominado valor actual o corriente, se denota por x_t.

Para calcular el valor del índice del momento t–ésimo respecto al período base (I_t^0) se divide el valor corriente (x_t) por el que la variable toma en el período de referencia (x_0) y se multiplica el cociente por cien; es decir, el índice es el porcentaje que supone el valor corriente respecto al del período de referencia:

$$I_t^0 = \frac{x_t}{x_0} \times 100$$

Utilizando el mismo procedimiento para todos los valores de la variable se obtiene la serie de números índices, que facilita la descripción de su evolución y, al ser adimensional, permite efectuar comparaciones entre distintas variables aunque estén expresadas en diferentes unidades de medida.

Dada su definición, el valor del índice en el período base es igual a 100; en los períodos en los que es superior, la variable ha experimentado desde el período base un crecimiento en un porcentaje igual a la cantidad que excede de 100 y en los períodos en los que es inferior la variable ha experimentado desde el período base una disminución en un porcentaje igual a la cantidad que falta hasta 100.

A modo ilustrativo, en la siguiente tabla se presenta la serie de índices correspondientes a la variable x = número de viajeros, que refleja los cambios porcentuales que dicha variable ha experimentado en relación al año 1995, que ha sido el período tomado como referencia.

Tabla 3.1. Números índices de la variable x = número de viajeros, año base 1995

Años	N^{o} de viajeros	I_t^{95}
1995	273	100,00
1996	230	84,25
1997	291	106,59
1998	300	109,89
1999	323	118,32
2000	327	119,78
2001	320	117,22
2002	316	115,75

El valor del índice en el año 1996 se obtiene multiplicando por cien el cociente entre el valor de la variable en 1996 y su valor en el período base, 1995 (= (230 / 273) x 100 = 84,25). De forma similar, el valor del índice en el año 1997 se obtiene multiplicando por cien el cociente entre el valor de la variable en 1997 y su valor en el período base, 1995 (= (291 / 273) x 100 = 106,59), etcétera. Una vez completada la serie, puede observarse que con respecto a 1995, en 1996 se registra una disminución del número de viajeros del 15,75 (= 100 – 84,25) por ciento, en el año 1997 se registra un aumento del 6,59 (= 106,59 – 100) por ciento, etcétera.

Los números **índices** que se han descrito hasta ahora son **simples**, en el sentido de que sintetizan la evolución de una sola variable, y suelen utilizarse con precios, cantidades y valores relativos.

Cuando el objetivo es analizar la evolución conjunta de n variables se utilizan los números **índices complejos**, que se obtienen construyendo, en primer lugar, la serie de índices simples para cada una de las n variables, y promediándolos a continuación para unificar la información. Cuando la importancia que se le otorga a cada índice simple en el complejo es la misma (**índices complejos no ponderados**), los simples se promedian utilizando una media aritmética simple y cuando el peso que se le otorga a cada índice simple en el complejo es diferente (**índices complejos ponderados**), los simples se promedian utilizando una media aritmética ponderada.

Los **índices complejos no ponderados** pueden obtenerse también sumando los valores correspondientes a cada variable para cada período considerado y calculando a continuación los índices simples de los valores agregados. Este procedimiento se conoce como **método de la media agregativa simple** y para utilizarlo es necesario que todas las variables implicadas estén expresadas en las mismas unidades de medida, puesto que en caso contrario, la suma no está definida. En general, cuando los datos lo permiten, es preferible utilizar este método, ya que la serie obtenida refleja mejor la evolución real del conjunto de las n variables consideradas.

Veamos un ejemplo. Se dispone de los datos proporcionados por el Banco de Datos TEMPUS del Instituto Nacional de Estadística (http://www.ine.es) respecto a la variable x = número de turistas (miles de personas) que visitaron España en el período

comprendido entre los años 1995 y 2003 por carretera, x_1, en avión, x_2, en tren, x_3 o en barco, x_4.

Tabla 3.2. Número de turistas según medio de transporte

Años	x_1	x_2	x_3	x_4
1995	9.036,961	23.934,026	241,111	1.707,477
1996	9.369,452	24.657,933	379,070	1.814,553
1997	9.454,422	27.677,060	409,178	2.012,060
1998	10.581,058	30.457,074	421,761	1.936,190
1999	11.521,994	32.574,134	428,221	2.251,520
2000	10.669,405	34.379,930	445,405	2.403,175
2001	11.738,611	35.331,092	457,357	2.566,495
2002	13.872,392	34.946,554	458,066	3.049,755
2003	12.118,004	36.922,885	406,021	2.732,994

El objetivo planteado es obtener una serie de números índices que refleje la evolución del número total de turistas, tomando como referencia el año 1995.

Para ello, obtenemos, en primer lugar, los índices simples para cada una de las cuatro variables consideradas.

Tabla 3.3. Números índices simples, año base 1995

Años	x_1	x_2	x_3	x_4
1995	100,00	100,00	100,00	100,00
1996	103,68	103,02	157,22	106,27
1997	104,62	115,64	169,71	117,84
1998	117,09	127,25	174,92	113,39
1999	127,50	136,10	177,60	131,86
2000	118,06	143,64	184,73	140,74
2001	129,90	147,62	189,69	150,31
2002	153,51	146,01	189,98	178,61
2003	134,09	154,27	168,40	160,06

Si se supone que cada uno de estos índices simples tiene la misma importancia en la construcción del complejo, deben promediarse calculando la media aritmética simple para cada año.

Tabla 3.4. Números índices complejos no ponderados, año base 1995

Años	Índice
1995	100,00
1996	117,55
1997	126,95
1998	133,16
1999	143,27
2000	146,80
2001	154,38
2002	167,03
2003	154,21

Si, por el contrario, se supone que en la construcción del índice complejo ha de darse un mayor peso a los índices simples correspondientes a los medios de transporte más utilizados por los turistas, el promedio debe obtenerse utilizando una media aritmética ponderada, siendo las ponderaciones los porcentajes de participación de cada medio de transporte en el total de viajeros transportados, que se recogen en la tabla siguiente:

Tabla 3.5. Porcentaje de participación de cada medio de transporte en el total

Años	Carretera	Avión	Tren	Barco	Total
1995	25,88	68,54	0,69	4,89	100,00
1996	25,87	68,08	1,05	5,01	100,00
1997	23,90	69,98	1,03	5,09	100,00
1998	24,38	70,18	0,97	4,46	100,00
1999	24,63	69,64	0,92	4,81	100,00
2000	22,28	71,78	0,93	5,02	100,00
2001	23,43	70,53	0,91	5,12	100,00
2002	26,51	66,79	0,88	5,83	100,00
2003	23,22	70,76	0,78	5,24	100,00

Los valores de la media aritmética ponderada obtenida para cada año utilizando estas ponderaciones constituyen la serie de números índices complejos ponderados recogida en la tabla de la página siguiente. Comparando las Tablas 3.4 y 3.6, puede observarse que los resultados obtenidos al utilizar los índices complejos no ponderados y ponderados son muy distintos.

Tabla 3.6. Números índices complejos ponderados, año base 1995

Años	Índice
1995	100,00
1996	103,92
1997	113,68
1998	124,62
1999	134,16
2000	138,18
2001	143,99
2002	150,28
2003	150,00

En este caso, al disponer de los datos correspondientes a las cuatro variables consideradas y dado que todas ellas están expresadas en las mismas unidades de medida (miles de personas), los índices complejos no ponderados pueden obtenerse también utilizando el método de la media agregativa simple.

Tabla 3.7. Números índices complejos no ponderados
Método de la media agregativa simple

Años	$x_1 + x_2 + x_3 + x_4$	Índice
1995	34.919,575	100,00
1996	36.221,008	103,73
1997	39.552,720	113,27
1998	43.396,083	124,27
1999	46.775,869	133,95
2000	47.897,915	137,17
2001	50.093,555	143,45
2002	52.326,767	149,85
2003	52.179,904	149,43

Dado el procedimiento de cálculo, el método de la media agregativa simple proporciona una serie de números índices que refleja exactamente la evolución del número total de turistas, mientras que los índices complejos no ponderados y ponderados obtenidos anteriormente son sólo aproximaciones. Si se comparan los resultados obtenidos en los tres casos, puede observarse que la aproximación proporcionada por los índices ponderados es mejor que la obtenida si se utilizan los no

ponderados, de manera que cuando el índice complejo debe calcularse partiendo de los simples, es preferible promediarlos utilizando la media ponderada.

3.3. Cambio de base y enlace de series

Cuando se utilizan números índices es frecuente que una vez transcurrido cierto tiempo desde la elección del período base, el dato correspondiente a dicho período pierda su carácter de normal o representativo, de manera que resulte conveniente fijar un nuevo período de referencia más próximo al actual. Para efectuar un **cambio de base** simplemente se divide cada valor de la serie de números índices entre el que toma en el que se va a considerar como nuevo período de referencia.

Así, por ejemplo, en el epígrafe anterior se han obtenido los índices en base 1995 para los turistas que llegaron a España por carretera en el período comprendido entre 1995 y 2003. Dado el tiempo transcurrido entre el periodo de referencia y el actual, puede ser aconsejable realizar un cambio de base de 1995 a 2000. Para obtener los índices en base 2000 a partir de los que están en base 1995, se divide cada uno de los valores de la serie en base 1995 entre 118,06 que es el valor que en dicha serie corresponde al nuevo período de referencia, de manera que en el año 1995, el índice en base 2000 es 84,70 (= 100,00 / 118,06), en el año 1996 es 87,82 (= 103,68 / 118,06), etcétera.

Tabla 3.8. Números índices, en bases 1995 y 2000

Años	I_t^{95}	I_t^{00}
1995	100,00	84,70
1996	103,68	87,82
1997	104,62	88,62
1998	117,09	99,18
1999	127,50	107,99
2000	118,06	100,00
2001	129,90	110,03
2002	153,51	130,03
2003	134,09	113,58

También es frecuente en la práctica disponer de series de números índices para una variable, que abarcan un período de tiempo más o menos largo, pero que están en

diferentes bases, lo cual imposibilita su comparación. Si se dispone, al menos para un período, de información correspondiente a todas las series, el problema puede resolverse mediante un **enlace técnico de series**, que consiste en efectuar uno o varios cambios de base.

Veamos un ejemplo. Se dispone de dos series de números índices relativas a los precios de un determinado artículo, la primera de ellas en base 1995 y la segunda en base 1998.

Tabla 3.9. Números índices de la serie de precios, en bases 1995 y 1998

Años	Base 1995	Base 1998
1995	100,00	
1996	105,00	
1997	112,00	
1998	120,00	100,00
1999		106,00
2000		111,00
2001		116,00

A fin de disponer de una serie homogénea para el período comprendido entre 1995 y 2001 es necesario enlazar las dos series, pasando la que está en base 1995 a base 1998, que es más actual.

Tabla 3.10. Cambio de base 1995 a 1998

Años	I_t^{95}	I_t^{98}
1995	100,00	83,33
1996	105,00	87,50
1997	112,00	93,33
1998	120,00	100,00

Ahora las dos series de índices están en base 1998; es decir, son homogéneas, por tanto, la serie enlazada se obtiene simplemente pegándolas.

Tabla 3.11. Serie enlazada

Años	Base 1998
1995	83,33
1996	87,50
1997	93,33
1998	100,00
1999	106,00
2000	111,00
2001	116,00

3.4. Índices de precios. Deflación de series

El índice de precios más conocido y utilizado en España es el índice de precios al consumo (IPC) que elabora el Instituto Nacional de Estadística (INE), que tiene por objeto medir la evolución general de los precios.

Se trata de un índice complejo ponderado, que sintetiza la información correspondiente a los índices simples de un amplio conjunto de bienes y servicios que integran lo que se ha denominado **cesta de la compra**, que refleja las pautas de consumo de las familias españolas. En la actualidad, los bienes y servicios a los que se refiere constituyen los siguientes grupos:

1. Alimentos y bebidas no alcohólicas
2. Bebidas alcohólicas y tabaco
3. Vestido y calzado
4. Vivienda
5. Menaje
6. Medicina
7. Transporte
8. Comunicaciones
9. Ocio y cultura
10. Enseñanza
11. Hoteles, cafés y restaurantes
12. Otros bienes y servicios

que se dividen en subgrupos, éstos a su vez en clases y, éstas últimas, en subclases, cada vez más definidas o más concretas. En particular, los precios que se refieren a la actividad turística se recogen en el undécimo grupo, **hoteles, cafés** y **restaurantes**, que refleja la evolución de los precios de los viajes organizados con "todo incluido".

Una vez que se determinan los bienes y servicios que integran la cesta, se valoran las correspondientes cantidades consumidas de cada uno de ellos a precios del período base y del actual y se elaboran las series de números índices simples correspondientes. Dado que no todos los productos tienen el mismo peso en el gasto total de una familia, para obtener el índice complejo, el promedio se obtiene a través de la media aritmética ponderada.

El IPC que se obtiene de esta forma es el más comúnmente utilizado para deflactar variables, proceso que consiste en convertir las series de valores nominales en series de valores reales. El término nominal hace alusión al valor de un bien o servicio expresado en unidades monetarias de cada año (por ejemplo, euros corrientes), mientras que con el término real se hace referencia a su valor expresado en unidades monetarias de un año base (por ejemplo, euros constantes).

Para **deflactar una serie** se calcula el cociente entre sus valores nominales y el índice de precios más apropiado en cada caso que, con frecuencia, puede aproximarse mediante el IPC. Generalmente, este cociente se multiplica por cien, porque también el índice de precios suele estar en porcentaje. Puesto que en términos nominales la variable está valorada en unidades monetarias de cada año y el índice de precios es el cociente entre los precios en unidades monetarias de cada año y los precios en unidades monetarias del año base, al efectuar el cociente se obtiene el valor de la variable en unidades monetarias del año base; es decir, en términos reales o una vez eliminado el efecto de la inflación.

La serie de valores reales permite analizar la evolución real de la variable, en el sentido de que sus valores en términos nominales son el resultado de multiplicar cantidades por precios de cada año, por lo que un aumento (o una disminución) entre dos períodos cualesquiera puede ser debido al aumento (o disminución) que han experimentado las cantidades o que han experimentado los precios, sin que sea posible distinguir cual de los dos factores ha cambiado o si han cambiado ambos simultáneamente, mientras que sus valores en términos reales son el resultado de multiplicar cantidades por precios de un año base, por lo que un aumento (o una disminución) entre dos períodos cualesquiera sólo puede ser debido al aumento (o disminución) que han experimentado las cantidades.

Veamos un ejemplo. En la siguiente tabla se presentan los valores nominales de la variable x = ingresos por turismo expresados en miles de euros corrientes, tomados de la web del INE anteriormente citada, y la serie de índices en base 2001.

Tabla 3.12. Ingresos por turismo en miles de euros corrientes e índices en base 2001

Años	Ingresos	Índices
1990	11.289,334	30,84
1991	11.967,025	32,69
1992	13.613,539	37,19
1993	15.100,001	41,25
1994	17.297,037	47,26
1995	19.038,779	52,02
1996	20.975,158	57,31
1997	23.667,833	64,66
1998	26.806,378	73,24
1999	30.415,594	83,10
2000	33.749,655	92,21
2001	36.602,357	100,00
2002	35.543,440	97,11
2003	36.871,040	100,73

Esta serie indica que, por ejemplo, con respecto al año 2001, en el año 2003 los ingresos por turismo en términos nominales han experimentado un crecimiento del 0,73 por ciento.

Puesto que los índices se han obtenido a partir de los valores nominales de la variable, las variaciones observadas pueden deberse a cambios en los precios o en las cantidades. Para cuantificar los cambios que experimentan las cantidades, que son los que realmente dan una idea de la evolución de la actividad turística, debe eliminarse el efecto de los precios, y calcular los índices una vez deflactada la serie.

Para obtener la serie de valores reales de los ingresos por turismo basta con dividirla por algún deflactor adecuado que, en este caso, podría ser el índice de precios turísticos. Dado que no se dispone de información respecto a él, utilizaremos, como aproximación, el índice de precios al consumo en base 2001. En la tabla siguiente se recogen las series de ingresos nominales, el IPC y los ingresos reales.

Tabla 3.13. Ingresos por turismo en términos nominales y reales
Serie deflactada con el IPC, en base 2001

Años	Ingresos nominales	IPC	Ingresos reales
1990	11.289,334	65,95	17.116,771
1991	11.967,025	69,86	17.129,724
1992	13.613,539	74,01	18.394,564
1993	15.100,001	77,44	19.498,528
1994	17.297,037	81,04	21.343,419
1995	19.038,779	84,83	22.443,647
1996	20.975,158	87,85	23.876,547
1997	23.667,833	89,58	26.421,075
1998	26.806,378	91,22	29.385,651
1999	30.415,594	93,33	32.589,217
2000	33.749,655	96,54	34.961,036
2001	36.602,357	100,00	36.602,357
2002	35.543,440	103,54	34.328,884
2003	36.871,040	106,68	34.560,984

Tabla 3.14. Ingresos por turismo en miles de euros constantes e índices en base 2001

Años	Ingresos	Índices
1990	17.116,771	46,76
1991	17.129,724	46,80
1992	18.394,564	50,26
1993	19.498,528	53,27
1994	21.343,419	58,31
1995	22.443,647	61,32
1996	23.876,547	65,23
1997	26.421,075	72,18
1998	29.385,651	80,28
1999	32.589,217	89,04
2000	34.961,036	95,52
2001	36.602,357	100,00
2002	34.328,884	93,79
2003	34.560,984	94,42

Como hemos comentado, con respecto al año 2001, en el año 2003 en los ingresos por turismo en términos nominales se observa un crecimiento del 0,73 por ciento. Pero este crecimiento es sólo aparente, puesto que, a precios constantes; es decir, una vez eliminado el efecto de la inflación, la variable realmente experimenta una disminución del 5,60 por ciento.

TEMA 4
SERIES TEMPORALES

4.1. Definición y representación gráfica

Hemos dedicado el capítulo anterior al análisis de los cambios que una variable experimenta con respecto a un período de referencia. Nuestro objetivo es, ahora, el estudio de los métodos más sencillos para observar su evolución.

Una **serie temporal**, y, es el conjunto de valores de una variable ordenados en el tiempo. Cada uno de dichos valores se designa genéricamente como y_t, donde el subíndice t representa el período al que se refiere, que puede ser un año, trimestre, mes, semana, día o cualquier otra unidad temporal.

Para analizar su evolución, generalmente se comienza por representar gráficamente la serie, situando sus valores en ordenadas y en abscisas los períodos temporales a los que corresponden. En cada caso, debe elegirse la escala adecuada para que el gráfico refleje la evolución de la serie sin distorsionarla y para que puedan observarse sus principales movimientos: las oscilaciones a corto, medio y largo plazo, la existencia de valores anómalos, etcétera.

Supongamos, por ejemplo, que se dispone de la serie temporal y = número de visitantes (millones de personas) que entran en España correspondiente al período comprendido entre los años 1995 y 2003 que figura en la Tabla 4.1.

Tabla 4.1. Número de visitantes

Años	y = Nº visitantes
1995	54,41
1996	57,27
1997	62,41
1998	67,76
1999	72,06
2000	74,46
2001	75,68
2002	79,88
2003	82,59

Siguiendo el procedimiento descrito en el epígrafe 1.3 para efectuar representaciones gráficas con EXCEL, el programa selecciona de forma automática el tramo del eje de ordenadas en el que se encuentran los valores mínimo y máximo de la variable y del eje de abscisas correspondiente al período temporal de referencia y elige las escalas apropiadas, proporcionando para la serie el siguiente gráfico:

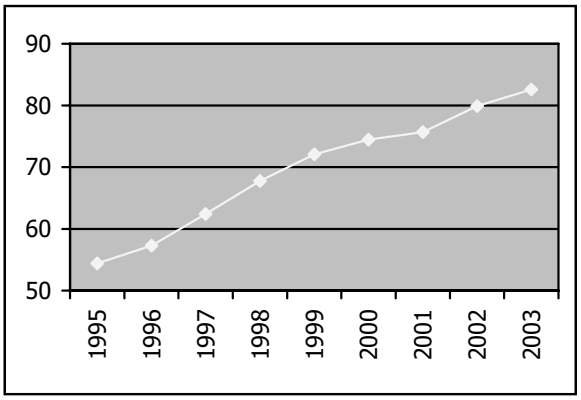


Gráfico 4.1. Número de visitantes

En él puede observarse que la variable crece, sin irregularidades importantes, en forma aproximadamente lineal, y se percibe una ralentización del crecimiento en el período comprendido entre 1999 y 2001.

De forma análoga, puede efectuarse el gráfico de la serie temporal x = ingresos (miles de euros) por turismo, recogida en la Tabla 4.2.

Tabla 4.2. Ingresos por turismo

Años	x = Ingresos
1995	19.038,78
1996	20.975,16
1997	23.667,83
1998	26.806,38
1999	30.415,59
2000	33.749,66
2001	36.602,36
2002	35.543,44
2003	36.871,04

Con estos datos, se obtiene el Grafico 4.2. En él se observa que la serie experimenta un crecimiento aproximadamente lineal con un ligero descenso en 2002, de manera que en 2003 toma un valor similar al registrado en 2001.

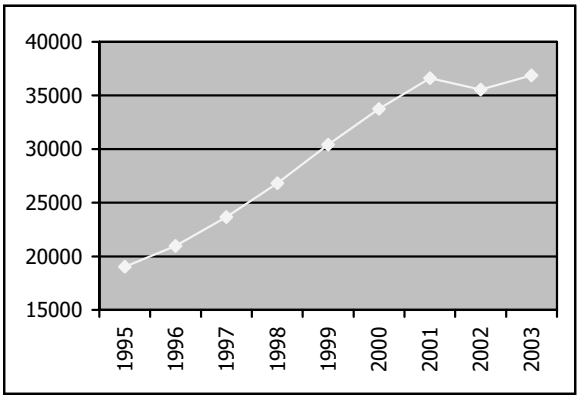


Gráfico 4.2. Ingresos por turismo

Ahora bien, dado que las dos series corresponden al mismo período temporal, puede ser interesante representar ambas en un solo gráfico, para observar las similitudes o disimilitudes en su comportamiento. Al hacerlo, se obtiene un resultado tal como el Gráfico 4.3.

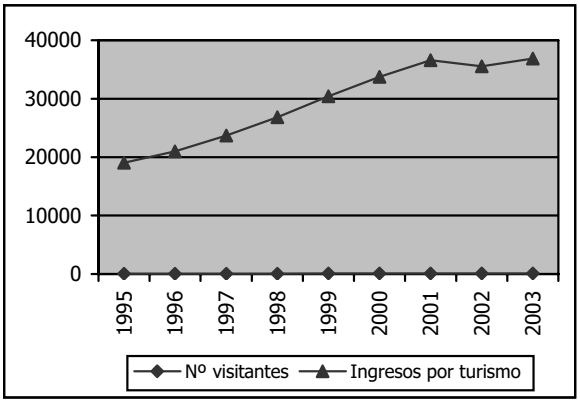


Gráfico 4.3. Escala simple

Como puede verse, si en el gráfico conjunto se utiliza una sola escala, como (dadas las unidades de medida de las variables) los valores del número de visitantes son mucho menores que los de los ingresos por turismo, la línea que los representa se superpone al eje de abscisas y no permite apreciar sus movimientos. Al mismo tiempo, al empezar la escala en valores mucho menores que el mínimo de los ingresos por turismo, la línea correspondiente se ve mucho más plana que en el gráfico individual, de manera que su crecimiento parece más lento.

Este problema se resuelve utilizando diferentes escalas para cada serie. Un gráfico de estas características puede efectuarse con EXCEL seleccionando la opción líneas en dos ejes en los tipos personalizados de gráfico.

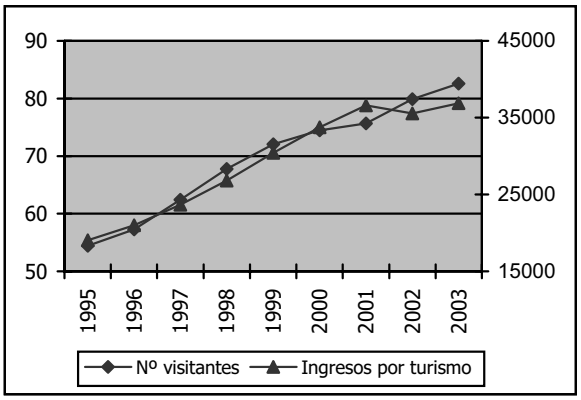


Gráfico 4.4. Doble escala

Así, en el Gráfico 4.4, la escala de la izquierda corresponde a los valores del número de visitantes y la de la derecha a los de los ingresos por turismo, de manera que pueden apreciarse con claridad los movimientos de ambas series.

4.2. Componentes de una serie de tiempo

El análisis clásico de las series temporales parte de la idea de que cada valor de la variable es el resultado de la combinación de cuatro factores no observables que determinan su evolución.

Dichos factores son el movimiento a largo plazo o **tendencia,** que refleja la marcha general del fenómeno analizado; los movimientos a medio plazo o **variaciones cíclicas**, que generalmente están relacionados con los ciclos económicos de prosperidad y recesión; las variaciones a corto plazo (trimestrales, mensuales, semanales, diarias o de otra periodicidad) o **variaciones estacionales**, ocasionadas por causas periódicas, y finalmente los movimientos esporádicos o **variaciones accidentales o residuales** derivados de acontecimientos ocasionales.

Las series temporales no siempre contienen los cuatro elementos sino que pueden estar formadas sólo por alguno o algunos de ellos. Además, dichos elementos pueden combinarse teóricamente de diferentes maneras, siendo los esquemas de formación más habituales el aditivo y el multiplicativo, que suponen que los valores de la variable son el resultado de sumar o de multiplicar sus componentes.

En las series de turismo, que son las que nos interesan principalmente, suele utilizarse el esquema multiplicativo. Las componentes de mayor interés son la tendencia y las variaciones estacionales. Describimos, a continuación, algunos de los métodos desarrollados para aislarlas.

4.3. Análisis de la tendencia mediante el método de las medias móviles

Uno de los métodos más utilizados para aislar la tendencia es el de las medias móviles, que consiste en suavizar las fluctuaciones de la serie sustituyendo cada observación por el promedio de dicha observación y las más cercanas.

El número de datos con el que se calculan las medias móviles debe corresponder a una unidad temporal más larga que la de la serie, teniendo en cuenta que cuanto mayor es el número de términos más se suavizan las irregularidades. Generalmente, si la serie es trimestral las medias son de cuatro términos, si es mensual de doce, etcétera.

Cuando para calcular la media se utilizan p términos, se dice que la media móvil es de orden p, en cuyo caso el promedio de p valores consecutivos de la serie corresponde a una observación de la variable. Para determinar la correspondencia entre los valores de la variable y los promedios calculados, hay que distinguir dos situaciones:

1. **p es impar**

En este caso, el primer valor de la media móvil sustituye a la observación correspondiente al momento ½ (p+1), el segundo al momento ½ (p+3), el tercero al momento ½ (p+5), y así sucesivamente. Es decir, cada una de las medias móviles sustituye a la observación correspondiente al momento ½ (p+n), siendo n la sucesión de los números naturales impares.

En la Tabla 4.3 se presenta el procedimiento para obtener la tendencia mediante una media móvil de orden p = 3, en el caso de una serie temporal, y, compuesta por siete observaciones.

Tabla 4.3. Serie de medias móviles de orden p = 3

Nº observación	y_t	Media móvil de orden 3 (MM3)
1	y_1	
2	y_2	$\bar{y}_{\frac{P+1}{2}} = \frac{y_1 + y_2 + ... + y_P}{P} = \frac{y_1 + y_2 + y_3}{3} = \bar{y}_2$
3	y_3	$\bar{y}_{\frac{P+3}{2}} = \frac{y_2 + y_3 + ... + y_{P+1}}{P} = \frac{y_2 + y_3 + y_4}{3} = \bar{y}_3$
4	y_4	$\bar{y}_{\frac{P+5}{2}} = \frac{y_3 + y_4 + ... + y_{P+2}}{P} = \frac{y_3 + y_4 + y_5}{3} = \bar{y}_4$
5	y_5	$\bar{y}_{\frac{P+7}{2}} = \frac{y_4 + y_5 + ... + y_{P+3}}{P} = \frac{y_4 + y_5 + y_6}{3} = \bar{y}_5$
6	y_6	$\bar{y}_{\frac{P+9}{2}} = \frac{y_5 + y_6 + ... + y_{P+4}}{P} = \frac{y_5 + y_6 + y_7}{3} = \bar{y}_6$
7	y_7	

En este caso, como puede observarse, no se dispone de información suficiente para el cálculo del primer y del último valor de la serie de medias móviles. En general, si el orden de la media móvil, p, es impar, se pierden las ½ (p+1) – 1 primeras y últimas observaciones.

2. **p es par**

En este caso ½ (p+1), ½ (p+3), ½ (p+5), etcétera, no son números enteros, por lo que a cada media móvil le corresponde un punto intermedio situado entre dos valores consecutivos de la variable y no se puede determinar a cuál de ellos debe

sustituir. Es decir, la serie de medias móviles queda "descentrada" con respecto a los valores de la variable.

Por ejemplo, al calcular la media móvil de orden p = 4 para la serie temporal y, se presenta la siguiente situación:

Tabla 4.4. Serie de medias móviles descentradas de orden p = 4

Nº observación	y_t	Media móvil de orden 4 descentrada ($MM4_d$)
1	y_1	
2	y_2	
		$\overline{y}_{\frac{P+1}{2}} = \frac{y_1 + y_2 + ... + y_P}{P} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \overline{y}_{2,5}$
3	y_3	
		$\overline{y}_{\frac{P+3}{2}} = \frac{y_2 + y_3 + ... + y_{P+1}}{P} = \frac{y_2 + y_3 + y_4 + y_5}{4} = \overline{y}_{3,5}$
4	y_4	
		$\overline{y}_{\frac{P+5}{2}} = \frac{y_3 + y_4 + ... + y_{P+2}}{P} = \frac{y_3 + y_4 + y_5 + y_6}{4} = \overline{y}_{4,5}$
5	y_5	
		$\overline{y}_{\frac{P+7}{2}} = \frac{y_4 + y_5 + ... + y_{P+3}}{P} = \frac{y_4 + y_5 + y_6 + y_7}{4} = \overline{y}_{5,5}$
6	y_6	
7	y_7	

Para resolver este inconveniente debe calcularse una media móvil entre cada dos medias móviles descentradas consecutivas, de tal forma que el primer valor de la serie de medias móviles centradas sustituye a la observación correspondiente al momento ½ (p+2), el segundo al momento ½ (p+4), el tercero al momento ½ (p+6) y así sucesivamente. Es decir, cada una de las medias móviles sustituye a la observación correspondiente al momento ½ (p+n), siendo n ahora la sucesión de los números naturales pares.

El procedimiento se describe en la Tabla 4.5, en la que $MM4_d$ y $MM4_C$ representan las medias móviles de orden 4 descentradas y centradas.

Tabla 4.5. Serie de medias móviles descentradas y centradas de orden p = 4

Obs.	y_t	$MM4_d$	$MM4_C$
1	y_1		
2	y_2		
		$\bar{y}_{\frac{P+1}{2}} = \frac{y_1 + y_2 + ... + y_P}{P} = \frac{y_1 + y_2 + y_3 + y_4}{4} = \bar{y}_{2,5}$	
3	y_3	----------▶	$\bar{\bar{y}}_{\frac{P+2}{2}} = \frac{\bar{y}_{\frac{P+1}{2}} + \bar{y}_{\frac{P+3}{2}}}{2}$
		$\bar{y}_{\frac{P+3}{2}} = \frac{y_2 + y_3 + ... + y_{P+1}}{P} = \frac{y_2 + y_3 + y_4 + y_5}{4} = \bar{y}_{3,5}$	
4	y_4	----------▶	$\bar{\bar{y}}_{\frac{P+4}{2}} = \frac{\bar{y}_{\frac{P+3}{2}} + \bar{y}_{\frac{P+5}{2}}}{2}$
		$\bar{y}_{\frac{P+5}{2}} = \frac{y_3 + y_4 + ... + y_{P+2}}{P} = \frac{y_3 + y_4 + y_5 + y_6}{4} = \bar{y}_{4,5}$	
5	y_5	----------▶	$\bar{\bar{y}}_{\frac{P+6}{2}} = \frac{\bar{y}_{\frac{P+5}{2}} + \bar{y}_{\frac{P+7}{2}}}{2}$
		$\bar{y}_{\frac{P+7}{2}} = \frac{y_4 + y_5 + ... + y_{P+3}}{P} = \frac{y_4 + y_5 + y_6 + y_7}{4} = \bar{y}_{5,5}$	
6	y_6		
7	y_7		

En este caso no se pueden calcular los dos primeros y los dos últimos valores de las series de medias móviles. En general, si el orden de la media móvil, p, es par, se pierden las ½ (p+2) – 1 primeras y últimas observaciones.

Como ejercicio para la aplicación del procedimiento en cada uno de los dos casos descritos, puede extraerse la tendencia de la serie y = ingresos (miles de euros) correspondientes a un grupo de establecimientos turísticos en el período comprendido

entre 1990 y 2003, mediante una media móvil de orden impar (p = 3), con la cual se obtienen los resultados recogidos en la Tabla 4.6, y otra de orden par (p = 4), que proporciona los que figuran en la Tabla 4.7.

Tabla 4.6. Medias móviles de orden impar (p = 3) para la serie y = ingresos

Años	Ingresos	MM3
1990	1.202,02	
1991	1.262,13	1.232,07
1992	1.232,07	1.292,17
1993	1.382,33	1.312,20
1994	1.322,23	1.352,27
1995	1.352,28	1.340,25
1996	1.346,27	1.356,28
1997	1.370,31	1.398,83
1998	1.479,93	1.468,05
1999	1.553,93	1.524,08
2000	1.538,39	1.538,44
2001	1.523,01	1.540,82
2002	1.561,08	1.563,99
2003	1.607,91	

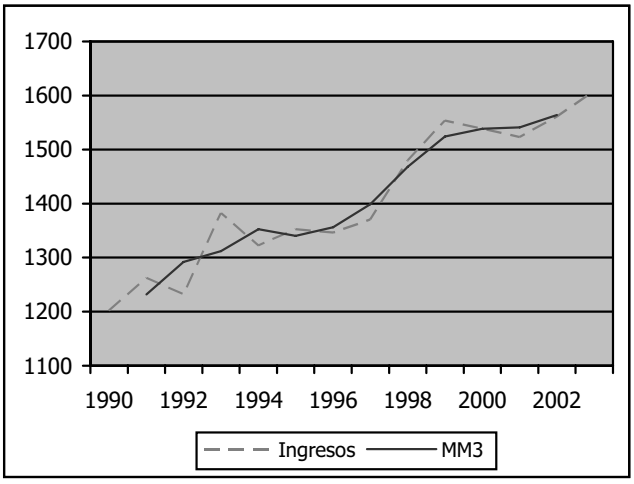


Gráfico 4.5. Serie ingresos y medias móviles de orden 3

Tabla 4.7. Medias móviles de orden par (p = 4) para la serie y = ingresos

Años	Ingresos	$MM4_d$	$MM4_c$
1990	1.202,02		
1991	1.262,13		
		1.269,63	
1992	1.232,07		1.284,66
		1.299,68	
1993	1.382,33		1.310,95
		1.322,22	
1994	1.322,23		1.336,50
		1.350,77	
1995	1.352,28		1.349,27
		1.347,76	
1996	1.346,27		1.367,48
		1.387,19	
1997	1.370,31		1.412,40
		1.437,60	
1998	1.479,93		1.461,62
		1.485,63	
1999	1.553,93		1.504,72
		1.523,81	
2000	1.538,39		1.533,95
		1.544,10	
2001	1.523,01		1.550,84
		1.557,59	
2002	1.561,08		
2003	1.607,91		

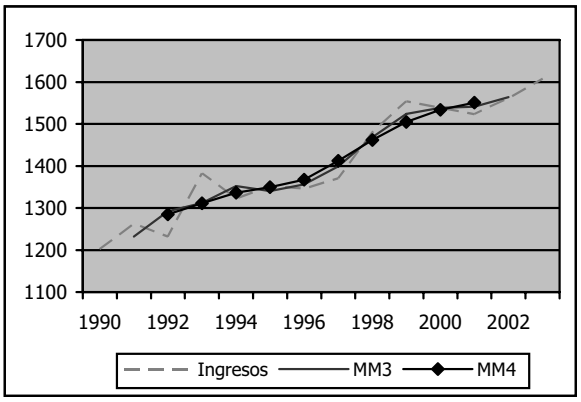


Gráfico 4.6. Serie ingresos y medias móviles de órdenes 3 y 4

En el Gráfico 4.5 puede observarse claramente como la serie de medias móviles, al haberse obtenido promediando los valores de los ingresos, suaviza sus irregularidades, aislando la tendencia.

El Gráfico 4.6 muestra que la línea que corresponde a las medias móviles de cuarto orden es menos irregular que la que corresponde a las de orden tres; es decir que, como hemos dicho, a medida que el valor de p aumenta, las fluctuaciones de la serie original se suavizan más.

4.4. Análisis de las variaciones estacionales mediante el método de la razón a la media móvil. Desestacionalización

Al analizar el comportamiento de los indicadores de la demanda de turismo, como consecuencia de la concentración temporal de la actividad turística en el tercer trimestre del año, que coincide con las vacaciones escolares y laborales para una gran parte de la población ocupada, suele observarse la existencia de una fuerte componente estacional. Resulta, entonces, de gran interés aislar esta componente, tanto para su análisis específico, como para evitar conclusiones erróneas al comparar datos correspondientes a diferentes meses o trimestres del mismo año.

Uno de los métodos más sencillos para hacerlo es el de de la razón a la media móvil, cuya aplicación comienza con el cálculo de una serie de medias móviles de cuatro términos si la serie es trimestral, de doce si es mensual, etcétera, teniendo en cuenta que si el orden de las medias móviles es par, como es habitual, ha de procederse al centrado de la serie.

Aunque teóricamente las series pueden formarse siguiendo los esquemas aditivo o multiplicativo, el que se adopta con mayor frecuencia para las series de turismo es el esquema multiplicativo, por lo que nos referiremos exclusivamente a él.

En este caso, las componentes estacional y residual se aíslan dividiendo los valores de la serie original entre las medias móviles calculadas previamente, y el cociente obtenido, o razón a la media móvil, se denomina **índice de variación estacional**, IVE.

$$IVE_t = \frac{y_t}{MMP_t} = \frac{(T \times C \times E \times R)_t}{(T \times C)_t} = (E \times R)_t$$

Aunque se puede expresar en tanto por uno, generalmente se multiplica el resultado por cien, y el índice se interpreta como el porcentaje que la serie original representa respecto a la de medias móviles, que recoge la tendencia y el ciclo, de manera que cuantifica los efectos de la estacionalidad y la irregularidad en cada período. Por ejemplo, un índice de variación estacional igual a 120 indica que, debido al efecto estacional y residual, en el período considerado se observa en la serie un valor un 20 por ciento superior al que corresponde a la tendencia y el ciclo.

Promediando los índices de variación estacional se obtiene el **índice medio de variación estacional**, IME, que, si la estacionalidad es estable y se admite que los errores en cada subperíodo tienden a compensarse, sólo incluye la componente estacional. Por ejemplo, en una serie mensual, el índice medio de variación estacional del mes de enero se obtiene como la media aritmética de los índices de variación estacional de todos los meses de enero disponibles en la muestra, y para completar la serie, el proceso se repite para cada mes del año.

Los índices medios de variación estacional de una serie mensual deben sumar 1200 ya que, si no existiese variación estacional, en cada mes su valor sería igual a 100. De forma similar, si la serie es trimestral deben sumar 400. Cuando esto no sucede, cada índice se ajuste mediante una simple regla de tres, denominándose ahora, **índice medio de variación estacional ajustado**, IMEA.

Al dividir la serie original entre estos índices ajustados, se elimina la componente estacional y multiplicando por cien el cociente, se obtiene la **serie desestacionalizada**, que es la que debe utilizarse, por ejemplo, para comparar datos correspondientes a distintos meses o trimestres del mismo año.

$$yd_t = \frac{y_t}{IMEA_t} \times 100 = \frac{(T \times C \times E \times R)_t}{E_t \times 100} \times 100 = (T \times C)_t \times R_t$$

Para ver un ejemplo se procede, a continuación, a aislar la componente estacional y a desestacionalizar la variable y = viajeros (miles de personas) alojados en establecimientos hoteleros, utilizando la serie de periodicidad mensual que proporciona el Banco de Datos TEMPUS del Instituto Nacional de Estadística (http://www.ine.es), que se recoge en la Tabla 4.8.

Tabla 4.8. Viajeros alojados en establecimientos hoteleros

t	y_t	t	y_t	t	y_t	t	y_t	t	y_t
94.01	1.632,00	96.01	1.807,00	98.01	2.120,00	00.01	2.809,00	02.01	2.873,00
94.02	1.926,00	96.02	2.051,00	98.02	2.480,00	00.02	3.391,00	02.02	3.508,00
94.03	2.608,00	96.03	2.756,00	98.03	3.164,00	00.03	4.244,00	02.03	4.757,00
94.04	2.895,00	96.04	3.442,00	98.04	3.999,00	00.04	5.335,00	02.04	4.853,00
94.05	3.479,00	96.05	3.841,00	98.05	4.492,00	00.05	5.592,00	02.05	5.677,00
94.06	3.527,00	96.06	3.887,00	98.06	4.455,00	00.06	5.807,00	02.06	5.726,00
94.07	3.898,00	96.07	4.139,00	98.07	4.940,00	00.07	6.374,00	02.07	6.408,00
94.08	4.242,00	96.08	4.574,00	98.08	5.474,00	00.08	6.995,00	02.08	7.384,00
94.09	3.984,00	96.09	4.205,00	98.09	5.039,00	00.09	6.328,00	02.09	6.225,00
94.10	3.183,00	96.10	3.543,00	98.10	4.259,00	00.10	5.369,00	02.10	5.360,00
94.11	2.126,00	96.11	2.362,00	98.11	2.813,00	00.11	3.691,00	02.11	3.715,00
94.12	1.950,00	96.12	2.124,00	98.12	2.547,00	00.12	3.348,00	02.12	3.383,00
95.01	1.787,00	97.01	1.867,00	99.01	2.776,00	01.01	3.005,00	03.01	2.971,00
95.02	1.996,00	97.02	2.212,00	99.02	3.206,00	01.02	3.519,00	03.02	3.464,00
95.03	2.586,00	97.03	3.075,00	99.03	4.143,00	01.03	4.333,00	03.03	4.422,00
95.04	3.527,00	97.04	3.329,00	99.04	4.931,00	01.04	5.385,00	03.04	5.409,00
95.05	3.639,00	97.05	4.081,00	99.05	5.725,00	01.05	5.373,00	03.05	5.985,00
95.06	3.680,00	97.06	4.085,00	99.06	5.834,00	01.06	5.914,00	03.06	6.277,00
95.07	4.009,00	97.07	4.554,00	99.07	6.415,00	01.07	6.406,00	03.07	6.796,00
95.08	4.370,00	97.08	5.091,00	99.08	6.986,00	01.08	7.179,00	03.08	7.698,00
95.09	4.075,00	97.09	4.593,00	99.09	6.350,00	01.09	6.307,00	03.09	6.518,00
95.10	3.418,00	97.10	3.816,00	99.10	5.448,00	01.10	5.310,00	03.10	5.537,00
95.11	2.282,00	97.11	2.533,00	99.11	3.571,00	01.11	3.658,00	03.11	3.868,00
95.12	2.060,00	97.12	2.362,00	99.12	3.204,00	01.12	3.215,00	03.12	3.589,00

El proceso comienza aislando los componentes tendencia y ciclo mediante una media móvil de doce términos centrada respecto a la serie original, cuyo valor no puede obtenerse ni para las seis primeras ni para las seis últimas observaciones. Los demás resultados figuran en la Tabla 4.9.

Tabla 4.9. Tendencia y ciclo de la serie y = viajeros

t	$MM12_c$	t	$MM12_c$	t	$MM12_c$	t	$MM12_c$	t	$MM12_c$
94.01		96.01	3.171,92	98.01	3.654,33	00.01	4.927,63	02.01	4.955,83
94.02		96.02	3.185,83	98.02	3.686,38	00.02	4.926,29	02.02	4.964,46
94.03		96.03	3.199,75	98.03	3.720,92	00.03	4.925,75	02.03	4.969,58
94.04		96.04	3.210,38	98.04	3.757,96	00.04	4.921,54	02.04	4.968,25
94.05		96.05	3.218,92	98.05	3.788,08	00.05	4.923,25	02.05	4.972,71
94.06		96.06	3.224,92	98.06	3.807,46	00.06	4.934,25	02.06	4.982,08
94.07	2.877,29	96.07	3.230,00	98.07	3.842,50	00.07	4.948,42	02.07	4.993,17
94.08	2.886,67	96.08	3.239,13	98.08	3.900,08	00.08	4.961,92	02.08	4.995,42
94.09	2.888,67	96.09	3.259,13	98.09	3.971,13	00.09	4.970,96	02.09	4.979,63
94.10	2.914,08	96.10	3.267,71	98.10	4.050,75	00.10	4.976,75	02.10	4.988,83
94.11	2.988,75	96.11	3.273,00	98.11	4.140,96	00.11	4.982,21	02.11	5.024,83
94.12	3.043,46	96.12	3.291,25	98.12	4.249,79	00.12	4.990,04	02.12	5.060,63
95.01	3.054,46	97.01	3.316,79	99.01	4.368,71	01.01	4.995,83	03.01	5.099,75
95.02	3.064,42	97.02	3.355,63	99.02	4.493,17	01.02	5.004,83	03.02	5.129,00
95.03	3.073,54	97.03	3.393,33	99.03	4.610,79	01.03	5.011,63	03.03	5.154,29
95.04	3.087,13	97.04	3.420,88	99.04	4.714,96	01.04	5.008,29	03.04	5.173,88
95.05	3.103,42	97.05	3.439,38	99.05	4.796,08	01.05	5.004,46	03.05	5.187,63
95.06	3.114,50	97.06	3.456,42	99.06	4.855,04	01.06	4.997,54	03.06	5.202,58
95.07	3.119,92	97.07	3.476,96	99.07	4.883,79	01.07	4.986,50	03.07	
95.08	3.123,04	97.08	3.498,75	99.08	4.892,88	01.08	4.980,54	03.08	
95.09	3.132,42	97.09	3.513,63	99.09	4.904,79	01.09	4.997,75	03.09	
95.10	3.135,96	97.10	3.545,25	99.10	4.925,83	01.10	4.993,25	03.10	
95.11	3.140,83	97.11	3.590,29	99.11	4.937,13	01.11	4.971,25	03.11	
95.12	3.157,88	97.12	3.622,83	99.12	4.930,46	01.12	4.963,58	03.12	

Para obtener los índices de variación estacional, IVE, recogidos en la Tabla 4.10, se divide la serie original entre la de medias móviles, de manera que también quedan sin determinar los seis primeros y los seis últimos datos.

Tabla 4.10. Índices de variación estacional

t	IVE_t	t	IVE_t	t	IVE_t	t	IVE_t	t	IVE_t
94.01		96.01	**56,97**	98.01	**58,01**	00.01	**57,01**	02.01	**57,97**
94.02		96.02	64,38	98.02	67,28	00.02	68,84	02.02	70,66
94.03		96.03	86,13	98.03	85,03	00.03	86,16	02.03	95,72
94.04		96.04	107,22	98.04	106,41	00.04	108,40	02.04	97,68
94.05		96.05	119,33	98.05	118,58	00.05	113,58	02.05	114,16
94.06		96.06	120,53	98.06	117,01	00.06	117,69	02.06	114,93
94.07	135,48	96.07	128,14	98.07	128,56	00.07	128,81	02.07	128,34
94.08	146,95	96.08	141,21	98.08	140,36	00.08	140,97	02.08	147,82
94.09	137,92	96.09	129,02	98.09	126,89	00.09	127,30	02.09	125,01
94.10	109,23	96.10	108,43	98.10	105,14	00.10	107,88	02.10	107,44
94.11	71,13	96.11	72,17	98.11	67,93	00.11	74,08	02.11	73,93
94.12	64,07	96.12	64,54	98.12	59,93	00.12	67,09	02.12	66,85
95.01	**58,51**	97.01	**56,29**	99.01	**63,54**	01.01	**60,15**	03.01	**58,26**
95.02	65,14	97.02	65,92	99.02	71,35	01.02	70,31	03.02	67,54
95.03	84,14	97.03	90,62	99.03	89,85	01.03	86,46	03.03	85,79
95.04	114,25	97.04	97,31	99.04	104,58	01.04	107,52	03.04	104,54
95.05	117,26	97.05	118,66	99.05	119,37	01.05	107,36	03.05	115,37
95.06	118,16	97.06	118,19	99.06	120,16	01.06	118,34	03.06	120,65
95.07	128,50	97.07	130,98	99.07	131,35	01.07	128,47	03.07	
95.08	139,93	97.08	145,51	99.08	142,78	01.08	144,14	03.08	
95.09	130,09	97.09	130,72	99.09	129,47	01.09	126,20	03.09	
95.10	108,99	97.10	107,64	99.10	110,60	01.10	106,34	03.10	
95.11	72,66	97.11	70,55	99.11	72,33	01.11	73,58	03.11	
95.12	65,23	97.12	65,20	99.12	64,98	01.12	64,77	03.12	

Los valores que en la tabla anterior se han destacado en negrita son los índices de variación estacional correspondientes a los meses de enero de todos los años observados. Por ejemplo, en el año 1995, el IVE del mes de enero es 58,51 lo que indica que, debido al efecto estacional, el valor de la serie es un 41,49 por ciento inferior al de la tendencia y el ciclo. La media aritmética de estos valores es el índice medio, IME, correspondiente al mes de Enero:

$$IME_{enero} = \frac{\Sigma IVE_{enero}}{\text{Nº meses enero en la muestra}} = \frac{526,71}{9} = 58,52$$

Realizando esta operación para los doce meses del año se han obtenido los índices medios de variación estacional que se muestran en la Tabla 4.11.

Tabla 4.11. Índices medios de variación estacional

Meses	IME_t
Enero	58,52
Febrero	67,93
Marzo	87,77
Abril	105,32
Mayo	115,96
Junio	118,41
Julio	129,85
Agosto	143,30
Septiembre	129,18
Octubre	107,97
Noviembre	72,04
Diciembre	64,74

La suma de estos índices resulta igual a 1200,99 y, dado que debería ser 1200, han de ajustarse, de manera que cada uno de ellos se multiplica por 1200 y se divide por 1200,99 para obtener los índices ajustados de la Tabla 4.12.

Tabla 4.12. Índices medios de variación estacional ajustados

Meses	$IMEA_t$
Enero	58,47
Febrero	67,88
Marzo	87,70
Abril	105,24
Mayo	115,87
Junio	118,31
Julio	129,74
Agosto	143,18
Septiembre	129,07
Octubre	107,88
Noviembre	71,98
Diciembre	64,69

Como puede observarse, en los meses de abril a octubre el índice medio ajustado es superior a cien, lo que indica que en estos meses el número de viajeros alojados en

establecimientos hoteleros es superior a la media mensual del año. Así, en el mes de julio, por ejemplo, es superior en un 29,74 por ciento. Sin embargo, en los meses de enero, febrero, marzo, noviembre y diciembre el índice medio ajustado es inferior a cien, lo que indica que en estos meses el número de viajeros alojados en establecimientos hoteleros es inferior a la media mensual del año. Por ejemplo, en el mes de enero, es inferior en un 41,53 por ciento.

Finalmente, para proceder a la desestacionalización, se dividen los valores de la serie original de viajeros por los correspondientes índices medios ajustados, y se obtiene la Tabla 4.13.

Tabla 4.13. Serie de viajeros desestacionalizada

t	yd_t	t	yd_t	t	yd_t	t	yd_t	t	yd_t
94.01	2790,96	96.01	3090,23	98.01	3625,51	00.01	4803,80	02.01	4913,25
94.02	2837,43	96.02	3021,59	98.02	3653,60	00.02	4995,71	02.02	5168,08
94.03	2973,93	96.03	3142,69	98.03	3607,94	00.03	4839,47	02.03	5424,45
94.04	2750,91	96.04	3270,68	98.04	3799,96	00.04	5069,46	02.04	4611,45
94.05	3002,55	96.05	3314,98	98.05	3876,82	00.05	4826,18	02.05	4899,54
94.06	2981,19	96.06	3285,48	98.06	3765,58	00.06	4908,35	02.06	4839,89
94.07	3004,48	96.07	3190,24	98.07	3807,63	00.07	4912,92	02.07	4939,13
94.08	2962,74	96.08	3194,62	98.08	3823,21	00.08	4885,52	02.08	5157,21
94.09	3086,62	96.09	3257,84	98.09	3903,99	00.09	4902,65	02.09	4822,85
94.10	2950,59	96.10	3284,30	98.10	3948,02	00.10	4976,97	02.10	4968,63
94.11	2953,53	96.11	3281,40	98.11	3907,95	00.11	5127,70	02.11	5161,04
94.12	3014,48	96.12	3283,46	98.12	3937,38	00.12	5175,63	02.12	5229,74
95.01	3056,03	97.01	3192,84	99.01	4747,36	01.01	5138,99	03.01	5080,84
95.02	2940,56	97.02	3258,78	99.02	4723,16	01.02	5184,28	03.02	5103,26
95.03	2948,84	97.03	3506,45	99.03	4724,30	01.03	4940,96	03.03	5042,45
95.04	3351,45	97.04	3163,31	99.04	4685,57	01.04	5116,97	03.04	5139,78
95.05	3140,64	97.05	3522,11	99.05	4940,96	01.05	4637,17	03.05	5165,35
95.06	3110,51	97.06	3452,83	99.06	4931,17	01.06	4998,79	03.06	5305,62
95.07	3090,04	97.07	3510,11	99.07	4944,52	01.07	4937,58	03.07	5238,19
95.08	3052,14	97.08	3555,71	99.08	4879,23	01.08	5014,03	03.08	5376,52
95.09	3157,13	97.09	3558,45	99.09	4919,69	01.09	4886,38	03.09	5049,85
95.10	3168,43	97.10	3537,37	99.10	5050,20	01.10	4922,28	03.10	5132,71
95.11	3170,26	97.11	3518,96	99.11	4960,99	01.11	5081,86	03.11	5373,60
95.12	3184,53	97.12	3651,39	99.12	4953,02	01.12	4970,03	03.12	5548,19

Por último, la serie original se representa junto con la de medias móviles de orden p = 12, y la serie desestacionalizada.

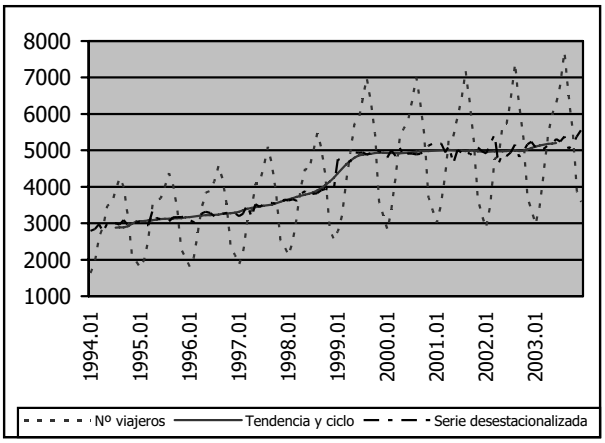


Gráfico 4.7. Número de viajeros, tendencia y ciclo, y serie desestacionalizada

En este gráfico, la serie de medias móviles refleja la tendencia y el ciclo y la desestacionalizada muestra la evolución del número de viajeros una vez eliminada la componente estacional.

4.5. Análisis de la evolución temporal de una serie. Tasa de variación

Hemos visto ya que los números índices permiten establecer comparaciones entre el valor que toma una serie en un determinado período y el correspondiente a una situación de referencia que se toma como base.

Pero, cuando lo que interesa es comparar cada valor observado con el anterior en el tiempo, la situación de referencia varía para cada observación, de manera que deben calcularse números índices con un período base variable (índices en cadena) o puede recurrirse al cálculo de la tasa de variación.

La **tasa de variación** o variación relativa de una serie en el período t respecto al anterior, t – 1, indica, en términos porcentuales, el cambio experimentado por la variable respecto al período anterior. Por tanto, se define como el cociente entre la

variación de la variable en el período t ($\Delta y_t = y_t - y_{t-1}$) dividida por su valor en el período inicial (y_{t-1}), multiplicado por cien.

$$TV_t = \frac{\Delta y_t}{y_{t-1}} \times 100 = \frac{y_t - y_{t-1}}{y_{t-1}} \times 100 = \left[\frac{y_t}{y_{t-1}} - 1 \right] \times 100$$

Dada esta definición, la tasa de variación es adimensional, por lo que es un instrumento adecuado para comparar la evolución de diferentes series, aunque estén expresadas en unidades de medida dispares.

Suponiendo que, como es habitual, $y_{t-1} > 0$, una tasa de variación positiva, indica que $y_t > y_{t-1}$; es decir, que del período t – 1 al período t la variable ha experimentado un aumento, y una tasa de variación negativa indica que $y_t < y_{t-1}$; es decir, que del período t – 1 al período t la variable ha experimentado una disminución.

Como ejemplo, se han obtenido las tasas de variación intermensual de la serie y = viajeros alojados en establecimientos hoteleros en el año 2002 y en el año 2003, utilizando parte de los datos del ejercicio de desestacionalización. Los resultados obtenidos figuran en la Tabla 4.14.

Tabla 4.14. Tasas de variación intermensual

Meses	y_t	TV_t	Meses	y_t	TV_t
2002.01	2.873	-10,64	2003.01	2.971	-12,18
2002.02	3.508	22,10	2003.02	3.464	16,59
2002.03	4.757	35,60	2003.03	4.422	27,66
2002.04	4.853	2,01	2003.04	5.409	22,32
2002.05	5.677	16,98	2003.05	5.985	10,65
2002.06	5.726	0,86	2003.06	6.277	4,88
2002.07	6.408	11,92	2003.07	6.796	8,27
2002.08	7.384	15,23	2003.08	7.698	13,27
2002.09	6.225	-15,69	2003.09	6.518	-15,33
2002.10	5.360	-13,89	2003.10	5.537	-15,05
2002.11	3.715	-30,69	2003.11	3.868	-30,14
2002.12	3.383	-8,94	2003.12	3.589	-7,21

Los valores que toma la tasa de variación, tanto en 2002 como en 2003, indican que el número de viajeros experimenta un crecimiento desde febrero hasta agosto y una disminución desde septiembre hasta enero. Por ejemplo, entre febrero y marzo de 2002 la variable aumenta un 35,60 por ciento, y entre noviembre y diciembre de 2003 disminuye un 7,21 por ciento.

Si calculásemos las tasas de variación intermensual para otros años, encontraríamos que esta situación se repite, porque el número de viajeros alojados en establecimientos hoteleros tiene todos los años un comportamiento similar, alcanzando valores mínimos en los meses de enero y máximos en los de agosto.

Por esta razón, cuando la serie que se desea analizar tiene una fuerte componente estacional, para analizar su evolución temporal suele calcularse, además, o bien la tasa de variación interanual, TV_t^{12}, que equivale a la tasa de variación intermensual una vez desestacionalizada la serie, TV_t, y refleja los cambios que la variable experimenta en un período de doce meses consecutivos.

Tabla 4.15. Tasa de variación interanual

Meses	2002	2003	TV_t^{12}
Enero	2.873	2.971	3,41
Febrero	3.508	3.464	-1,25
Marzo	4.757	4.422	-7,04
Abril	4.853	5.409	11,46
Mayo	5.677	5.985	5,43
Junio	5.726	6.277	9,62
Julio	6.408	6.796	6,05
Agosto	7.384	7.698	4,25
Septiembre	6.225	6.518	4,71
Octubre	5.360	5.537	3,30
Noviembre	3.715	3.868	4,11
Diciembre	3.383	3.589	6,09

Tabla 4.16. Tasa de variación intermensual de la serie desestacionalizada

Meses	2002	2003	TV_t
Enero	4.913,25	5.080,84	3,41
Febrero	5.168,08	5.103,26	-1,25
Marzo	5.424,45	5.042,45	-7,04
Abril	4.611,45	5.139,78	11,46
Mayo	4.899,54	5.165,35	5,43
Junio	4.839,89	5.305,62	9,62
Julio	4.939,13	5.238,19	6,05
Agosto	5.157,21	5.376,52	4,25
Septiembre	4.822,85	5.049,85	4,71
Octubre	4.968,63	5.132,71	3,30
Noviembre	5.161,04	5.373,60	4,11
Diciembre	5.229,74	5.548,19	6,09

Como hemos dicho, los resultados contenidos en cualquiera de estas dos tablas permiten cuantificar el cambio experimentado por el número de viajeros en un determinado mes del año con respecto al mismo mes del año anterior. Así, por ejemplo, en julio de 2003, con respecto a julio de 2002, el número de viajeros ha experimentado un crecimiento ligeramente superior al 6 por ciento.

TEMA 5
DISTRIBUCIONES BIDIMENSIONALES DE FRECUENCIAS

5.1. Introducción

Nos hemos ocupado hasta ahora de los métodos que permiten analizar el comportamiento de una sola variable. En el enfoque bivariante, que veremos a continuación, se procede al estudio conjunto de dos variables, analizando cómo se disponen sus datos en una distribución bidimensional de frecuencias y las técnicas estadísticas más sencillas para cuantificar la relación entre ellas.

Antes de nada, conviene señalar que cuando entre un par de variables se observa la existencia de alguna relación se suele interpretar como de causalidad, pero puede adoptar otras formas.

Una **relación** entre dos variables x e y es **de causalidad** sólo cuando una de ellas es la causa y la otra es el efecto. En particular se dice que x causa a y si y sólo si las variaciones en x provocan variaciones en y. Esta relación de causa y efecto puede ser simultánea o desfasada en el tiempo. Por ejemplo, el número de turistas que visita un lugar es en gran parte la causa de los ingresos por turismo en dicho lugar. En cuanto a la relación desfasada, el número de plazas en establecimientos hoteleros en un determinado período es en parte la causa del número de plazas en el período siguiente, especialmente si entre los dos períodos transcurre poco tiempo.

Una **relación de interdependencia** supone, sin embargo, que las dos variables son a la vez causa y efecto; es decir, que las variaciones de x provocan variaciones en y y simultáneamente las variaciones de y provocan variaciones en x. Por ejemplo, el precio de los alojamientos depende del número de turistas alojados en un destino y a la vez, el número de turistas alojados en un destino influye en el precio de los alojamientos.

En una **relación indirecta**, la asociación que se observa entre las variables x e y se debe a que ambas dependen de una tercera variable z. Por ejemplo, los niveles de ocupación mensuales en distintos destinos turísticos de "sol y playa" próximos están,

generalmente, muy relacionados, debido a que todos ellos dependen de la estación del año.

Finalmente, también puede suceder que la asociación entre los valores de x e y sea fruto de la casualidad, siendo ésta la **relación** que conocemos como **espuria**. Por ejemplo, es posible que el número de empleados en los establecimientos hoteleros de un país esté muy relacionado con la superficie forestal quemada, pero esto sucede, únicamente, por casualidad. No existe una relación de causalidad, de causa y efecto entre las variables. La asociación entre ellas se debe al azar.

Debe tenerse en cuenta que la información que proporcionan los datos de las variables no permite determinar cuál es la clase o el tipo de relación que existe entre ellas, que se supone a priori, a partir de los conocimientos teóricos respecto al comportamiento de un determinado fenómeno. Los métodos estadísticos que vamos a desarrollar son válidos, únicamente, para confirmar o rechazar la existencia de dicha relación y valorar su intensidad.

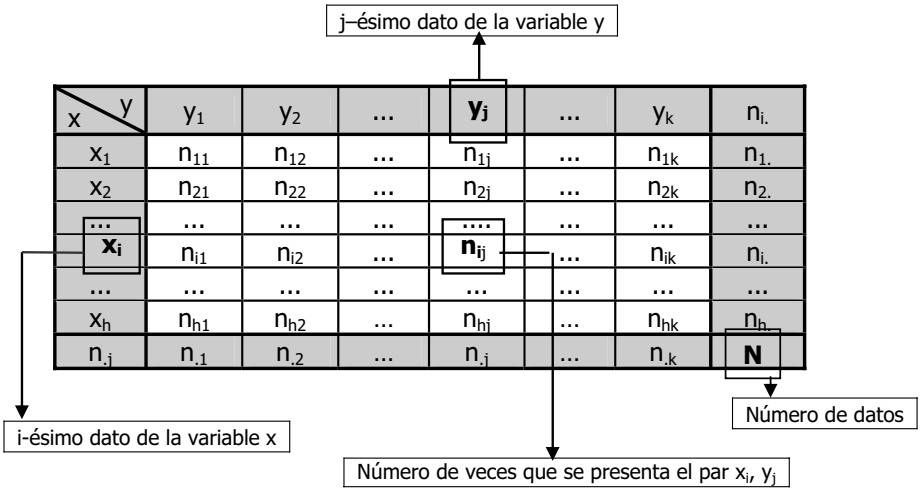
Para utilizar este tipo de técnicas, obviamente, es necesario disponer de algunas observaciones de las dos variables implicadas en la relación. Vamos a ver, en primer lugar, cómo se sintetiza la información correspondiente, mediante las tablas de correlación y contingencia.

5.1.1. Tablas de correlación y contingencia

Supongamos que se desea analizar simultáneamente dos caracteres x e y de una determinada población. En tal caso, la distribución de frecuencias correspondiente se representa como (x_i, y_j, n_{ij}), donde x_i e y_j simbolizan un par de valores cualquiera de las variables x e y, y n_{ij} es la **frecuencia bidimensional** o frecuencia absoluta conjunta, que indica el número de veces que conjuntamente se presenta el par (x_i, y_j).

Habitualmente, los datos se presentan en una tabla de doble entrada similar a la que figura en la página siguiente, que se denomina **tabla de correlación** cuando las observaciones se refieren a variables cuantitativas y **tabla de contingencia** cuando los datos corresponden a variables cualitativas o atributos.

Tabla 5.1. Tabla de correlación o de contingencia



En esta tabla:

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

simboliza el número de veces que la variable x toma el valor x_i independientemente del valor que tome la variable y. Análogamente el número de veces que la variable y toma el valor y_j con independencia del valor que toma la variable x se representa por:

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

y finalmente:

$$N = \sum_{i=1}^{h} n_{i.} = \sum_{j=1}^{k} n_{.j}$$

es el número total de datos disponibles.

Como ejemplos, se presentan a continuación una tabla de correlación y una de contingencia.

La tabla de correlación contiene la información correspondiente a dos caracteres de los hoteles ubicados en una provincia española: x = categoría del establecimiento, que se ha agrupado en dos intervalos de amplitud constante igual a dos, e y = número de plazas, que se ha agrupado en seis intervalos de distinta amplitud.

Tabla 5.2. Tabla de correlación de x e y

x \ y	0 – 30	30 – 60	60 – 90	90 – 200	200 – 900	900 – 3.000	$n_{i.}$
0 – 2	9	5	5	7	7	2	35
2 – 4	11	15	**12**	12	3	2	55
$n_{.j}$	20	20	17	19	10	4	90

En esta provincia existen por tanto 12 hoteles que tienen registradas entre 60 y 90 plazas y una categoría que oscila entre las 2 y las 4 estrellas.

$$n_{1.} = \sum_{j=1}^{6} n_{ij} = 9 + 5 + ... + 2 = 35$$

es el número hoteles que tienen entre 0 y 2 estrellas, independientemente del número de plazas que tienen registradas.

$$n_{2.} = \sum_{j=1}^{6} n_{ij} = 11 + 15 + ... + 2 = 55$$

es el número hoteles que tienen entre 2 y 4 estrellas, independientemente del número de plazas que tienen registradas.

$$n_{.1} = \sum_{i=1}^{2} n_{ij} = 9 + 11 = 20$$

es el número de hoteles que tienen entre 0 y 30 habitaciones independientemente de su categoría.

$$n_{.2} = \sum_{i=1}^{2} n_{ij} = 5 + 15 = 20$$

es el número de hoteles que tienen entre 30 y 60 habitaciones independientemente de su categoría.

(…)

$$n_{.7} = \sum_{i=1}^{2} n_{ij} = 2 + 2 = 4$$

es el número de hoteles que tienen entre 900 y 3.000 habitaciones independientemente de su categoría.

y finalmente:

$$N = \sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{7} n_{.j} = 35 + 55 = 20 + 20 + ... + 4 = 90$$

es el número total de hoteles de la provincia, con cualquier número de plazas y de cualquier categoría.

La tabla de contingencia se refiere a dos caracteres de los visitantes españoles que llegan a Galicia: x = Comunidad Autónoma de residencia, cuyos valores son los de todas las comunidades autónomas españolas excepto Ceuta y Melilla, cuyos datos se han agregado a los de Andalucía, e y = motivo del viaje, para el cual se han considerado los valores más frecuentes: el paisaje, la cultura, la familia y la tranquilidad, y se han agrupado los demás motivos en otros.

Tabla 5.3. Tabla de contingencia de x e y

x \ y	Paisaje	Cultura	Familia	Tranquilidad	Otros	$n_{i.}$
Andalucía	34	24	19	9	11	97
Aragón	18	5	5	4	2	34
Asturias	16	8	23	13	26	86
Baleares	6	3	3	1	13	26
Canarias	12	5	6	1	30	54
Cantabria	5	7	2	4	24	42
Castilla-León	47	13	40	28	198	326
Castilla-La Mancha	9	3	8	9	43	72
Cataluña	70	37	64	15	251	437

x \ y	Paisaje	Cultura	Familia	Tranquilidad	Otros	$n_{i.}$
Extremadura	2	3	2	3	15	25
Galicia	80	30	154	94	720	1.078
Rioja	8	4	2	1	17	32
Madrid	123	66	130	60	515	894
Murcia	8	4	3	1	22	38
Navarra	**7**	3	7	3	23	43
País Vasco	50	16	70	16	202	354
Valencia	57	18	12	5	120	212
$n_{.j}$	552	249	550	267	2.232	3.850

Por tanto, de los visitantes que llegan a Galicia procedentes de Navarra, 7 viajan para ver el paisaje.

$$n_{1.} = \sum_{j=1}^{5} n_{ij} = 34 + 24 + ... + 11 = 97$$

es el número de visitantes que llegan a Galicia procedentes de Andalucía, independientemente del motivo del viaje.

$$n_{2.} = \sum_{j=1}^{5} n_{ij} = 18 + 5 + ... + 2 = 34$$

es el número de visitantes que llegan a Galicia procedentes de Aragón, independientemente del motivo del viaje.

(...)

$$n_{17.} = \sum_{j=1}^{5} n_{ij} = 57 + 18 + ... + 120 = 212$$

es el número de visitantes que llegan a Galicia procedentes de la Comunidad Valenciana, independientemente del motivo del viaje.

$$n_{.1} = \sum_{i=1}^{17} n_{ij} = 34 + 18 + ... + 57 = 552$$

es el número de visitantes que vienen a Galicia para ver el paisaje, independientemente de cuál sea su Comunidad Autónoma de residencia.

$$n_{.2} = \sum_{i=1}^{17} n_{ij} = 24 + 5 + ... + 18 = 249$$

es el número de visitantes que vienen a Galicia para conocer su cultura, independientemente de cuál sea su Comunidad Autónoma de residencia.

(…)

$$n_{.5} = \sum_{i=1}^{17} n_{ij} = 11 + 2 + ... + 120 = 2.232$$

es el número de visitantes que vienen a Galicia por motivos diferentes de los anteriores, independientemente de cuál sea su Comunidad Autónoma de residencia.

y finalmente:

$$N = \sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{7} n_{.j} = 97 + 34 + ... + 212 = 552 + 249 + ... + 2.232 = 3.850$$

es el número total de visitantes que llegan a Galicia por cualquier motivo y procedentes de cualquier Comunidad Autónoma.

5.1.2. Distribuciones marginales

A partir de una distribución bidimensional de frecuencias (x_i, y_j, n_{ij}) puede analizarse cada una de las variables o atributos por separado. Para ello deben deducirse de la tabla de correlación dos distribuciones unidimensionales, que corresponden a las variables o atributos x e y respectivamente, que se denominan **distribuciones marginales**.

La distribución marginal de la variable x (y) queda definida por los valores que toma dicha variable junto con sus correspondientes frecuencias marginales, esto es el

número de veces que la variable x (y) toma cada uno de sus valores con independencia de los que tome la variable y (x), que pueden presentarse de la siguiente forma:

Tabla 5.4. Distribuciones marginales de x e y

x_i	$n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
...	...
x_j	$n_{j.}$
...	...
x_h	$n_{h.}$
	N

y_j	$n_{.j}$
y_1	$n_{.1}$
y_2	$n_{.2}$
...	...
y_j	$n_{.j}$
...	...
y_k	$n_{.k}$
	N

Partiendo de la información que contiene la Tabla 5.2, que se refiere a las variables x = categoría del establecimiento e y = número de plazas de los hoteles ubicados en una provincia española, se han obtenido las dos distribuciones marginales correspondientes.

Tabla 5.5. Distribución marginal de x = categoría del establecimiento

$L_{i-1} - L_i$	$n_{i.}$
0 – 2	35
2 – 4	55
	90

Tabla 5.6. Distribución marginal de y = número de plazas

$L_{i-1} - L_i$	$n_{.j}$
0 – 30	20
30 – 60	20
60 – 90	17
90 – 200	19
200 – 900	10
900 – 3.000	4
	90

Una vez obtenidas las distribuciones marginales, pueden aplicarse para cada variable todas las técnicas de análisis que hemos visto en el caso las distribuciones unidimensionales.

5.1.3. Distribuciones condicionadas

La distribución condicionada de la variable y dado un valor de $x = x_i$, queda definida por los valores que toma la variable y junto con sus correspondientes frecuencias condicionadas. De forma similar, la distribución condicionada de la variable x dado un valor de $y = y_j$, queda definida por los valores que toma la variable y junto con sus correspondientes frecuencias condicionadas.

A la frecuencia condicionada de y_j dado el valor de $x = x_i$ se la designa n (j/i), y se la define como el número de veces que se presenta el valor de $y = y_j$ siendo $x = x_i$. De igual forma, a la frecuencia condicionada de x_i dado el valor de $y = y_j$ se la designa n (i/j) y se la define como el número de veces que se presenta el valor de $x = x_i$ siendo $y = y_j$.

Las distribuciones condicionadas de y a $x = x_i$ y de x a $y = y_j$ pueden presentarse, entonces, de la siguiente forma:

Tabla 5.7. Distribuciones condicionadas de y a $x = x_i$ y de x a $y = y_j$

$y_j/x = x_i$	n (j/i)
y_1	n_{i1}
y_2	n_{i2}
...	...
y_j	n_{ij}
...	...
y_k	n_{ik}
	$n_{i.}$

$x_i/y = y_j$	n (i/j)
x_1	n_{1j}
x_2	n_{2j}
...	...
x_i	n_{ij}
...	...
x_h	n_{hj}
	$n_{.j}$

Siguiendo con el ejemplo anterior, se obtiene, en primer lugar, la distribución de y = número de plazas condicionada a que x = categoría del establecimiento tome el valor $x_i = 2 - 4$; es decir, la distribución del número de plazas de los establecimientos hoteleros condicionada a que dichos establecimientos tengan una categoría de entre 2 y 4 estrellas y, en segundo lugar, la distribución de x = categoría del establecimiento

condicionada a que y = número de plazas tome el valor y_j = 60 – 90; es decir, la distribución de la categoría de los establecimientos hoteleros condicionada a que dichos establecimientos tengan entre 60 y 90 plazas.

Tabla 5.8. Distribución condicionada de y a x = 2 – 4

y_j/x = 2 – 4	n (j/i)
0 – 30	11
30 – 60	15
60 – 90	12
90 – 200	12
200 – 900	3
900 – 3.000	2
	55

Tabla 5.9. Distribución condicionada de x a y = 60 – 90

x_i/y = 60 – 90	n (i/j)
0 – 2	5
2 – 4	12
	17

5.2. Covariación o variación conjunta

Para proceder al análisis de la relación existente entre dos variables, es interesante comenzar examinando el gráfico de los pares de valores observados de las variables, que se construye sobre un par de ejes cartesianos en los que se sitúan escalas para las dos variables.

Generalmente se mide la variable x sobre el eje horizontal y la variable y sobre el eje vertical. En el plano, se dibuja un punto por cada par de valores observados de x y de y. La representación gráfica del conjunto de puntos resultante o nube de puntos se denomina **diagrama de dispersión**, y su análisis visual constituye un buen punto de partida, puesto que proporciona una amplia información sobre la covariación y sus características.

Cuando la nube de puntos se concentra alrededor de una línea recta, se dice que la asociación entre x e y es lineal.

Si la asociación es lineal y la nube de puntos es creciente; es decir, si los valores pequeños de x están asociados a valores pequeños de y y los valores elevados de x están asociados a valores elevados de y, se dice que la asociación entre las variables es lineal y directa. En cambio, si la asociación es lineal y la nube de puntos es decreciente; es decir, si los valores pequeños de x están asociados a valores elevados de y y los valores elevados de x están asociados a valores pequeños de y, se dice que la asociación entre las variables es lineal e inversa.

Puede ocurrir, sin embargo, que el diagrama de dispersión no sugiera una recta, sino que los puntos pueden aparecer aproximadamente dispuestos sobre una curva. En tal caso, la asociación entre las variables es no lineal.

Finalmente, también es posible que la nube de puntos indique la no existencia de asociación entre las variables.

Los gráficos 5.1 a 5.4 son los diagramas de dispersión correspondientes a varias situaciones comunes: fuerte asociación lineal positiva, asociación lineal negativa más débil, asociación no lineal y ausencia de asociación.

En el gráfico 5.1, la nube de puntos es el diagrama de dispersión de las variables y = pernoctaciones en establecimientos hoteleros de viajeros procedentes de Alemania y x = número de viajeros procedentes de Alemania. Este gráfico muestra la existencia de una clara relación lineal (en forma de línea recta) entre ellas. Además, la relación es positiva o directa (cuando la variable x aumenta la variable y aumenta también) y los puntos del diagrama aparecen muy concentrados, lo que indica que la relación entre las variables es muy intensa.

El gráfico 5.2 es el diagrama de dispersión correspondiente a las variables y = pernoctaciones en establecimientos hoteleros de viajeros procedentes de Alemania y x = tipo de cambio en marcos por peseta. En él puede observarse la existencia de una relación lineal (en forma de línea recta), pero mucho menos clara que en el caso anterior, puesto que la nube de puntos está más dispersa. Además, la relación es de tipo inverso o negativa (cuando la variable x aumenta la variable y disminuye y viceversa).

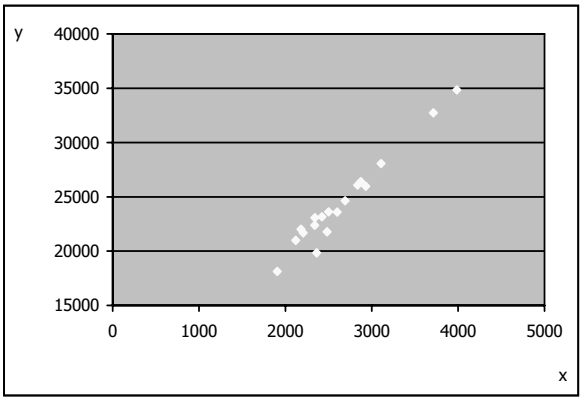


Gráfico 5.1. Asociación lineal positiva

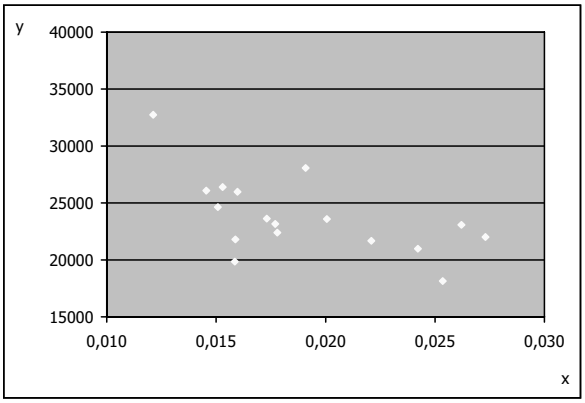


Gráfico 5.2. Asociación lineal negativa

El gráfico 5.3 es el diagrama de dispersión de las variables y = pernoctaciones en establecimientos hoteleros de viajeros procedentes de Alemania y x = tipo de cambio en pesetas por unidad de cuenta europea, y sugiere la existencia de una relación entre las variables de tipo no lineal o curvilínea.

Finalmente, el gráfico 5.4 es el diagrama de dispersión que corresponde a las variables y = pernoctaciones en establecimientos hoteleros de viajeros procedentes de Alemania y x = pernoctaciones en establecimientos hoteleros de viajeros procedentes de Francia, y muestra la inexistencia de relación entre las variables, ya que no parece que las variaciones de x supongan variaciones sistemáticas en y.

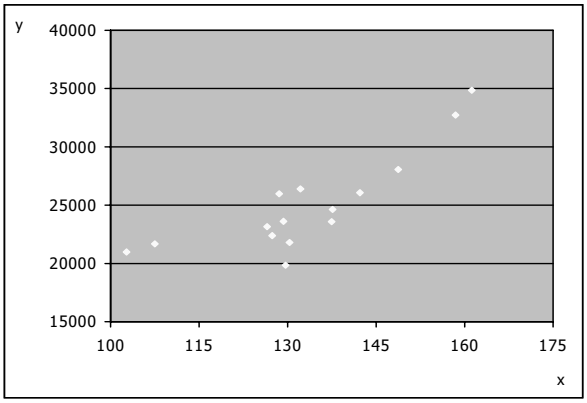


Gráfico 5.3. Asociación no lineal

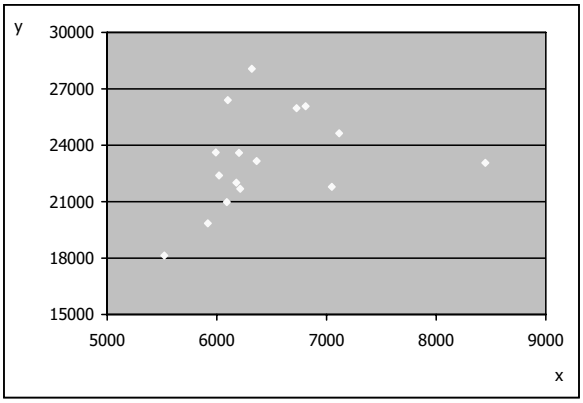


Gráfico 5.4. Ausencia de asociación

El método gráfico de análisis de la covariación sirve no sólo para detectar la posible relación entre dos variables sino también para conocer sus características. El procedimiento es, en el caso de dos variables, muy eficaz y, además, sumamente sencillo. Pero, aunque el diagrama de dispersión ofrece información acerca de la asociación entre dos variables, basada en la observación visual, en la práctica se acude a instrumentos más precisos que permiten la obtención de una cuantificación numérica de la intensidad y las características de dicha asociación.

La **covarianza** es uno de estos instrumentos. Se basa en las diferencias observadas entre los valores de x e y respecto a sus correspondientes medias.

Así, para obtener su valor, se calcula para cada par (x_i, y_j) de valores observados de x e y el producto de las desviaciones respecto a sus medias, que dejan por debajo de ellas los valores menores y por encima de ellas los valores mayores de las variables, de tal manera que para los valores pequeños de las variables estas diferencias son negativas y para los valores elevados son positivas.

El signo y la magnitud de los productos $(x_i - \bar{x})\ (y_j - \bar{y})$ obtenidos para todos los valores de i y de j determinan el valor de la covarianza.

Cuando la relación entre x e y es lineal y creciente, a los valores más pequeños de x ($x_i < \bar{x}$) les corresponden los valores más pequeños de y ($y_j < \bar{y}$) y a los valores más elevados de x ($x_i > \bar{x}$) les corresponden los valores más elevados de y ($y_j > \bar{y}$), de tal manera que la mayoría de estos productos son positivos y es positiva su suma. Además, cuanto más intensa sea la relación, mayor será la proporción de productos positivos en el total y, por tanto, la suma de los productos es elevada.

En cambio, si la relación es lineal y decreciente, a los valores más pequeños de x ($x_i < \bar{x}$) les corresponden los valores más elevados de y ($y_j > \bar{y}$) y a los valores más elevados de x ($x_i > \bar{x}$) les corresponden los valores más pequeños de y ($y_j < \bar{y}$), de tal manera que la mayoría de estos productos son negativos y su suma es negativa. Además, cuanto más intensa sea la relación, mayor será la proporción de productos negativos en el total y, por tanto, la suma de los productos es negativa, pero elevada en valor absoluto.

Cuando no existe relación entre las variables no se observa una respuesta sistemática de los valores de y a los valores de x, de tal manera que los productos de las desviaciones de las variables respecto a sus medias son tanto positivos como negativos y la suma de dichos productos es pequeña.

En definitiva, si la suma de los productos de las desviaciones de x respecto a su media por las desviaciones de y respecto a su media es positiva, las variables mantienen una relación directa, y si es negativa, las variables mantienen una relación inversa. Si las variables no tienen una relación de tipo lineal, la suma es igual a cero.

A partir de esta idea, la covarianza se define como la media aritmética de los productos de las desviaciones de las variables x e y respecto a sus respectivas medias; es decir, se define por cociente entre la suma de los productos de las variables en desviaciones respecto a sus medias y el número de observaciones:

$$s_{xy} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \overline{x})(y_j - \overline{y})\,n_{ij}}{N}$$

que de una forma más abreviada puede calcularse a través de la expresión:

$$s_{xy} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}x_i y_j n_{ij}}{N} - \overline{xy}$$

Este cociente es positivo si es positiva la suma y negativo cuando la suma es negativa, y es igual a cero si y sólo si es igual a cero la suma. Luego, el signo de la covarianza indica que la relación entre las variables es directa si es positivo, que es inversa si es negativo y que no existe asociación lineal si es nula.

Tabla 5.10. Tabla de correlación de x = gastos e y = beneficios

		y = Beneficios		
x = Gastos		250 – 350	350 – 450	$n_{i.}$
	10 – 15	4	5	9
	15 – 30	6	10	16
	$n_{.j}$	10	15	25

Con los datos que contiene esta tabla se calcula, como ejemplo, la covarianza entre las variables beneficios y gastos de las empresas turísticas ubicadas en una determinada comarca. Debe tenerse en cuenta que, al tratarse de una distribución agrupada en intervalos, tanto para el cálculo de las medias aritméticas, como para el de la covarianza, deben utilizarse las marcas de clase.

$$\frac{\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij}}{N} =$$

$$= \frac{(12{,}5 \times 300 \times 4) + (12{,}5 \times 400 \times 5) + (22{,}5 \times 300 \times 6) + (22{,}5 \times 400 \times 10)}{25} =$$

$$= \frac{170.500}{25}$$

$$\overline{x} = \frac{\sum_{i=1}^{h} x_i n_{i.}}{N} = \frac{12{,}5 \times 9 + 22{,}5 \times 16}{25} = 18{,}9$$

$$\overline{y} = \frac{\sum_{j=1}^{k} y_j n_{.j}}{N} = \frac{(300 \times 10 + 400 \times 15)}{25} = 360$$

Y la covarianza resulta:

$$s_{xy} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij}}{N} - \overline{xy} = \frac{170.500}{25} - 18{,}9 \times 360 = 16$$

Luego entre los beneficios y los gastos de las empresas existe una asociación lineal de tipo directo.

5.3. Correlación

Dada su definición, la covarianza está expresada en el producto de las unidades de medida de x e y, y no tiene límites inferior o superior. Por eso es necesario definir otra medida que no se vea afectada por problemas de escala o de comparabilidad. La medida adimensional más utilizada para cuantificar el grado de asociación lineal entre dos variables es el **coeficiente de correlación lineal simple**, que es el resultado de dividir la covarianza por el producto de las desviaciones típicas de x e y.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

El coeficiente de correlación lineal es una medida adimensional de la intensidad de la asociación lineal entre x e y que toma valores comprendidos entre – 1 y 1, porque la covarianza de las variables puede ser positiva o negativa, pero en valor absoluto es menor o igual que el producto de las desviaciones típicas. Cuando está próximo a sus valores extremos indica fuerte asociación lineal, que es perfecta si es exactamente igual a uno o a menos uno (positiva, si es positivo; negativa, si es negativo). Cuando está próximo a cero, indica que la asociación lineal entre las variables es débil, siendo nula cuando es exactamente igual a cero.

De la definición del coeficiente, se deduce inmediatamente que $r_{xy} = r_{yx}$; es decir, que el grado de asociación lineal entre y y x es el mismo que entre x e y. Además, $r_{xx} = r_{yy} = 1$; es decir, que cualquier variable tiene una correlación positiva perfecta consigo misma.

Siguiendo con el ejemplo anterior y dados los valores de las desviaciones típicas $s_x = 4{,}80$ y $s_y = 48{,}99$ se ha calculado el valor del coeficiente de correlación:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{16}{4{,}80 \times 48{,}99} = 0{,}07$$

resultado que indica que prácticamente no existe relación lineal entre los beneficios y los gastos de las empresas turísticas de la comarca seleccionada.

5.4. Correlación con EXCEL

Para obtener el valor del coeficiente de correlación lineal simple con EXCEL efectuando los cálculos intermedios, en primer lugar, se introducen los datos correspondientes a las dos variables y se calculan sus medias mediante la instrucción Insertar–Función–Estadísticas–Promedio.

Para obtener las diferencias entre y y su media y entre x y su media se procede de la misma forma: en la primera celda de la columna donde se desea que aparezca el resultado, se teclea el signo igual y con el ratón se pincha la primera celda de y, se escribe el signo menos, y se pincha con el ratón el valor de la media. Para dejar fijo en la diferencia este valor, se escribe el símbolo $ entre el nombre de la columna y el de la

fila. Se ejecuta la operación pulsando Intro. Se pincha esta celda con el ratón y se copia. Se pega la operación en el resto de las celdas de la columna.

Para obtener los productos de las diferencias de y respecto a su media y de x respecto a su media, en la primera celda de la columna donde se desea que aparezca el resultado, se teclea el signo igual y con el ratón se pincha la primera celda de y menos su media, se escribe el signo por, y se pincha con el ratón la primera celda de x menos su media. Se ejecuta la operación pulsando Intro. Se pincha esta celda con el ratón y se copia. Se pega la operación en el resto de las celdas de la columna.

Para obtener los cuadrados de las diferencias, en la primera celda en la que se desea que aparezca el resultado se teclea el signo igual, y con el ratón se pincha la primera celda de y (x) menos su media y se escribe el símbolo ^ (que indica potencia) seguido de 2. Se ejecuta la operación pulsando Intro. Se pincha esta celda con el ratón y se copia. Se pega la operación en el resto de las celdas de la columna.

Con Insertar–Función–Estadísticas–Promedio se obtienen los valores medios de las diferencias de y respecto a su media por x respecto a su media (o la covarianza de x e y), de los cuadrados de las diferencias de y respecto a su media (o varianza de y) y de los cuadrados de las diferencias de x respecto a su media (o varianza de x). Con Insertar–Función–Matemáticas–Raíz, se obtienen las desviaciones típicas.

Finalmente, con Inserta–Función–Matemáticas–Cociente, se obtiene el valor del coeficiente de correlación lineal simple entre las dos variables.

Obviamente, es mucho más sencillo ejecutar la instrucción **Inserta–Función–Estadísticas–Coeficiente de Correlación**, que proporciona el resultado de forma inmediata, aunque no da los resultados intermedios.

TEMA 6
ANÁLISIS DE REGRESIÓN

6.1. Introducción

La regresión es otro instrumento estadístico de análisis de la relación entre variables. En general, puede utilizarse para el análisis de la relación que una variable mantiene con un conjunto de otras k, pero en este texto nos referimos únicamente a la **regresión lineal simple**.

El término **simple** indica que en la relación sólo hay dos variables implicadas; es decir, sólo vamos a ocuparnos del caso bivariante. El término **lineal** hace referencia a la forma que adopta la relación que, en principio, puede expresarse mediante cualquier función matemática: lineal, polinómica, parabólica, exponencial, etcétera. Cuando la regresión es lineal, la relación se formaliza mediante una función lineal, lo que en términos gráficos equivale a que la nube de puntos del diagrama de dispersión que representa las observaciones de las variables esté, aproximadamente, sobre una línea recta. De hecho, lo que se pretende con la regresión lineal es encontrar la ecuación de la recta sobre la que aproximadamente se alinean los puntos de la nube.

En cuanto al término **regresión**, fue utilizado por Galton en un estudio relativo a las estaturas de una determinada población, con el objetivo de analizar si existía alguna relación entre las estaturas de los padres y las de los hijos. En dicho estudio se observó que los hijos de padres altos eran, por término medio, más altos que la media de la población, pero más bajos que sus padres. Igualmente, los hijos de padres bajos eran, por término medio, más bajos que la media de la población pero más altos que sus padres. Había, por tanto, según Galton, una tendencia de las estaturas de los hijos, tanto de padres altos como de padres bajos, a moverse, a volver o a "regresar" a la estatura media de la población.

En la actualidad, la palabra abarca un concepto más amplio. El análisis de regresión es el análisis de una dependencia **causal**, **unilateral**, **inexacta** o de tipo estadístico. Por tanto,

1. El análisis de regresión lineal simple se utiliza para analizar la relación entre dos variables bajo la hipótesis de que es causal; es decir, debe suponerse que una de las variables (x) es la causa de la otra (y), que es el efecto. Obviamente, la técnica es aplicable aunque la relación entre las variables sea no causal, pero no tiene sentido y puede conducir a conclusiones erróneas. La variable que se supone dependiente (y) se denomina también explicada, endógena o **regresando**. La variable que se supone independiente (x) se denomina también explicativa, exógena o **regresor**.

2. La relación entre las variables se supone unilateral, tiene un solo sentido; x es causa de y pero y no es causa de x; por tanto, no existe interdependencia.

3. La relación entre las variables es inexacta o de tipo estadístico. Como ya hemos visto a partir de los diagramas de dispersión, aunque las variables x e y estén muy relacionadas, si representamos en el plano pares de valores observados de x y de y, no encontramos una línea sino una nube de puntos que puede estar más o menos concentrada alrededor de una línea "ideal". Aunque la nube de puntos puede estar muy concentrada, por ejemplo, alrededor de una recta, no todos los puntos están sobre la recta. Luego, las observaciones de x y de y no se ajustan a una ley matemática, que asigna a cada valor de x un valor dado de y, sino que la relación que existe entre ellas es de tipo estocástico, existe un conjunto de posibles valores de y para cada valor dado de x.

Para centrar la idea, analizamos un ejemplo muy sencillo:

Una hipótesis razonable podría ser que el gasto en turismo de una familia (y) depende de la renta que percibe (x), de tal manera que responde a las variaciones de la renta con variaciones del mismo sentido. Estas variables mantienen, por tanto, una relación directa: el gasto aumenta si aumenta la renta y disminuye cuando la renta disminuye; es decir, en términos matemáticos, el gasto es una función creciente de la renta:

$$y = f(x)$$

O bien, suponiendo que f es una función lineal:

$$y = \alpha + \beta x$$

La hipótesis establecida respecto al comportamiento del gasto es una ley general que debería cumplirse para cualquier familia. Concretando la relación para una familia determinada, se tiene:

$$y_t = \alpha + \beta x_t$$

donde t toma valores comprendidos entre 1 y N, siendo N el número total de familias en el que estamos interesados.

En esta ecuación, entonces, y_t es el gasto en turismo correspondiente a la t–ésima familia y x_t es su renta. Además, al ser f una función creciente, el coeficiente β es positivo.

Dicho de otra forma: de acuerdo con esta igualdad, la única causa de que y varíe es que varíe x, y, como hemos dicho, las variaciones positivas (negativas) de x deben suponer variaciones positivas (negativas) en y.

Teniendo en cuenta que:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta x_t - (\alpha + \beta x_{t-1}) = \beta x_t - \beta x_{t-1} = \beta (x_t - x_{t-1}) = \beta \Delta x_t$$

Se deduce que la variación de y es igual al producto β por la variación de x, y para que las dos variaciones tengan el mismo signo, β debe ser mayor que cero.

La representación gráfica de la ecuación $y_t = \alpha + \beta x_t$ es una recta tal como la que se muestra en el Gráfico 6.1, que corta al eje de ordenadas a la altura de α y que tiene β como pendiente, ya que:

$x_t = 0 \Rightarrow y_t = \alpha$, luego la recta pasa por el punto $(0, \alpha)$

$m = \Delta y_t / \Delta x_t = \beta$ siendo $\beta > 0$, luego la recta está inclinada a la derecha y el ángulo que forma con una paralela al eje de abscisas es igual a β.

La ecuación $y_t = \alpha + \beta x_t$ supone la existencia de una relación exacta entre el gasto y la renta, puesto que, de acuerdo con ella, dos familias que tienen la misma

renta tienen exactamente el mismo gasto en turismo. Como vemos, fijado un valor de x_t a y_t le corresponde un único valor que viene dado por $\alpha + \beta x_t$. Esto significa, por ejemplo, que para un conjunto de familias con 100 unidades de renta, el gasto en turismo ha de ser, para todas ellas, exactamente igual a $\alpha + 100\beta$.

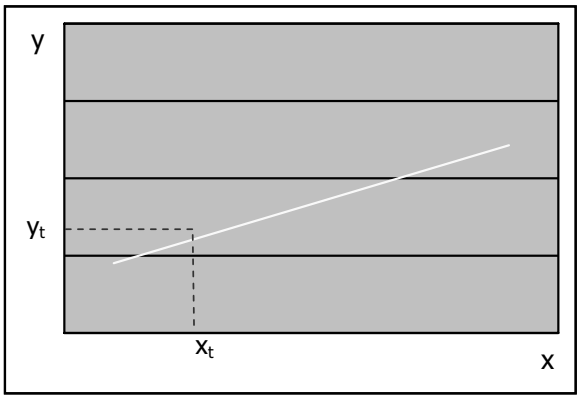


Gráfico 6.1. Representación gráfica de la ecuación $y_t = \alpha + \beta x_t$

Sabemos, sin embargo, que esto no sucede. Dos familias con la misma renta, pueden tener gastos turísticos diferentes. Por ejemplo, en un caso práctico de muestreo de 10 familias con renta igual a 100 unidades, en cuanto a su gasto en turismo, se obtuvieron como respuestas los siguientes valores ordenados: {0,00; 5,60; 5,65; 5,90; 6,10; 6,10; 6,20; 6,25; 6,35; 10,00} unidades.

Como vemos, no todas las familias encuestadas, que tienen una renta exactamente igual, tienen un gasto en turismo igual. En esta muestra, las familias declaran gastos comprendidos entre 0,00 y 10,00 unidades, que oscilan en torno a una media de 5,81 unidades. En la mayoría de los casos, los gastos son similares a la media, pero no iguales. Lo mismo sucedería si el proceso de muestreo se repitiese para un conjunto de familias con renta de 200, 300, 400 unidades, o cualquier otra cantidad.

Hay que tener en cuenta que, además de la renta, existen otras variables, tales como el número de miembros de la familia, sus circunstancias particulares, sus aficiones o sus gustos, que afectan también al gasto en turismo, de tal manera que, para una renta dada, el gasto no suele ser el mismo.

La existencia de esas variables omitidas es lo que hace que las relaciones entre las variables no sean exactas, sino aproximadas. La hipótesis establecida inicialmente es que el gasto en turismo depende de la renta. Pero, en la práctica, el gasto depende, además, de otras variables que tienen pequeños efectos sobre él, de tal manera que el gasto, aunque es función de la renta, no es función exacta de la renta, sino que la relación entre las variables es inexacta.

Esto significa que la representación gráfica de la verdadera relación existente entre el gasto y la renta no es una recta, sino una nube de puntos como la que muestra el Gráfico 6.2, en la que puede observarse como, dado un valor de la renta, el valor del gasto no es único, sino que existe un conjunto de diferentes valores del gasto que son compatibles con él.

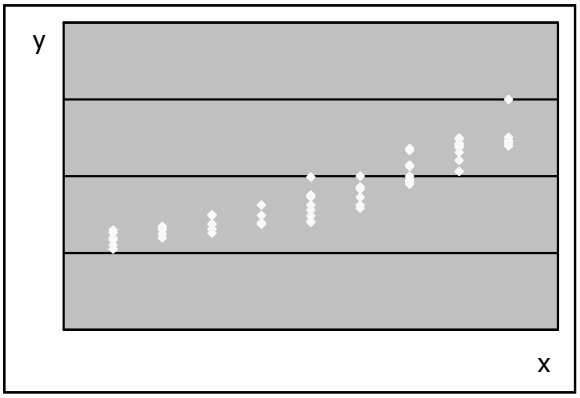


Gráfico 6.2. Conjunto de posibles valores de y fijado x

Lo que sucede es que, generalmente, en la práctica, de esos puntos de la nube sólo son conocidos algunos. Más aún, habitualmente, para cada valor observado de x se dispone de un solo valor observado para y. Los pares de valores observados de x e y son los que se representan en el diagrama de dispersión al que corresponde el Gráfico 6.3.

El tratamiento de los datos, de las observaciones disponibles para las variables, con determinadas técnicas, permite deducir unos valores numéricos a y b que estiman, de forma más o menos correcta, los coeficientes α y β que incluye la ecuación del modelo, que pueden utilizarse para representar la recta estimada, que también vemos

en el gráfico. Estos valores numéricos son, únicamente, estimaciones de los coeficientes teóricos, pero no son sus verdaderos valores que, en realidad, siguen siendo desconocidos.

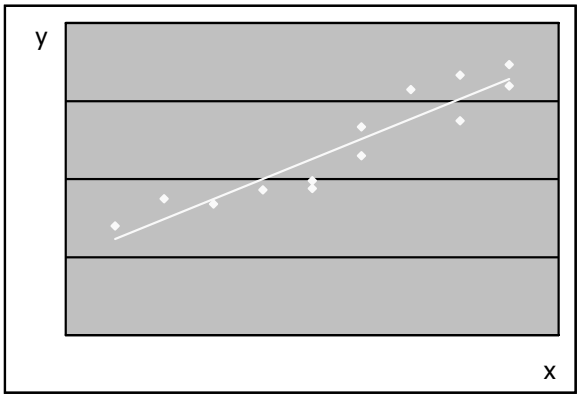


Gráfico 6.3. Diagrama de dispersión de x e y

Pero una vez que disponemos de estimaciones de los coeficientes, podemos analizar si los datos estadísticos de los que disponemos confirman la teoría propuesta; es decir, en este caso, si el gasto varía cuando varía la renta y si sus variaciones son del mismo sentido. El hecho de que las estimaciones obtenidas tengan los signos que teóricamente cabía esperar que tuviesen constituye un indicio de que la teoría es correcta. Los datos parecen confirmarla.

Además, con dichos coeficientes, se puede utilizar la ecuación para estimar o predecir el valor de la variable que se considera dependiente que, en este caso, es el gasto, en función de unos valores dados o conocidos de la variable independiente que es, en este caso, la renta.

Por otra parte, si la información muestral disponible no es compatible con la hipótesis establecida inicialmente, habrá que proceder a su revisión o a la elaboración de una nueva teoría, con lo que volveríamos al punto de partida.

En resumen, en la aplicación práctica del análisis de regresión, se establece un supuesto sobre el comportamiento de las variables, que indica qué variable puede resultar relevante para explicar a otra y si la influencia que tiene sobre ella es positiva

o negativa. Muy pocas veces, sin embargo, se conoce la forma funcional de la relación entre las variables por lo que, generalmente, al menos como primera aproximación, en la fase de planteamiento del modelo, suele elegirse la forma funcional más sencilla posible. Así, la inmensa mayoría de las ecuaciones son lineales o fácilmente linealizables.

Una vez planteada la ecuación, el proceso de estimación consiste en la obtención de valores numéricos representativos de los coeficientes que incluye.

Para proceder con esta etapa, es necesario disponer de datos; es decir, de un conjunto de observaciones de las variables que figuran en el modelo que, además, deben estar perfectamente definidas desde el punto de vista estadístico. En algunos casos puede ser necesario también efectuar algún tratamiento previo: deflactar las series, eliminar la tendencia o la estacionalidad, obtener tasas, porcentajes, logaritmos, etcétera. Puede ser necesario también, por la limitación de los datos disponibles, utilizar lo que denominamos variables "proxy" o aproximaciones a aquellas que serían teóricamente las más adecuadas.

En cualquier caso, la necesidad de disponer de datos para la estimación de estos modelos no significa que no se pueda incorporar en ellos información de tipo cualitativo. La única condición necesaria para que esto se pueda hacer es que, de alguna forma, la información cualitativa se pueda cuantificar. Generalmente, los atributos se cuantifican mediante la utilización de unas variables especiales, que llamamos ficticias, que toman únicamente los valores cero y uno, cero para indicar la ausencia de una cualidad y uno para indicar su presencia.

Una vez estimada la ecuación, se trata de analizar si los datos disponibles confirman la teoría de la que habíamos partido o si por el contrario dicha información no es compatible con el supuesto inicial.

En el caso de que la ecuación pase la prueba, podrá ser utilizada para la predicción y, por tanto, como instrumento de política económica. Si la ecuación no supera el contraste, el proceso debe ser revisado en todos sus puntos: ¿existe alguna razón para pensar que la teoría no es correcta?, ¿está la teoría correctamente representada con la ecuación planteada?, ¿los datos utilizados son los adecuados?,

¿están dados en las unidades correctas?, ¿el método o técnica utilizada para obtener los valores numéricos representativos de los coeficientes es la apropiada?, ¿se ha interpretado adecuadamente el resultado obtenido?, etcétera. Una vez efectuadas las correcciones oportunas, el modelo puede utilizarse para analizar la realidad, para obtener predicciones o como base para tomar decisiones.

6.2. El ajuste mínimo cuadrático ordinario

El problema que se plantea ahora es el de la obtención práctica de la recta de regresión. Como hemos visto en el ejemplo, en el análisis de regresión se supone que el regresando (y) y el regresor (x) mantienen una relación de dependencia que se representa formalmente mediante una ecuación tal como:

$$y_t = \alpha + \beta x_t$$

De acuerdo con esta igualdad, son los cambios en x los que explican los cambios en y. Normalmente se tiene también alguna hipótesis sobre cómo afectan al regresando los cambios en el regresor. Es posible que sepamos, por ejemplo, que entre ambas variables existe una relación directa (inversa), de manera que las variaciones en el regresor provocan en el regresando variaciones del mismo sentido (de sentido contrario). Pero se desconoce la magnitud del cambio, que es lo que se intenta deducir mediante el proceso de estimación del modelo.

Dicho proceso proporciona unos valores numéricos, a y b, que pueden considerarse representativos de los coeficientes desconocidos α y β, de manera que se cuantifica la relación que existe entre las variables. Así, puede obtenerse una aproximación o una estimación del valor de y, que representamos como $\hat{y}_t$, mediante el producto $a + bx_t$, de manera que se tiene:

$$\hat{y}_t = a + bx_t$$

que es la ecuación que estima la que expresa formalmente la relación entre el regresando y el regresor.

En términos gráficos, la estimación proporciona una aproximación de la recta "ideal" o teórica en torno a la cual se concentran los puntos del diagrama de dispersión, que denominamos recta muestral o estimada.

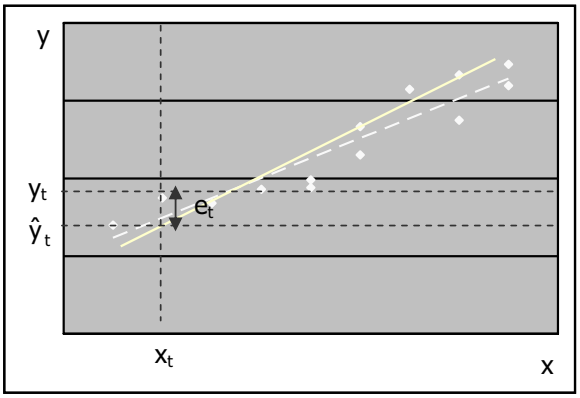


Gráfico 6.4. Diagrama de dispersión y rectas de regresión "ideal" o teórica (trazo discontinuo) y estimada (trazo continuo)

Para obtener esta recta, pueden utilizarse diferentes criterios. Uno de ellos, muy sencillo y ampliamente utilizado, es el que sugiere que los estimadores más adecuados son los que generan valores estimados del regresando lo más próximos posible a los observados.

Si llamamos **error** o **residuo** a la diferencia entre el valor observado y el estimado del regresando:

$$e_t = y_t - \hat{y}_t$$

los valores estimados del regresando son tanto más similares a los observados cuanto menor es el error, de manera que, de acuerdo con el criterio establecido, los estimadores se obtienen minimizando el error cometido; pero, puesto que t = 1, 2, …, N, debe tenerse en cuenta que en la ecuación hay tantos errores implicados como observaciones muestrales.

Puede minimizarse el conjunto de errores haciendo mínima su suma. Pero el problema de suma mínima no tiene solución única. Además, en la adición, los errores

de distinto signo se compensan entre sí, de manera que errores grandes con signos contrarios, generan sumas pequeñas y, por tanto, una suma mínima puede encubrir errores elevados.

Para evitar la compensación de los errores de distinto signo pueden tomarse los valores absolutos o los cuadrados. La suma de valores absolutos no es una función derivable, de manera que es más sencillo obtener el mínimo de la suma de los cuadrados de los errores. Este procedimiento tiene, además, la ventaja de que penaliza los errores elevados, puesto que si un error es el doble de otro, su cuadrado es cuatro veces mayor.

La estimación se efectúa, por tanto, mediante el **método de estimación mínimo–cuadrático ordinaria**, que consiste en la obtención de los valores de los coeficientes a y b que hacen mínima la suma de los cuadrados de los errores:

$$SCE = e_1^2 + e_2^2 + ... + e_N^2 = \sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{N} (y_t - (a + bx_t))^2$$

donde, para simplificar, se supone que, como es habitual, la frecuencia absoluta conjunta para todos los pares de valores (x_t, y_t) es igual a la unidad.

Como sabemos, la primera condición para el mínimo de una función es que sea igual a cero su primera derivada. Por tanto, para obtener los valores de a y b que hacen mínima SCE se calculan las correspondientes derivadas parciales:

$$\frac{\delta SCE}{\delta a} = -2\sum_{t=1}^{N} (y_t - (a + bx_t))$$

$$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N} (y_t - (a + bx_t))x_t$$

y se igualan a cero:

$$\sum_{t=1}^{N} (y_t - (a + bx_t)) = 0$$

$$\sum_{t=1}^{N} (y_t - (a + bx_t))x_t = 0$$

Operando en las dos igualdades anteriores:

$$\sum_{t=1}^{N} y_t = Na + b\sum_{t=1}^{N} x_t$$

$$\sum_{t=1}^{N} y_t x_t = a\sum_{t=1}^{N} x_t + b\sum_{t=1}^{N} x_t^2$$

Y los valores de a y b se obtienen resolviendo este sistema.

De la primera ecuación se deduce que:

$$\bar{y} = a + b\bar{x}$$

lo cual implica que:

$$a = \bar{y} - b\bar{x}$$

Sustituyendo en la segunda:

$$\sum_{t=1}^{N} y_t x_t = (\bar{y} - b\bar{x})N\bar{x} + b\sum_{t=1}^{N} x_t^2 = N\overline{xy} - bN\bar{x}^2 + b\sum_{t=1}^{N} x_t^2$$

de manera que:

$$\sum_{t=1}^{N} y_t x_t - N\overline{xy} = b(\sum_{t=1}^{N} x_t^2 - N\bar{x}^2)$$

es decir:

$$b = \frac{\sum_{t=1}^{N} y_t x_t - N\overline{xy}}{\sum_{t=1}^{N} x_t^2 - N\bar{x}^2} = \frac{s_{xy}}{s_x^2}$$

En cuanto a la interpretación de estos coeficientes:

El término independiente a es la ordenada en el origen de la recta estimada; es decir, es el valor estimado de y cuando x se anula, puesto que al ser $\hat{y}_t = a + bx_t$, si x_t es igual a cero, $\hat{y}_t$ es igual a a.

Desde el punto de vista económico, la interpretación de este coeficiente no tiene sentido cuando x es una variable que no se anula (población, producto interior bruto, precio, salario, etcétera). Por ejemplo, si estamos analizando los ingresos por turismo en función del número de turistas que entran en el país, a es el valor estimado de los ingresos si en el país no entra ningún turista, pero esta estimación no tiene, desde el punto de vista económico, ningún interés, porque esto es algo que no sucede.

El término independiente en la ecuación de la recta se introduce para evitar la fuerte restricción de que pase por el origen de coordenadas y, por tanto, en muchas ocasiones, aunque no en todas, carece de significado.

El coeficiente b es la pendiente de la recta de regresión estimada; por tanto, es la variación que se estima que se produce en el regresando cuando el regresor experimenta una variación unitaria.

Si b tiene signo positivo indica una relación directa (puesto que la covarianza de las variables es positiva), luego las variaciones de las dos variables son del mismo sentido; y si tiene signo negativo indica una relación inversa (puesto que la covarianza de las variables es negativa), por lo que las variaciones de las dos variables son de sentido contrario.

En el ejemplo propuesto para explicar el significado de a, b sería la variación estimada en los ingresos por turismo que se debe a una variación unitaria en el número de turistas. En este caso, el coeficiente b debería ser de signo positivo, puesto que se espera que a medida que aumenta el número de turistas aumenten también los ingresos por turismo.

Veamos una aplicación práctica, analizando la relación existente entre el gasto anual en turismo (y) y los ingresos familiares mensuales (x). Para ello, se dispone de la información que figura en la Tabla 6.1, en la que ambas variables vienen dadas en euros.

Tabla 6.1. Gasto anual en turismo e ingreso familiar mensual

Familia	Gasto (y)	Ingreso (x)
1	250	1.320
2	400	1.500
3	650	1.880
4	900	2.300
5	1.100	2.520
6	1.200	2.760
7	1.300	2.900
8	1.400	3.100
9	1.600	3.400
10	1.900	3.700

Si se efectúa la representación gráfica de los pares de valores observados de x e y que proporciona la muestra, se obtiene la nube de puntos del Gráfico 6.5, que está casi sobre una recta "ideal", cuya ecuación puede estimarse utilizando el procedimiento de mínimos cuadrados ordinarios.

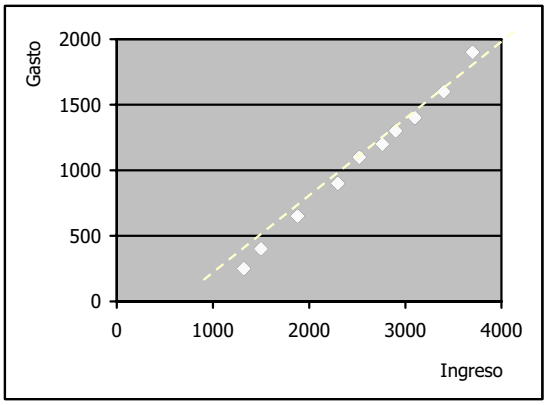


Gráfico 6.5. Diagrama de dispersión del gasto y el ingreso

Dado que es razonable suponer que el gasto turismo mantiene con el ingreso mensual una relación causal, unilateral y no exacta, se plantea el modelo de regresión lineal simple:

$$y_t = \alpha + \beta x_t$$

ecuación en la cual y es la variable dependiente, explicada o regresando, que representa el gasto en turismo y x es la variable independiente, explicativa o regresor, que representa los ingresos mensuales familiares.

Los valores a y b que estiman los coeficientes α y β minimizando la suma de los cuadrados de los errores son, como hemos visto:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{s_{xy}}{s_x^2}$$

Para obtener sus valores, por tanto, es necesario calcular previamente las medias aritméticas de x ($\bar{x} = 2.538$) y de y ($\bar{y} = 1.070$), la covarianza de x e y ($s_{xy} = 372.940$) y la varianza de x ($s_x^2 = 564.036$).

Dados estos datos, sustituyendo, se tiene:

$$b = \frac{s_{xy}}{s_x^2} = \frac{372.940}{564.036} = 0{,}661$$

$$a = \bar{y} - b\,\bar{x} = 1.070 - 0{,}661 \text{ x } 2.538 = -608{,}123$$

Y la recta de regresión estimada que resulta es:

$$\hat{y}_t = -608{,}123 + 0{,}661\ x_t$$

La ordenada en el origen de esta recta es igual a –608,123 e indica que si el ingreso familiar mensual fuese igual a cero, se estimaría un gasto anual en turismo de –608,123 euros. Obviamente, este resultado carece de sentido, puesto que el valor mínimo que cualquier gasto puede tomar es cero. Igualando a cero el gasto anual en turismo estimado:

$$0 = -608{,}123 + 0{,}661\ x_t$$

se deduce:

$$x_t = \frac{608{,}123}{0{,}661} = 920{,}004$$

Es decir, que el gasto anual en turismo estimado es igual a cero para un valor del ingreso familiar mensual igual a 920 euros y es positivo para valores superiores. Por tanto, se estima que sólo las familias con ingresos superiores a éstos gastan en turismo.

La pendiente de la recta de regresión estimada es igual a 0,661 e indica que por cada euro de variación en el ingreso familiar mensual se estima una variación en el mismo sentido del gasto anual en turismo de 0,661 euros. Es decir, que si el ingreso familiar mensual aumenta (disminuye) un euro, se estima que el gasto anual en turismo aumenta (disminuye) 0,661 euros.

El modelo de regresión lineal simple normalmente tiene un término independiente, tal y como hemos supuesto hasta ahora. Incluso cuando dicho término no tiene significado económico generalmente se incluye, porque al no imponer la restricción de que pase por el origen de coordenadas, la recta de regresión estimada puede adaptarse mejor a la nube de puntos del diagrama de dispersión. No obstante, no es necesario que se incluya siempre.

El modelo puede plantearse, también, sin ordenada en el origen, simplemente, haciendo α igual a cero en la ecuación que lo representa:

$$y_t = \beta x_t$$

Para estimar el valor de β en esta ecuación utilizando el procedimiento de mínimos cuadrados ordinarios, se iguala a cero la derivada de la suma de cuadrados de errores con respecto al único coeficiente de la ecuación:

$$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N}(y_t - bx_t)\, x_t$$

$$\sum_{t=1}^{N}(y_t - bx_t)\, x_t = 0$$

Operando:

$$\sum_{t=1}^{N} y_t x_t - b\sum_{t=1}^{N} x_t^2 = 0$$

de donde se deduce:

$$b = \frac{\sum_{t=1}^{N} y_t x_t}{\sum_{t=1}^{N} x_t^2}$$

Por ejemplo, si el modelo del gasto en turismo se hubiese planteado sin ordenada en el origen, lo habríamos representado mediante la ecuación:

$y_t = \beta x_t$.

y con los mismos datos que se han utilizado para estimar el modelo con ordenada en el origen, la estimación del parámetro β sería:

$$b = \frac{\sum_{t=1}^{N} y_t x_t}{\sum_{t=1}^{N} x_t^2} = \frac{30.886.000}{70.054.800} = 0{,}441$$

de manera que el modelo estimado que resulta es:

$\hat{y}_t = 0{,}441\ x_t$

En Gráfico 6.6 se ha reproducido el 6.5, pero añadiendo la recta estimada cuando no se incluye término independiente en el planteamiento del modelo. Así puede observarse que, como hemos dicho, la recta que no se fuerza a pasar por el origen de coordenadas se adapta mejor a la nube de puntos del diagrama de dispersión que la que contiene dicha restricción.

Cualquiera de los dos modelos estimados permite obtener estimaciones de los valores del gasto en turismo para cada valor dado del ingreso, sin más que sustituir en las correspondientes ecuaciones.

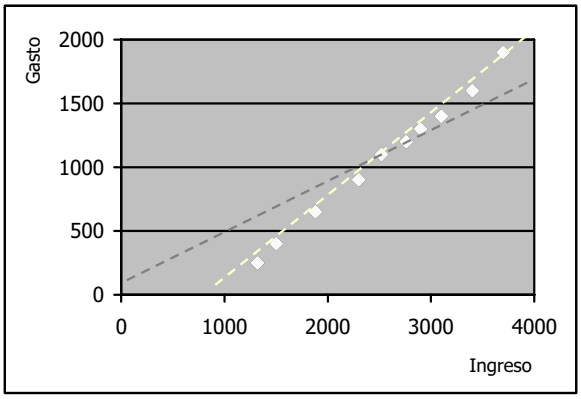


Gráfico 6.6. Diagrama de dispersión del gasto y el ingreso

En el modelo con ordenada en el origen, $\hat{y}_t = -608{,}123 + 0{,}661\ x_t$

Para la primera familia t es igual a 1, y se tiene:

$$\hat{y}_1 = -608{,}123 + 0{,}661\ x_1 = -608{,}123 + 0{,}661 \times 1.320 = 264{,}660$$

Para la segunda familia t es igual a 2, luego:

$$\hat{y}_2 = -608{,}123 + 0{,}661\ x_2 = -608{,}123 + 0{,}661 \times 1.500 = 383{,}676$$

De la misma forma, se obtienen los valores estimados del gasto en turismo para las demás familias de la muestra. En particular:

Para la última familia t es igual a 10, y resulta:

$$\hat{y}_{10} = -608{,}123 + 0{,}661\ x_{10} = -608{,}123 + 0{,}661 \times 3.700 = 1.831{,}310$$

En el modelo sin ordenada en el origen, $\hat{y}_t = 0{,}441\ x_t$

Para la primera familia t es igual a 1, y se tiene:

$\hat{y}_1 = 0{,}441\ x_1 = 0{,}441 \times 1.320 = 581{,}966$

Para la segunda familia t es igual a 2, luego:

$\hat{y}_2 = 0{,}441\ x_2 = 0{,}441 \times 1.500 = 661{,}325$

De la misma forma, se obtienen los valores estimados del gasto en turismo para las demás familias de la muestra. En particular:

Para la última familia t es igual a 10, y resulta:

$\hat{y}_{10} = 0{,}441\ x_{10} = 0{,}441 \times 3.700 = 1.631{,}27$

¿Cuál de los dos modelos estima mejor el gasto? Obviamente, el modelo que proporciona valores estimados del gasto más próximos a sus verdaderos valores, es decir, el modelo con el cual se cometen errores menores. Los errores son, como hemos dicho, las diferencias entre el valor observado y estimado del gasto en turismo; es decir, $e_t = y_t - \hat{y}_t$.

En el modelo con ordenada en el origen:

$e_1 = y_1 - \hat{y}_1 = 250 - 264{,}660 = -14{,}660$

$e_2 = y_2 - \hat{y}_2 = 400 - 383{,}676 = 16{,}324$

...

$e_{10} = y_{10} - \hat{y}_{10} = 1.900 - 1.838{,}310 = 61{,}687$

En el modelo sin ordenada en el origen:

$e_1 = y_1 - \hat{y}_1 = 250 - 581{,}966 = -331{,}966$

$e_2 = y_2 - \hat{y}_2 = 400 - 661{,}325 = -261{,}325$

...

$e_{10} = y_{10} - \hat{y}_{10} = 1.900 - 1.631{,}270 = 268{,}731$

Resumiendo en una tabla los valores estimados del gasto y los errores con ambos modelos:

Tabla 6.2. Valores estimados del gasto en turismo y errores de la estimación

(1) Modelo con término independiente: $\hat{y}_t = -608{,}123 + 0{,}661\ x_t$

(2) Modelo sin término independiente: $\hat{y}_t = 0{,}441\ x_t$

y	$\hat{y}$ (1)	$\hat{y}$ (2)	e (1)	e (2)
250	264,66	581,97	-14,66	-331,97
400	383,68	661,33	16,32	-261,33
650	634,93	828,86	15,07	-178,86
900	912,64	1.014,03	-12,63	-144,03
1.100	1.058,10	1.111,03	41,90	-11,02
1.200	1.216,79	1.216,84	-16,79	-16,84
1.300	1.309,35	1.278,56	-9,35	21,44
1.400	1.441,59	1.366,74	-41,59	33,26
1.600	1.639,95	1.499,00	-39,95	101,00
1.900	1.838,31	1.631,27	61,69	268,73

Comparando las columnas primera y segunda y primera y tercera de esta tabla, puede observarse que, tal como veíamos en el Gráfico 6.6, los valores estimados del gasto están considerablemente más próximos a los observados con el modelo que tiene ordenada en el origen; de manera que, como puede verse comparando las dos últimas columnas, los errores cometidos son significativamente menores utilizando dicho modelo.

6.3. Propiedades del ajuste

El desarrollo efectuado para la obtención de los estimadores permite demostrar que el ajuste mínimo cuadrático ordinario tiene algunas propiedades interesantes.

1. Si el modelo tiene término independiente:

Como hemos visto, derivando la suma de cuadrados de errores con respecto a los dos coeficientes que incluye la ecuación e igualando las derivadas a cero, resulta el siguiente sistema:

(1) $$\frac{\delta SCE}{\delta a} = -2\sum_{t=1}^{N}(y_t - (a + bx_t)) = 0$$

(2) $$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N}(y_t - (a + bx_t))x_t = 0$$

La ecuación (1) es tal como:

$$-2\sum_{t=1}^{N}(y_t - (a + bx_t)) = 0$$

que simplificando se reduce a:

$$\sum_{t=1}^{N}(y_t - (a + bx_t)) = 0$$

Dado que $a + bx_t = \hat{y}_t$ se deduce que:

$$\sum_{t=1}^{N}(y_t - \hat{y}_t) = 0$$

y como la diferencia entre los valores observados y estimados del regresando es igual al error, se tiene:

$$\sum_{t=1}^{N} e_t = 0$$

Es decir, **la suma de los errores de la estimación es nula**. Esto implica que también **es igual a cero la media de los errores de la estimación**:

$$\bar{e} = \frac{\sum_{t=1}^{N} e_t}{N} = \frac{0}{N} = 0$$

Además:

$$\sum_{t=1}^{N}(y_t - \hat{y}_t) = 0$$

permite deducir que **las sumas y las medias del regresando y del regresando estimado son iguales** ya que, aplicando el sumatorio:

$$\sum_{t=1}^{N} y_t = \sum_{t=1}^{N} \hat{y}_t$$

y dividiendo ambos lados de la igualdad por N, se tiene que $\overline{y} = \overline{\hat{y}}$.

Por otra parte:

$$\sum_{t=1}^{N}(y_t - (a + bx_t)) = 0$$

permite deducir que **la recta de regresión estimada pasa por el punto de coordenadas medias del regresando y el regresor** o **centro de gravedad de la nube de puntos**, ya que, aplicando el sumatorio:

$$\sum_{t=1}^{N} y_t - Na - b\sum_{t=1}^{N} x_t = 0$$

luego:

$$\sum_{t=1}^{N} y_t = Na + b\sum_{t=1}^{N} x_t$$

y al dividir ambos lados de esta expresión por N, se tiene que:

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

Esta igualdad indica que el punto de coordenadas $(\overline{x}, \overline{y})$ satisface la ecuación de la recta de regresión estimada, luego la recta pasa por dicho punto.

La ecuación (2) es tal como:

$$\sum_{t=1}^{N}(y_t-(a+bx_t))x_t=0$$

y, junto con los resultados anteriormente obtenidos, permite deducir que **los errores de la estimación están incorrelacionados con los valores del regresor;** es decir, que el coeficiente de correlación lineal entre el error y el regresor es igual a cero.

Dicho coeficiente se define como:

$$r_{ex}=\frac{S_{ex}}{S_e\cdot S_x}$$

donde:

$$S_{ex}=\frac{\sum_{t=1}^{N}(e_t-\overline{e})(x_t-\overline{x})}{N}=$$

y como la media de los errores es igual a cero:

$$=\frac{\sum_{t=1}^{N}e_t(x_t-\overline{x})}{N}=$$

operando y aplicando el sumatorio:

$$=\frac{\sum_{t=1}^{N}e_tx_t-\overline{x}\sum_{t=1}^{N}e_t}{N}=$$

y, finalmente, teniendo en cuenta que la suma de los errores es nula:

$$=\frac{\sum_{t=1}^{N}e_tx_t}{N}$$

Por tanto, para demostrar que el coeficiente de correlación lineal entre el error y el regresor es nulo, basta comprobar que es igual a cero el numerador de esta fracción.

Para ello, partiendo de la segunda ecuación del sistema, se tiene que:

$$\sum_{t=1}^{N}(y_t - (a + bx_t))x_t = 0$$

por tanto:

$$\sum_{t=1}^{N}(y_t - \hat{y}_t)x_t = 0$$

luego:

$$\sum_{t=1}^{N} e_t x_t = 0$$

Por último, puede demostrarse también que **los errores están incorrelacionados con los valores estimados del regresando**; es decir, que también es nulo el coeficiente de correlación entre ambas variables.

Dicho coeficiente se define como:

$$r_{e\hat{y}} = \frac{S_{e\hat{y}}}{S_e \cdot S_{\hat{y}}}$$

donde:

$$S_{e\hat{y}} = \frac{\sum_{t=1}^{N}(e_t - \bar{e})(\hat{y}_t - \bar{\hat{y}})}{N} =$$

al ser nula la media de los errores:

$$= \frac{\sum_{t=1}^{N} e_t(\hat{y}_t - \bar{\hat{y}})}{N} =$$

operando y aplicando el sumatorio:

$$= \frac{\sum_{t=1}^{N} e_t\hat{y}_t - \bar{\hat{y}}\sum_{t=1}^{N} e_t}{N} =$$

y teniendo en cuenta que la suma de los errores es igual a cero:

$$= \frac{\sum_{t=1}^{N} e_t\hat{y}_t}{N}$$

Por tanto, para demostrar que el coeficiente de correlación lineal entre el error y el regresando estimado es nulo, basta comprobar que es igual a cero el numerador de esta fracción:

$$\sum_{t=1}^{N} e_t\hat{y}_t =$$

sustituyendo el regresando estimado por la expresión que lo define:

$$\sum_{t=1}^{N} e_t(a + bx_t) =$$

operando:

$$= a\sum_{t=1}^{N} e_t + b\sum_{t=1}^{N} e_t x_t$$

que, dado que tanto la suma de los errores como la suma del producto de los errores por los valores del regresor son nulas, es igual a cero.

2. Si la ecuación no tiene término independiente:

Al derivar la suma de cuadrados de los errores respecto al único coeficiente desconocido que incluye la ecuación e igualarla a cero:

$$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N}(y_t - bx_t)x_t = 0$$

De esta ecuación se deduce que:

$$\sum_{t=1}^{N}(y_t - bx_t)x_t = 0$$

y en el modelo sin ordenada en el origen, $bx_t = \hat{y}_t$, de manera que:

$$\sum_{t=1}^{N}(y_t - \hat{y}_t)x_t = 0$$

y, por tanto:

$$\sum_{i=1}^{N}e_t x_t = 0$$

Es decir, **la suma del producto de los valores de los errores por los del regresor es nula**. Sin embargo, en este caso, **esto no implica que los errores y los regresores estén incorrelacionados**, ya que:

$$S_{ex} = \frac{\sum_{t=1}^{N}(e_t - \bar{e})(x_t - \bar{x})}{N} = \frac{\sum_{t=1}^{N}(e_t x_t - \bar{x}e_t - \bar{e}x_t + \overline{ex})}{N} =$$

$$= \frac{\sum_{t=1}^{N}e_t x_t}{N} - \bar{x}\frac{\sum_{t=1}^{N}e_t}{N} - \bar{e}\frac{\sum_{t=1}^{N}x_t}{N} + \frac{N\overline{ex}}{N} = \frac{\sum_{t=1}^{N}e_t x_t}{N} - \overline{xe} - \overline{ex} + \overline{ex} = \frac{\sum_{t=1}^{N}e_t x_t}{N} - \overline{ex}$$

donde el primer sumando es nulo, porque es nulo el numerador de la fracción, pero, si el modelo no tiene ordenada en el origen, el segundo sumando no es nulo, porque la propiedad de que la media de los errores es igual a cero no se puede

demostrar, de manera que la covarianza de los errores y el regresor es no nula y es no nulo el coeficiente de correlación.

También puede demostrase que:

$$\sum_{i=1}^{N} e_t \hat{y}_t = 0$$

es decir, que **la suma del producto de los valores de los errores por los valores estimados del regresando es nula**:

$$\sum_{i=1}^{N} e_t \hat{y}_t = \sum_{i=1}^{N} e_t b x_t = b \sum_{i=1}^{N} e_t x_t = 0$$

Pero de nuevo **esto no implica que los errores estén incorrelacionados con los valores estimados del regresando**, ya que:

$$s_{e\hat{y}} = \frac{\sum_{t=1}^{N} (e_t - \bar{e})(\hat{y}_t - \bar{\hat{y}})}{N} = \frac{\sum_{t=1}^{N} (e_t \hat{y}_t - \bar{\hat{y}} e_t - \bar{e} \hat{y}_t + \bar{e}\bar{\hat{y}})}{N} =$$

$$= \frac{\sum_{t=1}^{N} e_t \hat{y}_t}{N} - \bar{\hat{y}} \frac{\sum_{t=1}^{N} e_t}{N} - \bar{e} \frac{\sum_{t=1}^{N} \hat{y}_t}{N} + \frac{N \bar{e} \bar{\hat{y}}}{N} = \frac{\sum_{t=1}^{N} e_t \hat{y}_t}{N} - \bar{\hat{y}} \bar{e} - \bar{e} \bar{\hat{y}} + \bar{e} \bar{\hat{y}} = \frac{\sum_{t=1}^{N} e_t \hat{y}_t}{N} - \bar{e} \bar{\hat{y}}$$

donde el primer sumando es nulo, porque es nulo el numerador de la fracción, pero, si el modelo no tiene ordenada en el origen, el segundo sumando no es nulo, porque la propiedad de que la media de los errores es igual a cero no se puede demostrar, de manera que la covarianza de los errores y el regresando estimado es no nula y es no nulo el coeficiente de correlación.

Para comprobar empíricamente el cumplimiento de estas propiedades, pueden utilizarse los resultados obtenidos en el ejemplo anterior.

Tabla 6.3. Comprobación de las propiedades del ajuste

Modelo con término independiente: $\hat{y}_t = -608{,}123 + 0{,}661\ x_t$

Familia	y (1)	x (2)	$\hat{y}$ (3)	e (4)	ex = (4) x (2)	e$\hat{y}$ = (4) x (3)
1	250	1.320	264,66	-14,66	-19.350,81	-3.879,83
2	400	1.500	383,68	16,32	24.486,73	6.263,31
3	650	1.880	634,93	15,07	28.329,53	9.567,71
4	900	2.300	912,64	-12,63	-29.059,71	-11.530,82
5	1.100	2.520	1.058,10	41,90	105.591,98	44.336,00
6	1.200	2.760	1.216,79	-16,79	-46.329,81	-20.425,17
7	1.300	2.900	1.309,35	-9,35	-27.126,64	-12.247,71
8	1.400	3.100	1.441,59	-41,59	-128.940,78	-59.961,36
9	1.600	3.400	1.639,95	-39,95	-135.841,83	-65.521,85
10	1.900	3.700	1.838,31	61,68	228.241,32	113.399,7
Sumas	10.700		10.700	0,00	0,00	0,00

Esta tabla recoge, en la columna (1) los datos del gasto anual en turismo, en la columna (2) los ingresos familiares mensuales, en la columna (3) los ingresos familiares estimados utilizando el modelo con ordenada en el origen y en la columna (4) los errores de la estimación, que hemos obtenido previamente.

En ellas puede comprobarse que la suma de los errores es igual a cero (por tanto, también es igual a cero su media), y que las sumas de los valores observados y estimados del gasto en turismo son iguales (por tanto, también son iguales sus medias).

La columna siguiente se obtiene multiplicando las columnas (4) y (2), y proporciona los productos de los errores por los valores del ingreso familiar mensual. Y la última columna de la tabla se obtiene multiplicando las columnas (4) y (3), y proporciona los productos de los errores por los valores del gasto anual en turismo estimado.

Puede comprobarse que las sumas de ambas columnas es igual a cero. Como, además, ya hemos comprobado que también es nula la media de los errores, esto significa que los errores y el regresor y los errores y el regresando estimado están incorrelacionados.

Únicamente queda comprobar que la recta estimada, de ecuación:

$\hat{y}_t = -608{,}123 + 0{,}661\ x_t$

pasa por el centro de gravedad de la nube de puntos, que es el punto de coordenadas:

$(\bar{x}, \bar{y}) = (2.538, 1.070)$

y efectivamente, puede comprobarse que:

$1.070 = -608{,}123 + 0{,}661 \times 2.538$

luego, el punto de coordenadas medias de las variables satisface la ecuación de la recta estimada, por tanto, la recta estimada pasa por dicho punto.

Tabla 6.4. Comprobación de las propiedades del ajuste
Modelo sin término independiente: $\hat{y}_t = 0{,}441\ x_t$

Familia	y (1)	x (2)	$\hat{y}$ (3)	e (4)	ex = (4) x (2)	e$\hat{y}$ = (4) x (3)
1	250	1.320	581,97	-331,97	-438.195,28	-193.193,03
2	400	1.500	661,33	-261,33	-391.987,70	-172.820,88
3	650	1.880	828,86	-178,86	-336.258,37	-148.250,74
4	900	2.300	1.014,03	-144,03	-262.273,31	-115.631,95
5	1.100	2.520	1.111,03	-11,026	-27.786,09	-12.250,43
6	1.200	2.760	1.216,84	-16,84	-46.473,56	-20.489,42
7	1.300	2.900	1.278,56	21,44	62.170,42	27.409,91
8	1.400	3.100	1.366,74	33,26	103.110,31	45.459,63
9	1.600	3.400	1.499,00	101,00	343.387,63	151.393,92
10	1.900	3.700	1.631,27	268,73	994.305,94	438.373,01
Sumas	10.700		11.189,63	-519,63	0,00	0,00

Esta tabla recoge, en la columna (1) los datos del gasto anual en turismo, en la columna (2) los ingresos familiares mensuales, en la columna (3) los ingresos familiares estimados utilizando el modelo sin ordenada en el origen y en la columna (4) los errores de la estimación, que hemos obtenido previamente.

En ellas puede comprobarse que la suma de los errores no es igual a cero (por tanto, tampoco es nula su media), y que las sumas de los valores observados y estimados del gasto en turismo no son iguales (por tanto, tampoco son iguales sus medias).

La columna siguiente se obtiene multiplicando las columnas (4) y (2), y proporciona los productos de los errores por los valores del ingreso familiar mensual. Y la última columna de la tabla se obtiene multiplicando las columnas (4) y (3), y proporciona los productos de los errores por los valores del gasto anual en turismo estimado.

Puede comprobarse que las sumas de ambas columnas es igual a cero. Sin embargo, al no ser nula la media de los errores, esto no significa que los errores y el regresor y los errores y el regresando estimado estén incorrelacionados.

Únicamente queda comprobar que la recta estimada, de ecuación:

$$\hat{y}_t = 0{,}441\ x_t$$

no pasa por el centro de gravedad de la nube de puntos, que es el punto de coordenadas:

$$(\overline{x}, \overline{y}) = (2.538,\ 1.070)$$

y efectivamente, puede comprobarse que:

$$1.070 \neq 0{,}441 \times 2.538 = 1.119{,}258$$

luego, el punto de coordenadas medias de las variables no satisface la ecuación de la recta estimada, por tanto, la recta estimada no pasa por dicho punto.

6.4. El coeficiente de determinación

Una vez que se ha obtenido la ecuación de la recta estimada, es conveniente analizar hasta qué punto dicha recta representa correctamente la nube de puntos que

recoge la información muestral. Como primera aproximación, dicho análisis puede efectuarse en términos gráficos, superponiendo la recta estimada sobre el diagrama de dispersión.

En el ejemplo que hemos analizado, la recta que se obtiene si el modelo se plantea con término independiente casi se superpone a los puntos del diagrama de dispersión; por tanto, representa correctamente la nube de puntos; es decir, se ha efectuado un buen ajuste. Sin embargo, la recta que se obtiene cuando el modelo se plantea sin término independiente está muy alejada de la mayoría de los puntos del diagrama de dispersión, de manera que es poco representativa de la nube de puntos; es decir, no se ha efectuado un buen ajuste.

En definitiva, el ajuste es tanto mejor cuanto menor es la distancia desde los puntos de la nube a la recta estimada o, lo que es lo mismo, cuanto más similares son los valores estimados del regresando a los verdaderos; es decir, cuanto más parecidas son las medias y las varianzas de ambas variables.

1. **Si el modelo tiene término independiente**, las medias del regresando y del regresando estimado son iguales. En cuanto a las varianzas, partiendo de la siguiente igualdad:

$$y_t - \overline{y} = \hat{y}_t - \overline{y} + y_t - \hat{y}_t$$

se deduce que:

$$y_t - \overline{y} = \hat{y}_t - \overline{y} + e_t$$

de manera que:

$$(y_t - \overline{y})^2 = ((\hat{y}_t - \overline{y}) + e_t)^2$$

$$(y_t - \overline{y})^2 = (\hat{y}_t - \overline{y})^2 + e_t^2 + 2e_t(\hat{y}_t - \overline{y})$$

$$(y_t - \overline{y})^2 = (\hat{y}_t - \overline{y})^2 + e_t^2 + 2e_t\hat{y}_t + 2\overline{y}e_t$$

sumando en t en ambos lados de la igualdad:

$$\sum_{t=1}^{N}(y_t-\overline{y})^2=\sum_{t=1}^{N}((\hat{y}_t-\overline{y})^2+e_t^2+2e_t\hat{y}_t+2\overline{y}e_t)$$

y aplicando sumatorio el:

$$\sum_{t=1}^{N}(y_t-\overline{y})^2=\sum_{t=1}^{N}(\hat{y}_t-\overline{y})^2+\sum_{t=1}^{N}e_t^2+2\sum_{t=1}^{N}e_t\hat{y}_t+2\overline{y}\sum_{t=1}^{N}e_t$$

Dado que, en el modelo con ordenada en el origen $\overline{y}=\overline{\hat{y}}$, $\sum_{t=1}^{N}e_t\hat{y}_t=0$ y $\sum_{t=1}^{N}e_t=0$ la igualdad anterior se reduce a:

$$\sum_{t=1}^{N}(y_t-\overline{y})^2=\sum_{t=1}^{N}(\hat{y}_t-\overline{\hat{y}})^2+\sum_{t=1}^{N}e_t^2$$

y teniendo en cuenta que la media de los errores es nula:

$$\sum_{t=1}^{N}(y_t-\overline{y})^2=\sum_{t=1}^{N}(\hat{y}_t-\overline{\hat{y}})^2+\sum_{t=1}^{N}(e_t-\overline{e})^2$$

dividiendo por N en ambos lados de la igualdad, se tiene:

$$\frac{\sum_{t=1}^{N}(y_t-\overline{y})^2}{N}=\frac{\sum_{t=1}^{N}(\hat{y}_t-\overline{\hat{y}})^2}{N}+\frac{\sum_{t=1}^{N}(e_t-\overline{e})^2}{N}$$

es decir:

$$s_y^2=s_{\hat{y}}^2+s_e^2$$

igualdad que se conoce como descomposición de la varianza, puesto que indica que la varianza del regresando puede descomponerse en la suma de las varianzas del regresando estimado y de los errores; de manera que las varianzas del regresando y del regresando estimado son tanto más similares cuanto menor es la varianza del error.

De la descomposición de la varianza, se deduce que:

$$1 = \frac{s_{\hat{y}}^2}{s_y^2} + \frac{s_e^2}{s_y^2}$$

y el ajuste será tanto mejor cuanto más próxima a la unidad esté la primera fracción o cuanto más próxima a cero esté la segunda.

Para valorar, entonces, la calidad del ajuste, puede utilizarse el coeficiente de determinación, R^2, definido por cociente entre las varianzas del regresando y el regresando estimado, que es igual a la diferencia entre la unidad y el cociente de varianzas del error y del regresando:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Puesto que R^2 es un cociente de varianzas no toma valores menores que cero. Como también es igual a uno menos un cociente de varianzas, tampoco toma valores superiores a la unidad. El coeficiente de determinación toma, entonces, valores comprendidos entre cero y uno.

Si el coeficiente de determinación es igual a uno, las varianzas del regresando y el regresando estimado son iguales y la varianza del error es igual a cero, de manera que el ajuste es perfecto. Si el coeficiente de determinación es igual a cero, la varianza del error es tan elevada que coincide con la varianza del regresando, y la varianza del regresando estimado es igual a cero, de manera que el ajuste es pésimo. Por tanto, valores del coeficiente de determinación próximos a la unidad son indicativos de buenos ajustes y valores del coeficiente de determinación próximos a cero son indicativos de malos ajustes. En definitiva, **el ajuste es tanto mejor cuanto más próximo a la unidad está el coeficiente de determinación**.

Además, ya que $\mathbf{R^2}$ es el cociente de varianzas del regresando estimado y el regresando, indica qué proporción de la varianza del regresando es la varianza del regresando estimado, de manera que, **multiplicado por cien, es el porcentaje de variaciones observadas de y que quedan explicadas con el modelo estimado**. De forma similar, como $1 - R^2$ es igual al cociente de varianzas del error y el regresando, indica qué proporción de la varianza del regresando es la varianza del

error, de manera que, multiplicado por cien, es el porcentaje de variaciones observadas del regresando que la ecuación estimada deja sin explicar.

Luego, si se obtiene un buen ajuste, la ecuación explica un elevado porcentaje de las variaciones observadas del regresando, mientras que si el ajuste es malo la ecuación deja sin explicar la mayor parte de dichas variaciones, generalmente a consecuencia de un error en el planteamiento del modelo.

Finalmente, en el caso de la regresión lineal simple, puede comprobarse que el coeficiente de determinación es igual al cuadrado del coeficiente de correlación lineal simple entre las dos variables de la ecuación, puesto que:

$$R^2 = \frac{s^2_{\hat{y}}}{s^2_y} = \frac{\sum_{t=1}^{N}(\hat{y}_t - \bar{\hat{y}})^2}{\sum_{t=1}^{N}(y_t - \bar{y})^2}$$

siendo:

$$\hat{y}_t = a + bx_t$$

$$\bar{y} = a + b\bar{x}$$

entonces:

$$\sum_{t=1}^{N}(\hat{y}_t - \bar{\hat{y}})^2 = \sum_{t=1}^{N}(a + bx_t - a - b\bar{x})^2 = b^2 \sum_{t=1}^{N}(x_t - \bar{x})^2 =$$

$$= \left[\frac{s_{xy}}{s^2_x}\right]^2 \sum_{t=1}^{N}(x_t - \bar{x})^2 = \frac{s^2_{xy} \sum_{t=1}^{N}(x_t - \bar{x})^2}{(\sum_{t=1}^{N}(x_t - \bar{x})^2 / N)^2}$$

Por tanto:

$$\frac{\sum_{t=1}^{N}(\hat{y}_t-\bar{\hat{y}})^2}{\sum_{t=1}^{N}(y_t-\bar{y})^2}=\frac{s_{xy}^2\sum_{t=1}^{N}(x_t-\bar{x})^2}{((\sum_{t=1}^{N}(x_t-\bar{x})^2)^2/N^2)\times(\sum_{t=1}^{N}(y_t-\bar{y})^2)}=$$

$$=\frac{s_{xy}^2}{(\sum_{t=1}^{N}(x_t-\bar{x})^2/N)\times(\sum_{t=1}^{N}(y_t-\bar{y})^2/N)}=\frac{s_{xy}^2}{s_x^2s_y^2}=r^2$$

es decir, $R^2 = r^2$.

2. **Si el modelo no tiene término independiente**, aunque $\sum_{t=1}^{N}e_t\hat{y}_t=0$, en general, la suma de los errores no es nula y las medias del regresando y el regresando estimado no son iguales, de manera que no se puede descomponer la varianza.

$$s_y^2 \neq s_{\hat{y}}^2+s_e^2$$

luego:

$$1\neq\frac{s_{\hat{y}}^2}{s_y^2}+\frac{s_e^2}{s_y^2}$$

es decir, el cociente de varianzas del regresando estimado y el regresando no es igual a uno menos el cociente de varianzas del error y el regresando, por lo que el coeficiente de determinación se define sólo de esta última forma:

$$R^2=1-\frac{s_e^2}{s_y^2}$$

y **no se puede interpretar como el porcentaje de variaciones del regresando que quedan explicadas por el modelo estimado**.

Además, aunque su valor máximo posible es la unidad, su valor mínimo posible no es cero, puesto que como no se puede descomponer la varianza, no se tiene ninguna garantía de que la varianza del error sea menor que la varianza del regresando y si es mayor, el coeficiente de determinación es negativo.

A modo ilustrativo, se calcula e interpreta el valor del coeficiente de determinación para el ejemplo del epígrafe anterior.

Para ello es necesario obtener previamente las varianzas del regresando, del regresando estimado y de los errores.

Tabla 6.5. Cálculos intermedios para la obtención de R^2

Modelo con término independiente: $\hat{y}_t = -608{,}123 + 0{,}661\ x_t$

Familia	y	y^2	$\hat{y}$	$\hat{y}^2$	e	e^2
1	250	62.500	264,66	70.044,76	-14,66	214,91
2	400	160.000	383,68	147.206,90	16,32	266,49
3	650	422.500	634,93	403.137,51	15,07	227,07
4	900	810.000	912,64	832.902,01	-12,63	159,63
5	1.100	1.210.000	1.058,10	1.119.572,26	41,90	1.755,74
6	1.200	1.440.000	1.216,79	1.480.568,57	-16,79	281,78
7	1.300	1.690.000	1.309,35	1.714.407,93	-9,35	87,50
8	1.400	1.960.000	1.441,59	2.078.192,68	-41,59	1.730,04
9	1.600	2.560.000	1.639,95	2.689.447,41	-39,95	1.596,28
10	1.900	3.610.000	1.838,31	3.379.395,26	61,69	3.805,27
Sumas		13.925.000		13.914.875,29		10.124,71

$$s_y^2 = \frac{\sum_{t=1}^{N} y_t^2}{N} - \bar{y}^2 = \frac{13.925.000}{10} - 1.070^2 = 247.600$$

$$s_{\hat{y}}^2 = \frac{\sum_{t=1}^{N} \hat{y}_t^2}{N} - \bar{\hat{y}}^2 = \frac{13.914.875{,}29}{10} - 1.070^2 = 246.587{,}53$$

$$s_e^2 = \frac{\sum_{t=1}^{N} e_t^2}{N} = \frac{10.124{,}71}{10} = 1.012{,}47$$

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{246.587{,}53}{247.600} = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{1.012{,}47}{247.600} = 0{,}99$$

Este valor del coeficiente de determinación tan próximo a la unidad indica que la recta estimada es muy representativa de la nube de puntos; es decir, que se ha

obtenido un muy buen ajuste. La regresión efectuada explica prácticamente la totalidad de las variaciones muestrales del gasto en turismo.

Este resultado es coherente con las conclusiones deducidas de la representación gráfica, que ya hemos comentado.

Tabla 6.6. Cálculos intermedios para la obtención de R^2

Modelo sin término independiente: $\hat{y}_t = 0{,}441\ x_t$

Familia	y	y^2	$\hat{y}$	$\hat{y}^2$	e	e^2
1	250	62.500	581,97	338.684,56	-331,97	110.201,50
2	400	160.000	661,33	437.350,93	-261,33	68.290,83
3	650	422.500	828,86	687.010,28	-178,86	31.991,20
4	900	810.000	1.014,03	1.028.260,64	-144,03	13.003,27
5	1.100	1.210.000	1.111,03	1.234.379,27	-11,026	121,58
6	1.200	1.440.000	1.216,84	1.480.695,32	-16,84	283,53
7	1.300	1.690.000	1.278,56	1.634.720,60	21,44	459,59
8	1.400	1.960.000	1.366,74	1.867.974,43	33,26	1.106,32
9	1.600	2.560.000	1.499,00	2.247.011,90	101,00	10.200,27
10	1.900	3.610.000	1.631,27	2.661.041.81	268,73	72.216,53
Sumas		13.925.000		13.617.125,39		307.874,61

$$s_y^2 = \frac{\sum_{t=1}^{N} y_t^2}{N} - \bar{y}^2 = \frac{13.925.000}{10} - 1.070^2 = 247.600$$

$$s_e^2 = \frac{\sum_{t=1}^{N} e_t^2}{N} = \frac{307.874,61}{10} = 30.787,46$$

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{30.787,461}{24.7600} = 0,87$$

Este valor del coeficiente de determinación es significativamente menor que el que resulta cuando el modelo tiene término independiente.

Como se ha deducido a partir de la representación gráfica, el ajuste no resulta satisfactorio.

6.5. Regresión con EXCEL

Para obtener la recta de regresión lineal simple con EXCEL, en primer lugar, deben introducirse en columnas los valores de las variables x e y.

A continuación, se selecciona un espacio en blanco, de dos filas por cinco columnas, y se ordena **Insertar–Función–Estadísticas–Estimación lineal**, que abre un cuadro de diálogo respecto a los argumentos de la función. En la primera ventana deben introducirse los nombres de la primera y última celda que contienen los valores del regresando, separados con dos puntos. En la segunda ventana se introducen los nombres de la primera y última celda que contienen los valores del regresor, igualmente separados con dos puntos. En la tercera ventana debe escribirse la palabra "verdadero" si la ecuación incluye término independiente y la palabra "falso" si no lo incluye. En la última ventana, se escribe la palabra "falso" si sólo se desea obtener el valor estimado de la pendiente de la recta, y la palabra "verdadero" si se desea obtener algunos resultados adicionales, que es lo habitual.

Si el modelo tiene ordenada en el origen ("verdadero" en la tercera ventana), al pinchar el botón aceptar, aparece cubierta con un valor numérico una única celda. Para obtener los demás, debe sombrearse la orden en la línea de mandatos y presionar, simultáneamente, las teclas Ctrl ↑(Mayús) e Intro. Al ejecutar la orden, las celdas seleccionadas muestran el siguiente resultado:

Tabla 6.7. Resultados de la estimación lineal con EXCEL

b	a
s_b	s_a
R^2	s_e
F	gl
SCR	SCE

Si el modelo tiene ordenada en el origen ("falso" en la tercera ventana), siguiendo este procedimiento se obtiene el mismo resultado, con un valor de a igual a cero y un valor s_a no disponible (#NA).

En esta tabla, los valores b y a que figuran en la primera fila son las estimaciones de la pendiente y la ordenada en el origen de la recta de regresión.

Los valores s_b y s_a que figuran en la segunda fila corresponden a las desviaciones típicas estimadas de los coeficientes de regresión, b y a. Estas desviaciones típicas, son unos indicadores del grado de fiabilidad de los coeficientes a y b. Si cada uno de los coeficientes contiene al menos dos veces su desviación estimada, las estimaciones se consideran precisas. De acuerdo con la teoría de la regresión, el valor verdadero del coeficiente está comprendido entre el coeficiente estimado $\pm$ (aproximadamente) dos veces su desviación típica, por tanto, la desviación típica del estimador determina el radio del intervalo que contiene al valor verdadero del coeficiente con una probabilidad del 95 por ciento, que es estrecho cuando el radio del intervalo es pequeño en relación con el centro.

En la tercera fila, R^2 es el coeficiente de determinación.

El valor de gl en esta tabla es el número de observaciones "libres" para calcular los errores. Los errores se obtienen por diferencia entre el valor real y el valor estimado del regresando. Para que se puedan obtener, es necesario conocer los coeficientes a y b, lo cual requiere resolver un sistema de dos ecuaciones, que establecen dos restricciones en los datos. Los grados de libertad se obtienen por diferencia entre el número de observaciones y el número de restricciones que se les imponen (dos). Es decir, que en el caso de la regresión lineal simple, gl siempre es igual a $N - 2$.

s_e es el error estándar. Es la raíz cuadrada de la suma de los cuadrados de los errores dividida por los grados de libertad que tiene que, como hemos visto, son $N - 2$. Es una aproximación del error típico cometido al estimar los valores del regresando utilizando la ecuación de la recta de regresión. Proporciona una medida "resumen" del tamaño de los errores. Viene dado en las mismas unidades de medida que el regresando, así que comparándolo con la media muestral del regresando, se obtiene el porcentaje de error tipo medio cometido al efectuar la estimación.

El valor de F es un indicador de la capacidad que tiene el regresor para explicar las variaciones del regresando. Si toma valores elevados, indica que realmente el

regresor influye en el regresando. Si toma valores pequeños, existen dudas respecto a que x sea la variable que explica el comportamiento de y.

Finalmente, SCR es la suma de los cuadrados de las desviaciones de $\hat{y}$ respecto a su media muestral, que se denomina suma de los cuadrados de la regresión. SCE es la suma de los cuadrados de los errores; es decir, proporciona el valor de la suma de los errores elevados al cuadrado.

En el ejemplo del gasto en turismo que hemos utilizado para ilustrar la regresión, al efectuar el procedimiento indicado, EXCEL devuelve los siguientes resultados:

Si el modelo tiene ordenada en el origen ("verdadero" en la tercera ventana):

0,66119893	-608,122886
0,01497935	39,6471345
0,99591086	35,5751113
1.948,40203	8
2.465.875,29	10.124,7084

Si el modelo no tiene ordenada en el origen ("falso" en la tercera ventana):

0,44088342	0
0,02209768	#N/A
0,87565646	184,954831
63,3801168	9
2.168.125,39	307.874,607

Resultados que, como puede comprobarse, coinciden con los que se han obtenido previamente.

Para completar el manual se incluyen en este Apéndice algunas cuestiones y ejercicios planteados a los alumnos en los exámenes de la asignatura Estadística Aplicada al Sector Turístico que se imparte en la Diplomatura de Turismo de la Universidad de A Coruña.

Hemos tratado de ordenarlos respetando, en la medida de lo posible, la estructura del texto, aunque en ocasiones con la misma tabla de datos se han propuesto cuestiones relativas a distintas partes de la materia.

EJERCICIO 1

La tabla siguiente recoge el resultado de una encuesta efectuada en distintos establecimientos hoteleros respecto a la inversión (euros) efectuada en el año 2002 para el mantenimiento de las instalaciones.

Inversión ($L_{i-1} - L_i$)	Nº establecimientos (n_i)
3.000 – 4.500	125
4.500 – 6.000	185
6.000 – 7.500	245
7.500 – 8.500	212
8.500 – 12.500	155
12.500 – 20.000	78

La distribución de la inversión **está**, por tanto, **agrupada en** 6 **intervalos de amplitud variable**.

Dada esta información, calcule:

1.1. El porcentaje de establecimientos que invierte menos de 8.500 euros.
1.2. El número de establecimientos que invierte entre 6.000 y 12.500 euros.
1.3. La inversión más frecuente.
1.4. La cantidad que como mínimo invierte un establecimiento del grupo del 25 por ciento de los que más invierten.

Para responder a estas cuestiones, se elabora la siguiente tabla de resultados intermedios:

$L_{i-1} - L_i$	n_i	f_i	N_i	F_i	c_i	d_i
3.000 – 4.500	125	0,125	125	0,125	1.500	0,083
4.500 – 6.000	185	0,185	310	0,310	1.500	0,123
6.000 – 7.500	245	0,245	555	0,555	1.500	0,163
7.500 – 8.500	212	0,212	767	0,767	1.000	0,212
8.500 – 12.500	155	0,155	922	0,922	4.000	0,038
12.500 – 20.000	78	0,078	1.000	1,000	7.500	0,010
	1.000	1,000				

1.1. El porcentaje de establecimientos que invierte menos de 8.500 euros es la frecuencia acumulada relativa correspondiente al intervalo de límite superior 8.500, multiplicada por cien. Como puede observarse, $F_4 = 0{,}767$; por tanto, **aproximadamente el 77 por ciento de los establecimientos invierte menos de 8.500 euros**.

1.2. El número de establecimientos que invierte entre 6.000 y 12.500 euros se obtiene sumando las frecuencias absolutas correspondientes a los intervalos comprendidos entre el de límite inferior 6.000 y el de límite superior 12.500. De la tabla se deduce que $n_3 + n_4 + n_5 = 245 + 212 + 155 = 612$, luego **612 establecimientos invierten entre 6.000 y 12.500 euros**.

1.3. La inversión más frecuente es la moda de la distribución, que está en el intervalo de mayor densidad de frecuencia; es decir, en este caso, en el intervalo 7.500 – 8.500, para el cual $d_i = 0{,}212$. Para obtener su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 7.500 + \frac{0{,}038}{0{,}163 + 0{,}038} \times 1.000 = 7.689{,}054$$

Entonces **la inversión más frecuente es 7.689 euros**.

1.4. La cantidad que como mínimo invierte un establecimiento del grupo del 25 por ciento de los que más invierten es el tercer cuartil de la distribución, que está en el intervalo al que le corresponde la primera frecuencia absoluta acumulada mayor o

igual que 3 ¼ N = 3 ¼ 1000 = 750; es decir, en el intervalo 7.500 – 8.500, ya que N_4 = 767. Para obtener su valor:

$$Q_3 = L_{i-1} + \frac{\frac{3N}{4} - N_{i-1}}{n_i} \times c_i = 7.500 + \frac{750 - 555}{212} \times 1.000 = 8.419{,}811$$

Por tanto, **la cantidad que como mínimo invierte un establecimiento del grupo del 25 por ciento de los que más invierten es 8.419 euros**.

EJERCICIO 2

Con los datos relativos a la edad de 210 visitantes de un municipio de la provincia de A Coruña se han calculado la media aritmética, $\bar{x}$ = 39,04; la mediana, M_e = 37,25 y el primer y tercer cuartiles, Q_1 = 27,69 y Q_3 = 47,39. Dada esta información, indique si son correctas las siguientes afirmaciones:

2.1. La edad más frecuente es 39,04 años y su media es 37,25 años.
2.2. El 25 por ciento de los visitantes más jóvenes tienen como máximo 47,39 años.
2.3. El 25 por ciento de los visitantes más jóvenes tienen como máximo 27,69 años.
2.4. El 75 por ciento de los visitantes son mayores de 47,39 años.

La afirmación **2.1 no es correcta**. La edad más frecuente es la moda (cuyo valor, con esta información, es desconocido) y no la media (= 39,04), y se da como media de la edad su valor mediano (= 37,25).

La afirmación **2.2 no es correcta**. La edad máxima del 25 por ciento de los visitantes más jóvenes es el primer cuartil (= 27,69) de la distribución y no el tercero (= 47,39). Esto indica que la afirmación **2.3 es correcta**. El 25 por ciento de los visitantes más jóvenes tienen como máximo 27,69 años.

La afirmación **2.4 no es correcta**. El tercer cuartil de la distribución (= 47,39) es la edad máxima del 75 por ciento de los visitantes más jóvenes o la edad mínima del 25 por ciento de los visitantes mayores. Es decir, el 75 por ciento de los visitantes tienen

una edad inferior a 47,39 años o, lo que es lo mismo, el 25 por ciento de los visitantes tienen una edad superior a 47,39 años.

EJERCICIO 3

La siguiente tabla corresponde a la distribución de los salarios mensuales (cientos de euros) de los trabajadores de una empresa de hostelería:

Salarios ($L_{i-1} - L_i$)	N° de trabajadores (n_i)
7 – 9	10
9 – 13	30
13 – 15	40
15 – 21	15
21 – 31	5

La distribución de los salarios **está**, por tanto, **agrupada en** 5 **intervalos de amplitud variable**.

3.1. ¿Es el salario mensual medio representativo de la distribución?
3.2. ¿Cuál sería el salario máximo que percibiría un trabajador clasificado en el grupo en el que se encuentra el 25 por ciento de los que menos ganan?
3.3. Calcule e interprete el coeficiente de asimetría de Fisher.
3.4. ¿Podría afirmarse que esta distribución es platicúrtica? ¿Por qué? Indique cuál es el significado de esta expresión.
3.5. ¿Cuál es el salario más frecuente de los trabajadores de esta empresa?

Para responder a estas cuestiones, se elabora la siguiente tabla de resultados intermedios:

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	$x_i^2 n_i$	d_i	N_i	$(x_i - \bar{x})^3 n_i$	$(x_i - \bar{x})^4 n_i$
7 – 9	8	10	80	640	5,0	10	-1.851,93	10.556,00
9 – 13	11	30	330	3.630	7,5	40	-590,49	1.594,32
13 – 15	14	40	560	7.840	20,0	80	1,08	0,32
15 – 21	18	15	270	4.860	2,5	95	1.192,60	5.128,20
21 – 31	26	5	130	3.380	0,5	100	9.304,33	114.443,32
		100	1.370	20.350			8.055,59	131.722,16

3.1. Dada esta información:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{1.370}{100} = 13,7$$

El salario mensual medio es 13,7 cientos de euros; es decir, **1.370 euros**.

La representatividad de la media del salario puede valorarse calculando el coeficiente de variación de Pearson:

$$CV = \frac{s}{\bar{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{20.350}{100} - 13,7^2 = 15,81 \Rightarrow s = 3,976$$

$$= \frac{3,976}{13,7} = 0,29$$

Es decir, la dispersión relativa de la distribución es del 29 por ciento, y **el salario mensual medio es representativo**.

3.2. El salario máximo que percibiría un trabajador clasificado en el grupo en el que se encuentra el 25 por ciento de los que menos ganan es el primer cuartil de la distribución, que está en el intervalo de extremos 9 – 13, al que le corresponde la primera frecuencia absoluta acumulada, $N_2 = 40$, mayor o igual que ¼ N = ¼ 100 = 25. Para calcular su valor:

$$Q_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i = 9 + \frac{25 - 10}{30} \times 4 = 11$$

Luego **el salario máximo de un trabajador del grupo del 25 por ciento de los que menos ganan es** 11 cientos de euros; es decir, **1.100 euros**.

3.3. El coeficiente de Fisher es nulo si la distribución es simétrica, positivo si es asimétrica a la derecha y negativo si es asimétrica a la izquierda. En este caso:

$$A_F = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 n_i \ / \ N}{s^3} = \frac{8.055{,}59/100}{(3{,}976)^3} = 1{,}28$$

Por tanto, **la distribución de los salarios es asimétrica a la derecha**; es decir, en términos gráficos, la distribución tiene cola a la derecha.

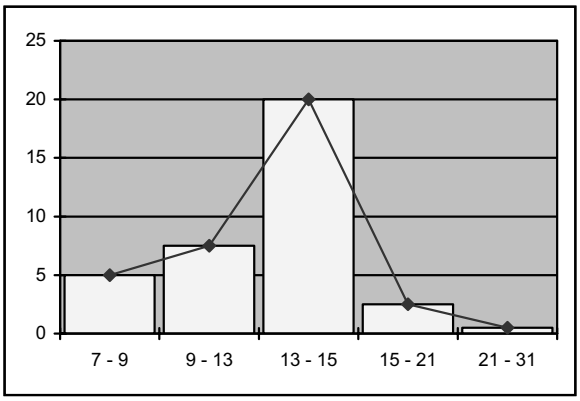


Gráfico de la distribución de los salarios

3.4. La distribución es platicúrtica si el coeficiente de exceso de apuntamiento es negativo. En este caso:

$$g = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 n_i \ / \ N}{s^4} - 3 = \frac{131.722{,}16/100}{(3{,}976)^4} - 3 = 2{,}27$$

Luego el coeficiente de exceso no es menor que cero y **la distribución no es platicúrtica**.

Si la distribución fuese platicúrtica sería menos apuntada que una distribución normal de igual media y varianza. En este caso es más apuntada que una normal con la misma media y varianza; por tanto, **es leptocúrtica**.

3.5. El salario más frecuente es la moda de la distribución, que está en el intervalo con mayor densidad de frecuencia que, en este caso, es $d_3 = 20$; es decir, la moda está en el intervalo de extremos 13 – 15. Para calcular su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 13 + \frac{2,5}{7,5 + 2,5} \times 2 = 13,5$$

Entonces, **el salario más frecuente de los trabajadores de esta empresa es** 13,5 cientos de euros; es decir, **1.350 euros**.

EJERCICIO 4

Para realizar un estudio relativo a los precios (euros) por habitación en temporada alta en los hoteles de una localidad costera española se dispone de los datos que figuran en la siguiente tabla:

Precio ($L_{i-1} – L_i$)	Nº hoteles (n_i)
0 – 70	8
70 – 100	3
100 – 120	4
120 – 200	3
200 – 300	2

La distribución de los precios **está**, por tanto, **agrupada en** 5 **intervalos de amplitud variable**.

4.1. ¿Cuál es el precio medio por habitación? ¿Es representativo de la distribución?
4.2. ¿Cuál es el precio por habitación más frecuente?
4.3. De acuerdo con el valor que toma el coeficiente de Fisher, ¿qué puede decirse respecto a la asimetría de la distribución?
4.4. Obtenga el precio mediano.
4.5. Indique cuál es el precio medio en temporada baja suponiendo que:
    4.5.1. En todos los hoteles los precios se reducen un 15 por ciento.
    4.5.2. En todos los hoteles los precios se reducen 25 euros.
4.6. ¿Cuál es el mínimo precio por habitación en un hotel del grupo en el que se encuentran el 75 por ciento de los más caros?

4.7. El valor obtenido para el coeficiente de exceso es positivo. ¿Qué información proporciona este resultado?

Para responder a estas cuestiones se elabora la siguiente tabla de resultados intermedios:

$L_{i-1} - L_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	c_i	d_i	$(x_i - \bar{x})^3 n_i$	N_i
0 – 70	8	35	280	9.800	70	0,11	-1.976.656,375	8
70 – 100	3	85	255	21.675	30	0,10	-6.218,015	11
100 – 120	4	110	440	48.400	20	0,20	7.353,063	15
120 – 200	3	160	480	76.800	80	0,04	723.667,922	18
200 – 300	2	250	500	125.000	100	0,02	7.058.329,032	20
	20		1.955	281.675			5.806.475,627	

4.1. Dada la información que contiene:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{1.955}{20} = 97,75$$

Es decir, **el precio medio es 97,75 euros por habitación**.

Para analizar si es representativo:

$$CV = \frac{s}{\bar{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{281.675}{20} - 97,75^2 = 4.528,69 \Rightarrow s = 67,29$$

$$= \frac{67,29}{97,75} = 0,69$$

La dispersión relativa de esta distribución es del 69 por ciento, por tanto **el precio medio no es representativo**.

4.2. El precio más frecuente es la moda de la distribución. El intervalo modal es aquel al que le corresponde una mayor densidad de frecuencia que, en este caso, es d_3 = 0,20. La moda está, por tanto, en el intervalo 100 – 120. Para obtener su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 100 + \frac{0,04}{0,10 + 0,04} \times 20 = 105,71$$

Luego **el precio por habitación más frecuente es 105,71 euros**.

4.3. El valor que toma el coeficiente de asimetría de Fisher es:

$$A_F = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^3 / N}{s^3} = \frac{5.806.475,627 / 20}{67,29^3} = 0,953$$

Por tanto, **la distribución es asimétrica a la derecha o positiva**; es decir, el histograma de frecuencias tiene cola a la derecha.

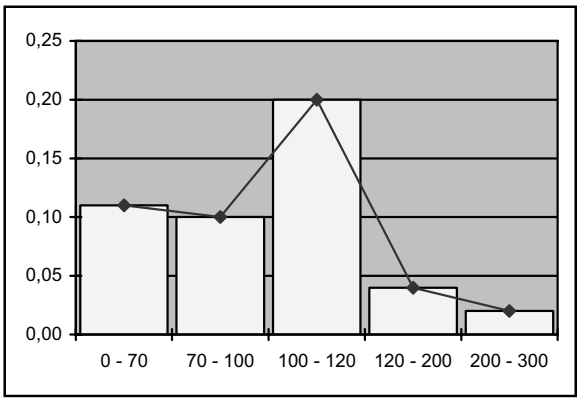


Gráfico de la distribución de los precios

4.4. La primera frecuencia absoluta acumulada igual o superior a ½ N = ½ 20 = 10 es N_2 = 11. La mediana está, entonces, en el intervalo 70 – 100. Para obtener su valor:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i = 70 + \frac{10 - 8}{3} \times 30 = 90$$

El precio mediano es 90 euros.

4.5.1. Si en temporada baja los precios se reducen un 15 por ciento:

$$x_i^e = x_i - 0{,}15x_i = (1 - 0{,}15)x_i = 0{,}85x_i$$

es decir, son iguales a los precios de temporada alta multiplicados por 0,85, lo que equivale a efectuar un cambio de escala de la variable.

$$\overline{x}^e = 0{,}85\overline{x} = 0{,}85 \times 97{,}75 = 83{,}08$$

El precio medio en temporada baja es 83,08 euros por habitación.

4.5.2. Si en temporada baja los precios se reducen 25 euros:

$$x_i^o = x_i - 25$$

es decir, son iguales a los de temporada alta menos 25 euros, lo que equivale a efectuar un cambio de origen de la variable.

$$\overline{x}^o = \overline{x} - 25 = 97{,}75 - 25 = 72{,}75$$

El precio medio en temporada baja es 72,75 euros por habitación.

4.6. El precio mínimo por habitación en un hotel del grupo del 75 por ciento de los más caros es el primer cuartil de la distribución, que está en el intervalo de extremos 0 – 70, porque la primera frecuencia absoluta acumulada mayor que ¼ N = ¼ 20 = 5 es $N_1 = 8$. Para obtener su valor:

$$Q_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i = 0 + \frac{5-0}{8} \times 70 = 43{,}75$$

Por tanto, **un hotel del grupo del 75 por ciento de los más caros cobra como mínimo 43,75 euros por habitación**.

4.7. El coeficiente de exceso indica el grado de apuntamiento o curtosis de la distribución. Si su valor es positivo, **la distribución es leptocúrtica**; es decir, es más apuntada que una normal de igual media y varianza.

EJERCICIO 5

Un grupo de turistas visitó durante unas horas un establecimiento comercial. Al final de la visita se les preguntó sobre el gasto (euros) realizado, y la información recogida figura en la tabla siguiente:

Gasto (L_{i-1} - L_i)	Nº turistas (n_i)
50 – 75	5
75 – 125	7
125 – 150	20
150 – 225	12
225 – 300	4
300 – 500	2

La distribución del gasto **está**, entonces, **agrupada en** 6 **intervalos de amplitud variable**.

5.1. Calcule la mediana de la distribución.
5.2. ¿Cuál es gasto más frecuente?
5.3. Calcule el valor que toma el coeficiente de asimetría de Pearson e indique qué información proporciona.
5.4. Obtenga el índice de Gini y explique su significado.

Con la información muestral se elabora la siguiente tabla de resultados intermedios:

$L_{i-1} – L_i$	n_i	x_i	$x_i n_i$	N_i	c_i	d_i	$x_i^2 n_i$	F_i	p_i	U_i	q_i	$p_i – q_i$
50 – 75	5	62,5	312,5	5	25	0,20	19.531,25	0,10	10	312,5	3,97	6,02
75 – 125	7	100,0	700,0	12	50	0,14	70.000,00	0,24	24	1.012,5	12,87	11,12
125 – 150	20	137,5	2.750,0	32	25	0,80	378.125,00	0,64	64	3.762,5	47,85	16,14
150 – 225	12	187,5	2.250,0	44	75	0,16	421.875,00	0,88	88	6.012,5	76,47	11,52
225 – 300	4	262,5	1.050,0	48	75	0,05	275.625,00	0,96	96	7.062,5	89,82	6,17
300 – 500	2	400,0	800,0	50	200	0,01	320.000,00	1,00	100	7.862,5	100,00	0,00
	50		7.862,5				1.485.156,25					

5.1. La primera frecuencia acumulada absoluta igual o superior a ½ N = ½ 50 = 25 es N_3 = 32; por tanto, la mediana está en el intervalo 125 – 150. Para calcular su valor:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i = 125 + \frac{25-12}{20} \times 25 = 141{,}25$$

Es decir, **la mediana de la distribución es 141,25 euros**.

5.2. El gasto más frecuente es la moda, que se encuentra en el intervalo de mayor densidad de frecuencia. En este caso, la mayor densidad es d_3 = 0,80 y el intervalo modal es 125 – 150. Para obtener su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 125 + \frac{0{,}16}{0{,}14 + 0{,}16} \times 25 = 138{,}33$$

El gasto más frecuente es 138,33 euros.

5.3. El coeficiente de asimetría de Pearson es:

$$A_P = \frac{\bar{x} - M_o}{s} =$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{7.862{,}5}{50} = 157{,}25$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{1.485.156{,}25}{50} - 157{,}25^2 = 4.975{,}56 \Rightarrow s = 70{,}53$$

$$= \frac{157{,}25 - 138{,}33}{70{,}53} = 0{,}268$$

Puesto que su valor es mayor que cero, indica que **la distribución es asimétrica a la derecha o positiva**.

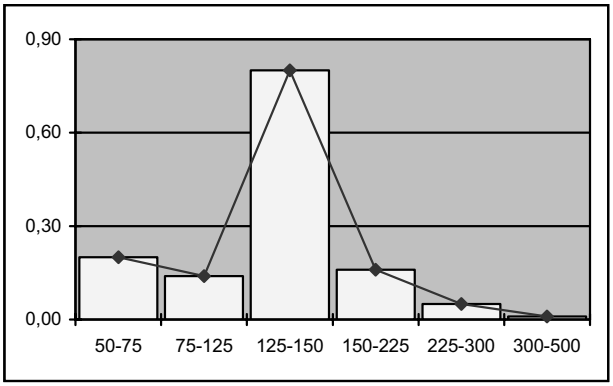


Gráfico de la distribución del gasto

5.4. El índice de Gini es:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{50,97}{282} = 0,18$$

Como su valor no está alejado de cero, **la concentración de la distribución del gasto es escasa**.

EJERCICIO 6

En la siguiente tabla figuran los sueldos anuales (cientos de euros) del personal de un hotel determinado.

Sueldos ($L_{i-1} - L_i$)	Nº empleados (n_i)
100 – 150	30
150 – 200	20
200 – 250	10
250 – 300	15
300 – 350	8

La **distribución** de los sueldos está **agrupada**, por tanto, **en** 5 **intervalos de amplitud constante**, igual a 50.

6.1. ¿Es el sueldo anual medio representativo de la distribución?
6.2. Si los sueldos estuviesen expresados en euros, ¿cuál sería la media aritmética de la distribución? ¿Variaría su representatividad? Demuéstrelo.
6.3. ¿Cuál es el sueldo anual más frecuente?
6.4. Sin hacer ningún cálculo adicional, ¿qué se puede decir respecto a la asimetría de la distribución?
6.5. Calcule el primer cuartil y explique su significado.
6.6. El coeficiente de exceso de apuntamiento de la distribución de los salarios es negativo. ¿Qué información proporciona este resultado?
6.7. Indique (sin efectuar los cálculos) cómo se obtendría el índice de Gini. Si su valor resultase igual a 0,12 ¿cómo se interpretaría?

Tabla de cálculos intermedios

$L_{i-1} - L_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	N_i
100 – 150	30	125	3.750	468.750	30
150 – 200	20	175	3.500	612.500	50
200 – 250	10	225	2.250	506.250	60
250 – 300	15	275	4.125	1.134.375	75
300 – 350	8	325	2.600	845.000	83
	83		16.225	3.566.875	

6.1. De la información que contiene esta tabla se deduce que:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{16.225}{83} = 195{,}48$$

El sueldo anual medio es 195,48 cientos de euros; es decir, **19.548 euros**.

Para analizar si es representativo:

$$CV = \frac{s}{\bar{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 ni}{N} - \bar{x}^2 = \frac{3.566.875}{83} - 195{,}48^2 = 4.761{,}97 \Rightarrow s = 69{,}00$$

$$= \frac{69{,}00}{195{,}48} = 0{,}35$$

La dispersión relativa de la distribución de los sueldos es del 35 por ciento; por tanto, **la media es representativa**.

6.2. Expresar los sueldos mensuales en euros equivale a efectuar un cambio de escala, multiplicando los sueldos por 100:

$$x_i^e = 100x_i$$

$$\bar{x}^e = 100\bar{x} = 100 \times 195{,}48 = 19.548$$

Como ya hemos dicho en el apartado anterior, **el sueldo anual medio, expresado en euros, es 19.548**.

Al efectuar el cambio de escala la representatividad de la media no varía, ya que dicho cambio tiene el mismo efecto en la desviación típica que en la media y el coeficiente de variación se mantiene invariante.

Demostración:

El cambio de escala se efectúa multiplicando por una constante todos los valores de la variable; es decir:

$$x_i^e = kx_i$$

Entonces:

$$\bar{x}^e = \frac{\sum x_i^e n_i}{N} = \frac{\sum kx_i n_i}{N} = \frac{k\sum x_i n_i}{N} = k\bar{x}$$

$$s^{2e} = \frac{\sum (x_i^e - \bar{x}^e)^2 n_i}{N} = \frac{\sum (kx_i - k\bar{x})^2 n_i}{N} = \frac{k^2 \sum (x_i - \bar{x})^2 n_i}{N} = k^2 s^2 \Rightarrow s^e = ks$$

y, por tanto:

$$CV^e = \frac{s^e}{\bar{x}^e} = \frac{ks}{k\bar{x}} = CV$$

6.3. El sueldo más frecuente es la moda de la distribución, que está en el intervalo 100 – 150, que es el de frecuencia absoluta, $n_1 = 30$, más elevada. Para obtener su valor:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \times c_i = 100 + \frac{20}{0 + 20} \times 50 = 150{,}00$$

El sueldo anual más frecuente es 150 cientos de euros, o **15.000 euros**.

6.4. **La distribución es asimétrica a la derecha**, puesto que, como hemos visto, el sueldo anual medio es 195,48 cientos de euros y el sueldo más frecuente es 150,00 cientos de euros; es decir, la media es mayor que la moda.

Esto significa que en el coeficiente de asimetría de Pearson:

$$A_P = \frac{\overline{x} - M_o}{s}$$

tanto el numerador como el denominador de la fracción son positivos, de manera que también es positivo el cociente.

6.5. La primera frecuencia absoluta acumulada mayor o igual que ¼ N = ¼ 83 = 20,75 es $N_1 = 30$, luego el primer cuartil se encuentra en el intervalo de extremos 100 – 150. Para obtener su valor:

$$Q_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i = 100 + \frac{20{,}75 - 0}{30} \times 50 = 134{,}58$$

Entonces, **el sueldo anual máximo para un trabajador del 25 por ciento de los de menor retribución es** 134,58 cientos de euros; es decir, **13.458 euros**.

6.6. Si el coeficiente de exceso es negativo, el apuntamiento de la distribución es menor que el de una normal de igual media y varianza; es decir, **la distribución es platicúrtica**.

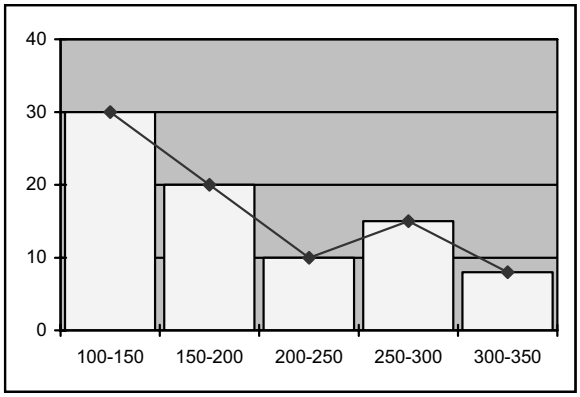


Gráfico de la distribución de los sueldos

6.7. El índice de Gini se define como:

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

siendo $p_i = 100\ F_i$, es decir, que p_i es el porcentaje de frecuencia relativa acumulada, que indica el peso que tiene cada frecuencia absoluta acumulada sobre el total de los datos, en porcentaje, y $q_i = 100\ (U_i / U_n)$, donde U_i es el valor acumulado del producto $x_i n_i$; es decir, que q_i es el valor acumulado de la variable, obtenido como suma, para cada valor, del producto $x_i n_i$ más los anteriores, expresado en porcentaje.

Para obtenerlo, entonces, hay que elaborar las siguientes tablas de cálculos intermedios:

x_i	n_i	N_i	F_i	$p_i = 100F_i$	$x_i n_i$
x_1	n_1	N_1	F_1	$p_1 = 100F_1$	$x_1 n_1$
x_2	n_2	N_2	F_2	$p_2 = 100F_2$	$x_2 n_2$
…	…	…	…	…	…
x_n	n_n	$N_n = N$	$F_n = 1$	$p_n = 100F_n = 100$	$x_n n_n$

$U_i = Ac\ (x_i n_i)$	$q_i = 100(U_i / U_n)$	$p_i - q_i$
$U_1 = x_1 n_1$	$q_1 = 100(U_1 / U_n)$	$p_1 - q_1$
$U_2 = x_1 n_1 + x_2 n_2$	$q_2 = 100(U_2 / U_n)$	$p_2 - q_2$
…	…	…
$U_n = x_1 n_1 + x_2 n_2 + \ldots + x_n n_n = \Sigma x_i n_i$	$q_n = 100(U_n / U_n) = 100$	$p_n - q_n = 0$

En cuanto a su interpretación, el índice de Gini permite conocer el grado de equidistribución de la variable. Hace referencia, por tanto, al mayor o menor grado de igualdad en el reparto del total de los valores de la variable.

Si la distribución de la variable es totalmente igualitaria, el α por ciento del total de sus valores observados corresponde al α por ciento del total de observaciones y p_i y q_i son iguales. Por el contrario, si la distribución está muy concentrada, a un porcentaje elevado del total de observaciones le corresponde sólo un porcentaje pequeño del total de los valores observados. Entonces, valores del índice de Gini próximos a cero (que implican valores de p_i aproximadamente iguales a los de q_i) indican una concentración débil y valores del índice de Gini próximos a la unidad (que implican valores de p_i elevados en relación con los de q_i) indican una fuerte concentración.

Luego si el índice de Gini tomase el valor 0,12 **la distribución de los sueldos sería bastante equitativa**.

EJERCICIO 7

Para hacer un estudio sobre el coste de alquiler (cientos de euros) de los locales destinados a discotecas en una zona turística española se dispone de los siguientes datos:

Coste ($L_{i-1} - L_i$)	Nº discotecas (n_i)
0 – 25	10
25 – 50	15
50 – 75	60
75 – 100	10
100 – 200	5

Como puede verse, **la distribución** del coste de alquiler **está agrupada en** 5 **intervalos de amplitud variable**.

A partir de esta información:

7.1. Obtenga la tabla de distribución de frecuencias.

7.2. ¿Cuál es el coste de alquiler medio? ¿Y el más frecuente? ¿Qué puede decirse respecto a la asimetría de la distribución?

7.3. ¿Es el coste medio representativo?

7.4. Obtenga el coste mediano y explique su significado.

7.5. Indique cuál es, por término medio, el coste de alquiler en cada una de estas dos situaciones:

7.5.1. Los alquileres aumentan 100 euros

7.5.2. Los alquileres aumentan un 10 por ciento

Demuéstrelo.

7.6. Calcule el primer cuartil de la distribución y explique su significado.

7.1. Tabla de distribución de frecuencias:

$L_{i-1} - L_i$	n_i	N_i	f_i	F_i
0 – 25	10	10	0,10	0,10
25 – 50	15	25	0,15	0,25
50 – 75	60	85	0,60	0,85
75 – 100	10	95	0,10	0,95
100 – 200	5	100	0,05	1,00
	100		1,00	

7.2. Para obtener el coste medio y el más frecuente se efectúan los siguientes cálculos:

$L_{i-1} - L_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$	c_i	d_i
0 – 25	10	12,5	125,0	1.562,50	25,00	0,40
25 – 50	15	37,5	562,5	21.093,75	25,00	0,60
50 – 75	60	62,5	3.750,0	234.375,00	25,00	2,40
75 – 100	10	87,5	875,0	76.562,50	25,00	0,40
100 – 200	5	150,0	750,0	112.500,00	100,00	0,05
	100		6.062,5	446.093,75		

Entonces:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{6.062,5}{100} = 60,625$$

La media del coste de alquiler es 60,625 cientos de euros; es decir, **6.062,5 euros**.

El coste de alquiler más frecuente es la moda de la distribución, que está en el intervalo de extremos 50–75, al que le corresponde la densidad de frecuencia más alta, $d_3 = 2,40$. Para obtener su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 50 + \frac{0,40}{0,60 + 0,40} \times 25 = 60$$

El coste de alquiler más frecuente es 60 cientos de euros; es decir, **6.000 euros**.

Como la media es mayor que la moda, el coeficiente de asimetría de Pearson es mayor que cero y **la distribución del coste de alquiler es asimétrica a la derecha** o positiva.

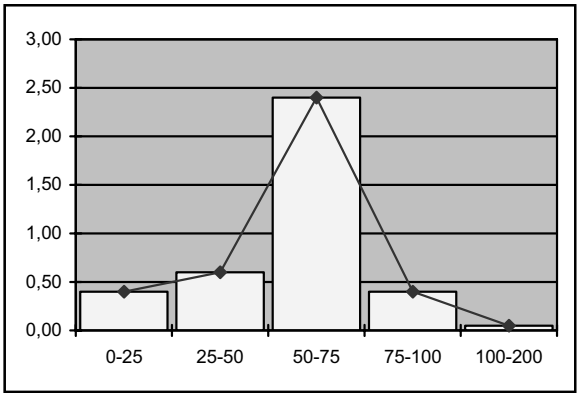


Gráfico de la distribución del coste de alquiler

7.3. Para analizar si el coste de alquiler medio es representativo, se calcula el coeficiente de variación:

$$CV = \frac{s}{\bar{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \bar{x}^2 = \frac{446.093,75}{100} - 60,625^2 = 785,546 \Rightarrow s = 28,02$$

$$= \frac{28,02}{60,625} = 0,462$$

Luego la dispersión relativa de la distribución es del 46 por ciento, y **el coste de alquiler medio es representativo**.

7.4. El coste mediano se sitúa en el intervalo de extremos 50–75, al que le corresponde el primer valor de la frecuencia absoluta acumulada igual o superior a ½ N = ½ 100 = 50, que es $N_3 = 85$. Para obtener su valor:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i = 50 + \frac{50 - 25}{60} \times 25 = 60,416$$

La mediana del coste de alquiler es, por tanto, 60,416 cientos de euros; es decir, **6.041,6 euros**. Este es el alquiler máximo para el 50 por ciento de los locales más baratos y el mínimo para el 50 por ciento de los locales más caros.

7.5.1. Un aumento de 100 euros en el coste de alquiler equivale a efectuar un cambio de origen de la variable, que consiste en sumar a todos sus valores una constante k.

En este caso, puesto que el coste de alquiler está expresado en cientos de euros, un aumento de 100 euros es un aumento de una unidad, por lo que la constante k es igual a 1. Entonces:

$$x_i^o = x_i + 1$$

$$\bar{x}^o = \bar{x} + 1 = 60,625 + 1 = 61,625$$

Igual que sucede con los valores de la variable, **el coste de alquiler medio** aumenta 100 euros, y **es** igual a 61,625 cientos de euros o **6.162,5 euros**.

7.5.2. Un aumento de un 10 por ciento en el coste de alquiler equivale a efectuar un cambio de escala de la variable, que consiste en multiplicar todos sus valores por una constante k. En este caso, la constante k es igual a 1,10 ya que si el coste de alquiler aumenta un 10 por ciento:

$$x_i^e = x_i + 0{,}10x_i = (1 + 0{,}10)x_i = 1{,}10x_i$$

$$\bar{x}^e = 1{,}10 \times \bar{x} = 1{,}10 \times 60{,}625 = 66{,}687$$

Igual que sucede con los valores de la variable, **el coste de alquiler medio** aumenta el 10 por ciento, y **es** igual a 67,787 cientos de euros o **6.778,7 euros**.

Demostración:

$$\bar{x}^o = \frac{\sum x_i^o n_i}{N} = \frac{\sum (x_i + k) n_i}{N} = \frac{\sum x_i n_i}{N} + \frac{k \sum n_i}{N} = \bar{x} + k$$

Lo que significa que la media de una variable para la que se ha efectuado un cambio de origen de k unidades es igual a la suma de la media original y las k unidades del cambio de origen.

$$\bar{x}^e = \frac{\sum x_i^e n_i}{N} = \frac{k \sum x_i n_i}{N} = k\bar{x}$$

Lo que significa que la media de una variable para la que se ha efectuado un cambio de escala de factor k, es igual al producto de dicho factor por la media de la variable original.

7.6. El primer cuartil se sitúa en el intervalo de extremos 25–50, al que le corresponde la primera frecuencia acumulada absoluta mayor o igual que ¼ N = ¼ 100 = 25, que es $N_2 = 25$. Para obtener su valor:

$$Q_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i = 25 + \frac{25 - 10}{15} \times 25 = 50$$

Luego **el precio que como máximo cuesta el alquiler de un local que forma parte del grupo del 25 por ciento de los más baratos es** 50 cientos de euros, o **5.000 euros**. Esto implica que **5.000 euros es también el precio que como mínimo cuesta el alquiler de un local que forma parte del grupo del 75 por ciento de los más caros**.

EJERCICIO 8

El número de empleados de las 30 agencias de viajes que un empresario tiene en una determinada Comunidad Autónoma presenta la siguiente distribución:

Nº Empleados ($L_{i-1} - L_i$)	Nº Oficinas (n_i)
0 – 4	3
4 – 8	11
8 – 14	12
14 – 35	4

Como puede observarse, **la distribución** del número de empleados **está agrupada en** 4 **intervalos de amplitud variable**.

8.1. Obtenga el número medio de empleados ¿Es representativo de la distribución?
8.2. Si en otras dos Comunidades Autónomas este empresario tiene 16 y 33 agencias que tienen una media de 10 y 14 empleados respectivamente ¿Cuál es la media de empleados de las oficinas en las tres Comunidades?

Tabla de cálculos intermedios

$L_{i-1} - L_i$	n_i	x_i	$x_i n_i$	$x_i^2 n_i$
0 – 4	3	2	6	12
4 – 8	11	6	66	396
8 – 14	12	11	132	1.452
14 – 35	4	24,5	98	2.401
	30		302	4.261

8.1. Entonces, el número medio de empleados es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{302}{30} = 10{,}066$$

Luego, **las agencias de viajes de este empresario en la primera Comunidad Autónoma tienen una media de 10 empleados, aproximadamente**.

Para analizar si este valor medio es representativo de la distribución, se calcula el coeficiente de variación de Pearson:

$$CV = \frac{s}{\bar{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_1^2 n_i}{N} - \bar{x}^2 = \frac{4.261}{30} - 10{,}066^2 = 40{,}708 \Rightarrow s = 6{,}38$$

$$= \frac{6{,}38}{10{,}066} = 0{,}634$$

La dispersión relativa de la distribución es del 63 por ciento; por tanto, **la media no es representativa de la distribución**.

8.2. Se sabe que en la primera Comunidad Autónoma, en la que el número de oficinas es $N_1 = 30$, la media del empleo es $\bar{x}_1 = 10{,}066$; en la segunda, en la que el número de oficinas es $N_2 = 16$, la media del empleo es $\bar{x}_2 = 10$, y en la tercera, en la que el número de oficinas es $N_3 = 33$, la media del empleo es $\bar{x}_3 = 14$. La distribución está formada por los $N = N_1 + N_2 + N_3 = 30 + 16 + 33 = 79$ datos del número de empleados en las tres Comunidades Autónomas, y el subconjunto de datos correspondiente a cada comunidad no tiene valores en común con los demás. Entonces:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N} = \frac{30 \times 10{,}066 + 16 \times 10 + 33 \times 14}{79} = 11{,}69$$

Y **la media de empleados en las oficinas de las tres Comunidades es aproximadamente igual a 12**.

EJERCICIO 9

El valor más frecuente de una variable x es 5,7 y el coeficiente de exceso de su distribución, que es campaniforme, es igual a 7,2. Se sabe, además, que:

$\sum_{i=1}^{n}(x_i - k)^2 n_i$ toma su valor mínimo cuando k es igual a 5,5

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N}} = 1,5$$

Dada esta información, ¿es la media aritmética representativa de la distribución? ¿Puede decirse que la distribución es asimétrica a la izquierda? ¿Y que es leptocúrtica? ¿Por qué? Indique claramente el significado de estas expresiones.

El valor medio de la variable es 5,5 ya que, como sabemos:

$$\sum_{i=1}^{n}(x_i - k)^2 n_i$$

es mínima cuando k es igual a $\bar{x}$.

Se conoce, también, su desviación típica:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{N}} = 1,5$$

Entonces:

$$CV = \frac{s}{\bar{x}} = \frac{1,5}{5,5} = 0,272$$

Es decir, la dispersión relativa es del 27 por ciento, y **la media es representativa de la distribución**.

La desviación típica de cualquier variable es positiva y, en este caso, el valor medio de la variable (= 5,5) es menor que su valor más frecuente (= 5,7). Entonces, el coeficiente de asimetría de Pearson:

$$A_P = \frac{\overline{x} - M_o}{S_X}$$

es negativo y **la distribución es**, efectivamente, **asimétrica a la izquierda**. Esto significa que **la distribución tiene cola a la izquierda**; es decir, que las frecuencias descienden más lentamente por la izquierda que por la derecha.

Respecto a la curtosis, se sabe que el coeficiente de exceso:

$$g = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4 n_i \ / \ N}{s^4} - 3$$

(= 7,2) es mayor que cero. El primer sumando de esta expresión (= g + 3 = 7,2 + 3 = 10,2) mide el apuntamiento correspondiente a esta distribución y el segundo (= 3) es el apuntamiento para una distribución normal con la misma media y varianza que esta variable. Luego el apuntamiento de esta distribución es mayor que el de la normal y **la distribución es**, efectivamente, **leptocúrtica**.

EJERCICIO 10

Se dispone de la siguiente información respecto a los beneficios (miles de euros) que los establecimientos de una cadena hotelera obtienen por servicio de restaurante:

Beneficios ($L_{i-1} - L_i$)	Nº hoteles (n_i)
0 – 300	32
300 – 600	16
600 – 1.200	8

Se sabe, además, que el beneficio medio es 342,85 miles de euros y la desviación típica es 262,44 miles de euros.

La distribución de los beneficios **está agrupada en** 3 **intervalos de amplitud variable**.

10.1. Calcule los beneficios que como máximo obtuvieron el 50 por ciento de los servicios de restaurante menos rentables.
10.2. Indique cuáles son los beneficios que como mínimo y cómo máximo han obtenido el 75 por ciento de los servicios de restaurante más rentables.
10.3. Se sabe que el número índice del año 2004 de los beneficios en términos nominales con respecto al año 1999 es 107,42. ¿Cuál es, entonces, la media de los beneficios obtenidos por los servicios de restaurante en el año 2004? ¿Y la desviación típica? Justifique su respuesta.
10.4. Sabiendo que al 57,14 por ciento de los servicios de restaurante les corresponde el 25 por ciento de los beneficios totales y que al 85,71 por ciento de dichos servicios les corresponde el 62,5 por ciento de los beneficios, ¿puede decirse que la distribución está muy concentrada?

Tabla de cálculos intermedios

$L_{i-1} - L_i$	n_i	N_i
0 – 300	32	32
300 – 600	16	48
600 – 1.200	8	56
	56	

10.1. La primera frecuencia acumulada absoluta igual o superior a ½ N = ½ 56 = 28 es $N_1 = 32$, luego la mediana está en el intervalo de extremos 0 – 300. Para obtener su valor:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times c_i = 0 + \frac{28 - 0}{32} \times 300 = 262{,}5$$

El beneficio máximo obtenido por el 50 por ciento de los servicios de restaurante menos rentables es 262,5 miles de euros; es decir, **262.500 euros**.

10.2. Puesto que la mediana está en el primer intervalo de la distribución, también está en dicho intervalo el primer cuartil. Para obtener su valor:

$$Q_1 = L_{i-1} + \frac{\frac{N}{4} - N_{i-1}}{n_i} \times c_i = 0 + \frac{14-0}{32} \times 300 = 131{,}25$$

El beneficio que como mínimo obtiene un servicio de restaurante del 75 por ciento de los más rentables es 131,25 miles de euros, o **131.250 euros**. **El máximo** beneficio que dicho servicio puede obtener es el valor máximo posible de los beneficios, que **es** 1.200 miles de euros o **1.200.000 euros**.

10.3. El número índice del momento t–ésimo respecto al período base (I_t^0) se obtiene multiplicando por cien el cociente entre el valor corriente (x_t) y el que la variable toma en el período de referencia (x_0); es decir:

$$I_t^0 = \frac{x_t}{x_0} \times 100$$

Entonces:

$$I_{04}^{99} = \frac{x_{04}}{x_{99}} \times 100 \Rightarrow x_{04} = \frac{I_{04}^{99} \times x_{99}}{100} = \frac{I_{04}^{99}}{100} \times x_{99}$$

Es decir, los beneficios en el año 2004 se obtienen multiplicando sus valores en el año 1999 por el índice expresado en tanto por uno, lo que equivale a efectuar un cambio de escala de factor k = I_{04}^{99} / 100 = 1,0742:

$$\bar{x}^e = k\bar{x} = 1{,}0742 \times 342{,}85 = 368{,}29$$

$$s^e = ks = 1{,}0742 \times 262{,}44 = 281{,}91$$

El beneficio medio y la desviación típica de los beneficios en el año 2004 son 368,29 y 281,91 miles de euros; o **368.290 y 281.910 euros respectivamente**.

10.4. El grado de concentración de la distribución puede analizarse calculando el índice de Gini. La información disponible se ha recogido en la tabla siguiente:

p_i	q_i	$p_i - q_i$
57,14	25	32,14
85,71	62,5	23,21
100	100	0

El índice de Gini, resulta:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{55,35}{142,85} = 0,387$$

Y su valor indica que **la distribución de los beneficios es bastante desigual**, ya que, como puede verse en la tabla, el 57 por ciento de los restaurantes se reparten sólo la cuarta parte del total de beneficios, mientras que el 38 por ciento de los beneficios le corresponden al 15 por ciento de los restaurantes.

EJERCICIO 11

La distribución del número de habitaciones por establecimiento en los alojamientos de una determinada localidad es la que se presenta en la siguiente tabla:

Nº habitaciones ($L_{i-1} - L_i$)	Nº establecimientos (n_i)
0 – 18	112
18 – 28	679
28 – 40	833
40 – 50	449
50 – 65	303
65 – 95	120

11.1. Dado que en esta distribución la media y la moda coinciden puede decirse que es simétrica. ¿Es correcta esta afirmación? Justifique su respuesta.
11.2. ¿Qué porcentaje de establecimientos tienen entre 20 y 30 habitaciones?

La distribución del número de habitaciones **está agrupada en** 6 **intervalos de amplitud variable**.

Tabla de resultados intermedios

$L_{i-1} - L_i$	x_i	n_i	$x_i n_i$	N_i	c_i	d_i
0 – 18	9,0	112	1.008,0	112	18	6,22
18 – 28	23,0	679	15.617,0	791	10	67,90
28 – 40	34,0	833	28.322,0	1.624	12	69,42
40 – 50	45,0	449	20.205,0	2.073	10	44,90
50 – 65	57,5	303	17.422,5	2.376	15	20,20
65 – 95	80,0	120	9.600,0	2.496	30	4,00
		2.496	92.174,5			

11.1. Dada la información que contiene la tabla, la media de la variable es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{92.174,5}{2.496} = 36,93$$

La moda está en el intervalo de extremos 28–40, puesto que $d_3 = 69,42$ es la mayor densidad de frecuencia. Para obtener su valor:

$$M_o = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \times c_i = 28 + \frac{44,90}{44,90 + 67,90} \times 12 = 32,78$$

La media de habitaciones por establecimiento es 36,93 y el número de habitaciones por establecimiento más frecuente es 32,78. **La afirmación no es correcta**. Al ser la media mayor que la moda, el coeficiente de asimetría de Pearson es positivo, y **la distribución es asimétrica a la derecha**.

11.2. La densidad de frecuencia en el intervalo de extremos 18–28 es 67,9 y los establecimientos que tienen entre 20 y 28 habitaciones son 9 x 67,9 = 611,1. De la misma forma se deduce que los que tienen entre 28 y 30 son 138,84.

Por tanto, **los establecimientos que tienen entre 20 y 30 habitaciones** son 611,1 + 138,84 = 749,94, que **representan el 30,04** (= (749,94 / 2.496) x 100) **por ciento del total**.

EJERCICIO 12

Los gastos de apertura (miles de euros) de un restaurante fueron presupuestados por cinco evaluadores distintos, que proporcionaron las siguientes respuestas:

Evaluador	Presupuesto
1	30
2	37
3	25
4	31
5	40

Para resumir esta información, se asumió como presupuesto el valor medio de los cinco propuestos.

12.1. ¿Puede considerarse dicho valor representativo de la distribución?
12.2. Teniendo en cuenta que a cada una de las valoraciones deben añadirse unos gastos fijos de 10 mil euros, ¿cuál es el presupuesto definitivo? ¿Varía la dispersión relativa de la distribución? Demuéstrelo.
12.3. Se dispone de la siguiente información adicional respecto a los evaluadores:

Evaluador	Años de experiencia
1	7
2	15
3	3
4	6
5	14

¿Cuál sería el presupuesto definitivo si en la media se ponderase la importancia de cada uno de ellos en función de los años de experiencia del evaluador?

La distribución de los presupuestos **no está agrupada**, y **la frecuencia de cada valor de la variable es igual a la unidad**.

Tabla de cálculos intermedios

Evaluador	x_i	w_i	x_i^2	$x_i w_i$
1	30	7	900	210
2	37	15	1.369	555
3	25	3	625	75
4	31	6	961	186
5	40	14	1.600	560
	163	45	5.455	1.586

12.1. La media de la distribución es:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{163}{5} = 32{,}6$$

Luego **el presupuesto asciende a** 32,6 miles de euros o **32.600 euros**.

Para analizar su representatividad:

$$CV = \frac{s}{\overline{x}} =$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \overline{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{N} - \overline{x}^2 = \frac{5.455}{5} - 32{,}6^2 = 28{,}24 \Rightarrow s = 5{,}31$$

$$= \frac{5{,}31}{32{,}6} = 0{,}16$$

La dispersión relativa de la distribución es del 16 por ciento, y **la media es representativa**.

12.2. Añadir a cada valor del presupuesto unos gastos fijos de 10 mil euros (= 10 unidades) equivale a cambiar el origen de la variable. Luego:

$$\bar{x}^o = \bar{x} + k = 10 + 32{,}6 = 42{,}6$$

El presupuesto definitivo asciende a 42,6 miles de euros o **42.600 euros**.

La representatividad si **varía**, porque el cambio de origen modifica la media pero no afecta a la varianza, de manera que el coeficiente de variación es distinto.

Demostración:

El cambio de origen supone sumar a los valores de la variable una constante k, de manera que $x_i^o = x_i + k$. Entonces:

$$\bar{x}^o = \frac{\sum x_i^o n_i}{N} = \frac{\sum (x_i + k) n_i}{N} = \frac{\sum (x_i n_i + k n_i)}{N} = \frac{\sum x_i n_i}{N} + k\frac{\sum n_i}{N} = \bar{x} + k$$

$$s^{2o} = \frac{\sum_{i=1}^{n} (x_i^o - \bar{x}_i)^2 n_i}{N} = \frac{\sum_{i=1}^{n} (x_i + k - \bar{x} - k)^2 n_i}{N} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 n_i}{N} = s^2 \Rightarrow s^o = s$$

$$CV^o = \frac{s^o}{\bar{x}^o} = \frac{s}{\bar{x} + k} \neq \frac{s}{\bar{x}} = CV$$

12.3. Ponderando la importancia de cada presupuesto en función de los años de experiencia del evaluador:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{1.586}{45} = 35{,}24$$

Y teniendo en cuenta los gastos fijos:

$$\bar{x}_w^o = \bar{x}_w + k = 10 + 35{,}24 = 45{,}24$$

El presupuesto definitivo es de 45,24 miles de euros o **45.240 euros**.

EJERCICIO 13

En dos establecimientos de una cadena hotelera se ha realizado un estudio sobre el número de pernoctaciones diarias registradas. En el primero se ha obtenido una media igual a 125 y una varianza de 36. En el segundo la media es 285 y la desviación típica es igual a 46. ¿Cuál de las dos distribuciones es más homogénea?

Aunque las pernoctaciones están dadas en las mismas unidades, las medias de las dos distribuciones son diferentes; por tanto, no es correcto comparar la dispersión por medio de las desviaciones típicas. Deben calcularse los correspondientes coeficientes de variación de Pearson, para valorar las dispersiones relativas:

$$CV_1 = \frac{s_1}{\bar{x}_1} = \frac{6}{125} = 0{,}048$$

$$CV_2 = \frac{s_2}{\bar{x}_2} = \frac{46}{285} = 0{,}160$$

Luego **es más homogénea la distribución de las pernoctaciones en el primer establecimiento,** puesto que su dispersión relativa es menor.

EJERCICIO 14

¿Qué indica el hecho de que en una distribución el valor de la media sea superior al de la mediana?

Que en la distribución hay algún o algunos valores extremos, significativamente superiores a los demás.

Puesto que en el cálculo de la media intervienen todos los valores de la variable, si entre ellos hay alguno o algunos muy elevados, su valor resulta superior al que ocupa el lugar central de la distribución.

EJERCICIO 15

Respecto a dos distribuciones campaniformes y simétricas se dispone de la siguiente información:

Distribución A: $M_e = 15$ $s^2 = 36$
Distribución B: $M_o = 20$ $s = 6$

15.1. Obtenga la media aritmética de cada distribución.
15.2. ¿Sería correcto utilizar la desviación típica para comparar la dispersión de las distribuciones? ¿Por qué?
15.3. Indique cuál de las dos distribuciones presenta mayor variabilidad.

15.1. Al ser las distribuciones campaniformes y simétricas, la media, la mediana y la moda coinciden.

Luego **la media de la primera distribución es igual a 15 y la de la segunda es igual a 20**.

15.2. **No sería correcto**. Sólo es correcto comparar la dispersión de las distribuciones utilizando la desviación típica si ambas tienen la misma media y están expresadas en las mismas unidades de medida y, en este caso, las distribuciones tienen medias diferentes. Para hacer la comparación, deben obtenerse los coeficientes de variación de Pearson.

15.3. Efectuando los cálculos:

$$CV_A = \frac{s_A}{\bar{x}_A} = \frac{6}{15} = 0{,}4$$

$$CV_B = \frac{s_B}{\bar{x}_B} = \frac{6}{20} = 0{,}3$$

La dispersión relativa de la distribución A es del 40 por ciento y la de la distribución B es del 30 por ciento. **Tiene mayor variabilidad la distribución A**.

En este caso particular, puede llegarse a la misma conclusión sin necesidad de calcular el coeficiente de variación.

Las desviaciones típicas de las dos distribuciones son iguales, pero la distribución A tiene una media menor que la distribución B; por tanto, el coeficiente de variación de las dos distribuciones tiene el mismo numerador y el denominador de la fracción es menor para la distribución A que para la B, o lo que es lo mismo, es menor en el caso de la distribución B. No es necesario conocer su valor para deducir que presenta mayor variabilidad la distribución A.

EJERCICIO 16

En una empresa turística los empleados se clasifican en cinco categorías para las cuales se dispone de datos respecto a los salarios anuales (cientos de euros) pactados en el convenio colectivo y al número de empleados.

Categoría	Salario (x_i)	N^o empleados (n_i)
1	100	60
2	120	40
3	180	100
4	230	20
5	260	5

Calcule el índice de Gini e interprete el resultado obtenido.

La distribución de los salarios **no está agrupada**.

Tabla de cálculos intermedios

x_i	n_i	N_i	F_i	p_i	$x_i n_i$	U_i	q_i	$p_i - q_i$
100	60	60	0,26	26,66	6.000	6.000	17,29	9,37
120	40	100	0,44	44,44	4.800	10.800	31,12	13,32
180	100	200	0,88	88,88	18.000	28.800	82,99	5,89
230	20	220	0,97	97,77	4.600	33.400	96,25	1,52
260	5	225	1,00	100,00	1.300	34.700	100,00	0,00
					34.700			

El índice de Gini es:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{30{,}10}{257{,}75} = 0{,}116$$

Luego, como puede observarse en la tabla de cálculos intermedios, **la distribución de los salarios es bastante equitativa** o no está muy concentrada.

EJERCICIO 17

Con los datos de las pernoctaciones registradas en establecimientos hoteleros en las Comunidades Autónomas de Canarias y de Galicia correspondientes a los 12 meses del año 2003 se han obtenido unos valores del índice de Gini iguales a 0,041 y 0,674 respectivamente. ¿Qué podría concluir a la vista de estos resultados?

En el año 2003, **en Canarias, las pernoctaciones están prácticamente equidistribuidas a lo largo del año**, puesto que el índice de Gini está próximo a cero. **Sin embargo, en Galicia, el grado de concentración de las pernoctaciones es muy elevado**, ya que el índice de Gini está más próximo a la unidad que a cero.

EJERCICIO 18

Indique qué información proporciona el índice de Gini respecto a la distribución de la inversión (miles de euros) en los establecimientos de una cadena hotelera que se recoge en la tabla siguiente:

Inversión ($L_{i-1} - L_i$)	Nº establecimientos (n_i)
7 – 9	10
9 – 13	30
13 – 15	40
15 – 21	15
21 – 31	5

Tabla de resultados intermedios

$L_{i-1} - L_i$	n_i	N_i	F_i	p_i	x_i	$x_i n_i$	U_i	q_i	$p_i - q_i$
7 – 9	10	10	0,10	10,00	8	80	80	5,84	4,16
9 – 13	30	40	0,40	40,00	11	330	410	29,93	10,07
13 – 15	40	80	0,80	80,00	14	560	970	70,80	9,20
15 – 21	15	95	0,95	95,00	18	270	1.240	90,51	4,49
21 – 31	5	100	1,00	100,00	26	130	1.370	100,00	0,00
	100					1.370			

El índice de Gini es:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{27,92}{225} = 0,12$$

Luego su valor está próximo a cero y **el reparto de la inversión es bastante igualitario**.

EJERCICIO 19

Al analizar el gasto en hoteles en una determinada zona se ha observado que, una vez ordenado en orden creciente, al primer 30 por ciento de los clientes les corresponde el 15 por ciento del gasto total, mientras que el 90 por ciento de los clientes gasta el 75 por ciento del total. Valore el grado de concentración de la distribución utilizando el índice de Gini.

Tabla de cálculos intermedios

p_i	q_i	$p_i - q_i$
30	15	15
90	75	15
100	100	0

El índice de Gini es:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{30}{120} = 0{,}25$$

Su valor no está alejado de cero y **la distribución del gasto no está muy concentrada**.

EJERCICIO 20

Se dispone de la información relativa al número de plazas de los hoteles que se ubican en una determinada Comunidad Autónoma que figura en la siguiente tabla:

$x_i n_i$	f_i
175	0,2500
300	0,1429
625	0,1785
700	0,1429
1.250	0,1786
1.350	0,1071

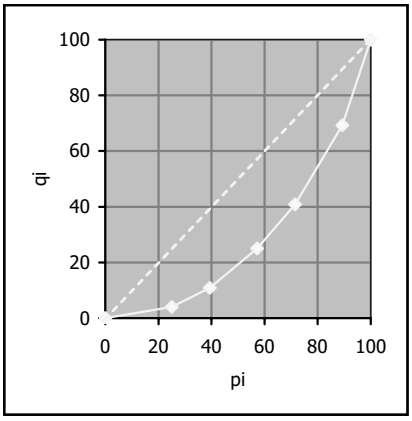
Represente la curva de Lorenz y obtenga el índice de Gini explicando con claridad qué información proporcionan.

Tabla de cálculos intermedios

$x_i n_i$	f_i	F_i	p_i	U_i	q_i	$p_i - q_i$
175	0,25	0,25	25,00	175	3,98	21,02
300	0,1429	0,39	39,29	475	10,80	28,49
625	0,1785	0,57	57,14	1.100	25,00	32,14
700	0,1429	0,71	71,43	1.800	40,91	30,52
1.250	0,1786	0,89	89,29	3.050	69,32	19,97
1.350	0,1071	1,00	100,00	4.400	100,00	0,00
4.400						

La curva de Lorenz y el índice de Gini permiten valorar el grado de concentración de la distribución.

La curva de Lorenz se inserta en un cuadrado en el que la diagonal que une el vértice inferior izquierdo con el superior derecho representa la mínima concentración, ya que en todos sus puntos se cumple que el porcentaje que representa la frecuencia absoluta acumulada sobre el total de los datos es igual al porcentaje que supone el valor acumulado de la variable sobre su valor total ($p_i = q_i$). Cuanto más elevada es la concentración, más pequeño es el porcentaje del valor acumulado de la variable sobre su valor total (q_i) que le corresponde a un porcentaje elevado de la frecuencia absoluta acumulada sobre el total de los datos (p_i), de manera que cuando la concentración es muy fuerte, la curva de Lorenz tiende a confundirse con los lados inferior e izquierdo del cuadrado.



Curva de Lorenz de la distribución del número de plazas

En este caso, la curva de Lorenz no está próxima a la diagonal, luego el reparto del número de plazas no es igualitario. Como muestra la tabla de cálculos intermedios, a la cuarta parte de los hoteles les corresponde sólo algo menos del 4 por ciento del total de las plazas, mientras que el 30 por ciento del total de las plazas se concentra en el 10 por ciento de los hoteles.

El índice de Gini es el doble del área comprendida entre la curva de Lorenz y la diagonal del cuadrado, bajo el supuesto convencional de que el área del cuadrado es igual a la unidad. Su valor está comprendido entre 0 y 1, puesto que si la curva

coincide con la diagonal el área vale cero, y si coincide con los lados del cuadrado es igual a ½. Dado que la concentración es mínima cuando la curva se superpone a la diagonal y máxima cuando la curva se superpone a los lados del cuadrado, el índice de Gini indica igualdad en el reparto si su valor está próximo a cero, y a medida que se aleja de cero indica un grado de concentración cada vez mayor.

Para obtenerlo:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{132{,}14}{282{,}15} = 0{,}468$$

En este caso, el **índice de Gini** está alejado de cero, luego es coherente con el resultado que, en términos gráficos, muestra la curva de Lorenz, **e indica que el reparto del número de plazas es poco igualitario**.

EJERCICIO 21

¿Varía la representatividad de la media aritmética de una distribución si se multiplican todos los valores de la variable por una constante k? Demuéstrelo.

No, **la representatividad de la media no varía**, porque el cambio de escala tiene el mismo efecto en la media que en la desviación típica, de manera que el coeficiente de variación se mantiene invariante.

Demostración:

$$x_i^e = kx_i$$

$$\bar{x}^e = \frac{\sum x_i^e n_i}{N} = \frac{k\sum x_i n_i}{N} = k\bar{x}$$

$$s^{2e} = \frac{\sum (x_i^e - \bar{x}^e)^2 n_i}{N} = \frac{\sum (kx_i - k\bar{x})^2 n_i}{N} = \frac{k^2 \sum (x_i - \bar{x})^2 n_i}{N} = k^2 s^2 \Rightarrow s^e = ks$$

Luego:

$$CV^e = \frac{s^e}{\bar{x}^e} = \frac{ks}{k\bar{x}} = CV$$

EJERCICIO 22

Si consideramos una transformación lineal de la variable x_i tal como $z_i = a + bx_i$ donde a y b son constantes, $\bar{z} = a + b\bar{x}$ ¿Es correcta esta afirmación? Demuéstrelo.

La afirmación es correcta.

Demostración:

$$\bar{z} = \frac{\sum z_i n_i}{N} = \frac{\sum (a + bx_i) n_i}{N} = \frac{a\sum n_i}{N} + b\frac{\sum x_i n_i}{N} = a + b\bar{x}$$

EJERCICIO 23

Si todos los valores de una variable se incrementan en una constante k, no se modifica ni la media de la distribución ni su representatividad. ¿Es correcta esta afirmación? Demuéstrelo.

La afirmación es falsa, porque el cambio de origen modifica la media y su representatividad, puesto que no afecta a la desviación típica pero sí modifica el coeficiente de variación.

Demostración:

$$x_i^o = x_i + k$$

$$\bar{x}^o = \frac{\sum x_i^o n_i}{N} = \frac{\sum (x_i + k) n_i}{N} = \frac{\sum (x_i n_i + k n_i)}{N} = \frac{\sum x_i n_i}{N} + k\frac{\sum n_i}{N} = \bar{x} + k$$

$$s^{2o} = \frac{\sum_{i=1}^{n}(x_i^o - \overline{x}_i)^2 n_i}{N} = \frac{\sum_{i=1}^{n}(x_i + k - \overline{x} - k)^2 n_i}{N} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 n_i}{N} = s^2 \Rightarrow s^{2o} = s$$

Entonces:

$$CV^o = \frac{s^o}{\overline{x}^o} = \frac{s}{\overline{x} + k} \neq \frac{s}{\overline{x}} = CV$$

EJERCICIO 24

De la distribución de frecuencias de la variable x sabemos que:

$\frac{\sum_{i=1}^{n}(x_i - k)^2 n_i}{N}$ es mínima cuando k = 48

¿Quiere decir esto que la desviación típica de la variable es 48? ¿Por qué?

No. Quiere decir que la media de la variable es 48, porque una de las propiedades de la media aritmética es que hace mínima la suma de los cuadrados de las desviaciones de los valores de la variable respecto a una constante.

EJERCICIO 25

En un hotel se clasifica a los trabajadores en dos categorías según su vinculación con el departamento de administración. El sueldo medio de la empresa es 170 unidades. Los empleados que trabajan en el departamento de administración cobran por término medio 250 unidades. El sueldo medio de los que prestan sus servicios en los restantes departamentos es de 130 unidades.

¿Qué porcentaje de trabajadores desempeñan su labor en el departamento de administración?

Se sabe que en el departamento de administración, que tiene N_1 empleados el sueldo medio es $\bar{x}_1 = 250$, y en los demás, en los que hay N_2 empleados es $\bar{x}_2 = 130$. Se conoce, además, el sueldo medio de la empresa $\bar{x} = 170$, que tiene un total de N = N_1 + N_2 empleados. El subconjunto de datos correspondiente al departamento de administración y el correspondiente a los demás no tienen valores en común.

Entonces:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N} =$$

y, teniendo en cuenta que $N = N_1 + N_2 \Rightarrow N_2 = N - N_1$, se tiene:

$$= \frac{N_1\bar{x}_1 + (N - N_1)\bar{x}_2}{N}$$

Sustituyendo:

$$170 = \frac{250N_1 + 130(N - N_1)}{N}$$

$$170N = 250N_1 + 130N - 130N_1$$

$$40N = 120N_1$$

$$N_1 = \frac{4}{12}N = 0{,}33N$$

Y el porcentaje de empleados en el departamento de administración es:

$$P_1 = \frac{N_1}{N} \times 100 = \frac{0{,}33N}{N} \times 100 = 33$$

En el departamento de administración trabaja el 33 por ciento de los empleados de la empresa.

EJERCICIO 26

En una cadena hotelera los empleados se clasifican en tres categorías para las cuales se dispone de la siguiente información correspondiente al año 1999:

Categoría	Nº empleados	Salario medio	Desviación típica
1	20	300	70
2	50	145	22,5
3	130	156	42

26.1. Calcule el salario medio del conjunto de los empleados.

26.2. ¿En que categoría es más homogéneo el salario?

26.3. Para fijar los salarios del año 2000 se proponen dos alternativas:

26.3.1. Un incremento lineal de 10 unidades para todos los trabajadores.

26.3.2. Un aumento del 10 por ciento en los salarios de los empleados de la categoría 3, del 8 por ciento en los de los empleados de la categoría 2 y del 4 por ciento en los de los empleados de la categoría 1.

Obtenga los salarios medios por categorías y para el total de empleados en cada una de las dos situaciones y comente los resultados obtenidos.

26.1. Se sabe que en la primera categoría, que tiene $N_1 = 20$ empleados, el salario medio es $\overline{x}_1 = 300$, en la segunda, que tiene $N_2 = 50$ empleados, el salario medio es $\overline{x}_2 = 145$ y en la tercera categoría, que tiene $N_3 = 130$ empleados, el salario medio es $\overline{x}_3 = 156$.

La cadena hotelera tiene un total de $N = N_1 + N_2 + N_3 = 20 + 50 + 130 = 200$ empleados, y el subconjunto de datos correspondiente a cada categoría no tiene valores en común con los demás.

Entonces:

$$\overline{x} = \frac{\overline{x}_1 N_1 + \overline{x}_2 N_2 + \overline{x}_3 N_3}{N} = \frac{300 \times 20 + 145 \times 50 + 156 \times 130}{200} = 167{,}65$$

Es decir, **el salario medio del conjunto de los empleados es 167,65 unidades**.

26.2. Calculando el coeficiente de variación de Pearson en cada categoría:

$$CV_1 = \frac{s_1}{\bar{x}_1} = \frac{70}{300} = 0{,}23$$

$$CV_2 = \frac{s_2}{\bar{x}_2} = \frac{22{,}5}{145} = 0{,}15$$

$$CV_3 = \frac{s_3}{\bar{x}_3} = \frac{42}{156} = 0{,}27$$

Luego **el salario es más homogéneo en la segunda**, que tiene menor dispersión relativa.

26.3.1. Para obtener las nuevas medias del salario en cada categoría se efectúa un cambio de origen, y los resultados que se obtienen son los siguientes:

Categoría 1: $\bar{x}_1^o = \bar{x}_1 + 10 = 300 + 10 = 310$
Categoría 2: $\bar{x}_2^o = \bar{x}_2 + 10 = 145 + 10 = 155$
Categoría 3: $\bar{x}_3^o = \bar{x}_3 + 10 = 156 + 10 = 166$

Y para el total de empleados, se tiene:

$$\bar{x}^o = \frac{\bar{x}_1^o N_1 + \bar{x}_2^o N_2 + \bar{x}_3^o N_3}{N} = \frac{310 \times 20 + 155 \times 50 + 166 \times 130}{200} = 177{,}65$$

26.3.2. Para obtener las nuevas medias del salario en cada categoría se efectúa un cambio de escala, y los resultados que se obtienen son los siguientes:

Categoría 1: $\bar{x}_1^e = k_1 \bar{x}_1 = 1{,}04 \times 300 = 312$
Categoría 2: $\bar{x}_2^e = k_2 \bar{x}_2 = 1{,}08 \times 145 = 156{,}6$
Categoría 3: $\bar{x}_3^e = k_3 \bar{x}_3 = 1{,}10 \times 156 = 171{,}6$

Y para el total de empleados, se tiene:

$$\bar{x}^e = \frac{\bar{x}_1^e N_1 + \bar{x}_2^e N_2 + \bar{x}_3^e N_3}{N} = \frac{312 \times 20 + 156{,}6 \times 50 + 171{,}6 \times 130}{200} = 181{,}89$$

Luego, por término medio, **tanto para cada categoría como para el conjunto de los empleados, es preferible la segunda alternativa**.

EJERCICIO 27

Disponemos de las series de números índices en base 1995 que reflejan la evolución del número de plazas registradas en diferentes medios de acomodación en Galicia a lo largo del período 1995–1997:

Años	Hoteles	Hostales	Campings	Apartamentos	E. turismo rural
1995	100,0	100,0	100,0	100,0	100,0
1996	106,1	99,8	101,3	106,2	130,2
1997	107,4	100,0	104,9	110,0	160,7

27.1. Puesto que en 1996 el índice para los hostales es 99,8, las plazas registradas en este tipo de alojamientos han experimentado desde 1995 un crecimiento del 99,8 por ciento ¿Es correcta esta afirmación? Razone su respuesta.
27.2. Construya una serie de números índices complejos no ponderados que refleje la evolución de las plazas ofertadas en el conjunto de los medios de acomodación utilizando la media aritmética simple.
27.3. Exprese la serie de números índices complejos en base 1997.

27.1. **La afirmación es falsa**. Dada su definición, **el valor del índice en el año base es igual a 100**. En los períodos en los que es superior, la variable ha experimentado desde el período base un crecimiento en un porcentaje igual a la cantidad que excede de 100 **y en los períodos en los que es inferior, la variable ha experimentado desde el período base una disminución en un porcentaje igual a la cantidad que falta hasta 100**. Por tanto, las plazas registradas en hostales han experimentado desde 1995 una disminución del 0,2 por ciento.

27.2. La serie de números índices complejos no ponderados para el conjunto de los medios de acomodación se obtiene calculando la media aritmética simple de los índices parciales.

El resultado obtenido figura en la tabla siguiente:

Años	Hoteles	Hostales	Campings	Apartamentos	E. turismo rural	**I. complejo**
1995	100,0	100,0	100,0	100,0	100,0	100,00
1996	106,1	99,8	101,3	106,2	130,2	108,72
1997	107,4	100,0	104,9	110,0	160,7	116,60

27.3. Para efectuar un cambio de base se divide cada valor de la serie de números índices entre el que toma en el que se va a considerar como nuevo período de referencia.

En este caso, dividiendo la serie en base 1995 por 116,6, que es el valor que el índice toma en 1997:

Años	Base 1995	**Base 1997**
1995	100,0	85,76
1996	108,7	93,22
1997	116,6	100,00

EJERCICIO 28

La siguiente tabla recoge el número de clientes (miles) de una empresa hostelera entre el primer trimestre de 1993 y el cuarto de 1996.

Trimestres	1993	1994	1995	1996
I	5	7	8	10
II	15	16	18	20
III	30	32	35	38
IV	10	11	11	12

Calcule los índices de variación estacional suponiendo que la serie sigue un esquema multiplicativo y obtenga la serie desestacionalizada.

Para calcular los índices de variación estacional se utiliza el método de la razón a la media móvil. Como la serie es trimestral, se obtienen las medias móviles de orden p = 4 descentradas, $MM4_d$, y centradas, $MM4_c$. Dividiendo la serie de clientes entre la de medias móviles se determinan los índices de variación estacional, IVE.

Los índices medios de variación estacional, IME, que se obtienen promediando los de variación estacional, se ajustan para que su suma sea igual a 400 y así resulta la serie IMEA. Por último, para obtener la serie desestacionalizada se divide la serie de clientes entre la de índices ajustados.

Tanto los resultados intermedios de este proceso como la serie desestacionalizada, SD, figuran en la siguiente tabla:

Período	y	$MM4_d$	$MM4_c$	IVE	IME	IMEA	**SD**
93.1	5				46,91	46,63	10,72
93.2	15				99,43	98,83	15,18
		15,00					
93.3	30		15,25	196,72	193,64	192,47	15,59
		15,50					
93.4	10		15,63	63,98	62,45	62,07	16,11
		15,75					
94.1	7		16,00	43,75	46,91	46,63	15,01
		16,25					
94.2	16		16,38	97,68	99,43	98,83	16,19
		16,50					
94.3	32		16,63	192,42	193,64	192,47	16,63
		16,75					
94.4	11		17,00	64,71	62,45	62,07	17,72
		17,25					
95.1	8		17,63	45,38	46,91	46,63	17,16
		18,00					
95.2	18		18,00	100,00	99,43	98,83	18,21
		18,00					
95.3	35		18,25	191,78	193,64	192,47	18,18
		18,50					
95.4	11		18,75	58,67	62,45	62,07	17,72
		19,00					
96.1	10		19,38	51,60	46,91	46,63	21,45
		19,75					
96.2	20		19,88	100,60	99,43	98,83	20,24
		20,00					
96.3	38				193,64	192,47	19,74
96.4	12				62,45	62,07	19,33

EJERCICIO 29

Los índices de variación estacional de la serie trimestral del número de personas que han solicitado información en una Oficina de Información Turística son los siguientes:

Trimestres	IVE
I	47,13
II	102,28
III	199,78
IV	50,73

¿Significa esto que en el segundo trimestre del año se concentra el 102,28 por ciento de las personas que solicitan información en dicha Oficina?

No. El índice de variación estacional indica el porcentaje que el valor de la variable representa en cada trimestre respecto a la media anual. Por tanto, **en el segundo trimestre del año el número de personas que solicitan información es un 2,28 por ciento superior a la media del año**.

EJERCICIO 30

Se dispone de la siguiente información relativa a las variables x = ingresos (cientos de euros) diarios por servicio de restaurante e y = ingresos (cientos de euros) diarios por servicio de bar correspondientes a diez hoteles de la misma categoría:

x_t	y_t
120	35
95	22
100	25
85	18
112	30
90	21
130	41
115	32
100	31
92	22

30.1. Obtenga el coeficiente de correlación lineal entre los ingresos diarios por servicio de restaurante y por servicio de bar y explique con claridad qué información proporciona.

30.2. ¿Son estas variables independientes desde el punto de vista estadístico?

30.3. Estime por mínimos cuadrados ordinarios la ecuación $y_t = \alpha + \beta x_t$ y explique el significado de la pendiente de la recta de regresión estimada.

30.4. Obtenga el coeficiente de determinación e interprete el resultado obtenido.

Tabla de cálculos intermedios

x_t	y_t	x_t^2	y_t^2	$x_t y_t$
120	35	14.400	1.225	4.200
95	22	9.025	484	2.090
100	25	10.000	625	2.500
85	18	7.225	324	1.530
112	30	12.544	900	3.360
90	21	8.100	441	1.890
130	41	16.900	1.681	5.330
115	32	13.225	1.024	3.680
100	31	10.000	961	3.100
92	22	8.464	484	2.024
1.039	277	109.883	8.149	29.704

30.1. Para obtener el coeficiente de correlación lineal calculamos previamente la covarianza y las desviaciones típicas de las variables:

$$s_{xy} = \frac{\sum_{t=1}^{N} x_t y_t}{N} - \overline{xy} = \frac{29.704}{10} - 103,9 \times 27,7 = 92,37$$

$$s_x^2 = \frac{\sum_{t=1}^{N} x_t^2}{N} - \overline{x}^2 = \frac{109.883}{10} - 103,9^2 = 193,09 \Rightarrow s_x = 13,90$$

$$s_y^2 = \frac{\sum_{t=1}^{N} y_t^2}{N} - \overline{y}^2 = \frac{8.149}{10} - 27,7^2 = 47,61 \Rightarrow s_y = 6,90$$

Entonces:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{92,37}{13,90 \times 6,90} = 0,96$$

El coeficiente de correlación lineal es positivo y está próximo a la unidad, por tanto, **entre los ingresos por servicio de restaurante y por servicio de bar existe una relación lineal de tipo directo muy intensa**.

30.2. La existencia de relación lineal o correlación implica la dependencia estadística, luego **las variables no son estadísticamente independientes**.

30.3. Estimación de la ecuación $y_t = \alpha + \beta x_t$:

$$b = \frac{s_{xy}}{s_x^2} = \frac{92,37}{193,09} = 0,4784$$

$$a = \bar{y} - b\bar{x} = 27,7 - 0,4784 \times 103,9 = -22,005$$

$$\hat{y}_t = -22,005 + 0,4784 x_t$$

La pendiente de la recta de regresión estimada indica la variación que se estima que se produce en el regresando, y, si el regresor, x, varía una unidad. En este caso, como el coeficiente es positivo, las variaciones de ambas variables tienen el mismo sentido, y las unidades de medida de las variables son cientos de euros.

Entonces, **b = 0,4784 significa que por cada aumento (disminución) de cien euros en los ingresos diarios por servicio de restaurante se estima un aumento (disminución) de los ingresos diarios por servicio de bar de 0,4784 cientos de euros**. O lo que es lo mismo, por cada euro adicional ingresado por servicio de restaurante se estima un ingreso adicional por servicio de bar de, aproximadamente, 0,5 euros.

30.3. El coeficiente de determinación es:

$R^2 = r^2 = 0{,}96^2 = 0{,}9216$

Dado que está próximo a la unidad, **con esta ecuación se ha obtenido un buen ajuste**. Su valor indica, además, que **el 92 por ciento de las variaciones de los ingresos diarios por servicio de bar en la muestra disponible son explicadas por las variaciones de los ingresos diarios por servicio de restaurante**.

EJERCICIO 31

Se dispone de los datos contenidos en la tabla siguiente respecto a las variables y = beneficios y x = gastos de las empresas turísticas de un municipio gallego, expresadas en miles de euros.

Gastos	Beneficios	
	250 – 350	350 – 450
10 – 15	4	5
15 – 30	6	10

31.1. Calcule las medias y las desviaciones típicas marginales de los beneficios y los gastos.
31.2. Indique si, dada esta información, puede afirmarse que los beneficios y los gastos mantienen una relación lineal intensa.

Distribución marginal de y = Beneficios

$L_{i-1} – L_i$	y_j	$n_{.j}$
250 – 350	300	10
350 – 450	400	15
		25

Tabla de cálculos intermedios

$L_{i-1} - L_i$	y_j	$n_{.j}$	$y_j n_{.j}$	$y_j^2 n_{.j}$
250 – 350	300	10	3.000	900.000
350 – 450	400	15	6.000	2.400.000
		25	9.000	3.300.000

Distribución marginal de x = Gastos

$L_{i-1} - L_i$	x_i	$n_{i.}$
10 – 15	12,5	9
15 – 30	22,5	16
		25

Tabla de cálculos intermedios

$L_{i-1} - L_i$	x_i	$n_{i.}$	$x_j n_{i.}$	$x_i^2 n_{i.}$
10 – 15	12,5	9	112,5	1.406,25
15 – 30	22,5	16	360	8.100
		25	472,5	9.506,25

31.1. Medias y desviaciones típicas marginales de los beneficios y los gastos:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_j n_{.j}}{\sum_{i=1}^{n} n_{.j}} = \frac{9.000}{25} = 360$$

$$s_y^2 = \frac{\sum_{i=1}^{n} y_j^2 n_{.j}}{\sum_{i=1}^{n} n_{.j}} - \bar{y}^2 = \frac{3.300.000}{25} - 360^2 = 2.400 \;\Rightarrow\; s_y = 48,99$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_{i.}}{\sum_{i=1}^{n} n_{i.}} = \frac{472,5}{25} = 18,9$$

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2 n_{i.}}{\sum_{i=1}^{n} n_{i.}} - \overline{x}^2 = \frac{9.506,25}{25} - 18,9^2 = 23,04 \Rightarrow s_x = 4,8$$

La media de los beneficios es 360 miles de euros o **360.000 euros y su desviación típica** es 48,99 miles de euros o **48.990 euros**. **La media de los gastos es** 18,9 miles de euros o **18.900 euros y su desviación típica** es 4,8 miles de euros o **4.800 euros**.

31.2. Para analizar si los beneficios están relacionados con los gastos se calcula la covarianza de las variables.

Gastos	Beneficios	
	300	400
12,5	4	5
22,5	6	10

$$s_{xy} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k} x_i y_j n_{ij}}{N} - \overline{xy} =$$

$$= \frac{12,5 \times 300 \times 4 + 12,5 \times 400 \times 5 + 22,5 \times 300 \times 6 + 22,5 \times 400 \times 10}{4 + 5 + 6 + 10} - 18,9 \times 360 =$$

$$= \frac{170.500}{25} - 18,9 \times 360 = 16$$

Por tanto, **entre los beneficios y los gastos existe una relación lineal de tipo directo**, ya que la covarianza es positiva. Sin embargo, dicha relación es **muy débil**, puesto que el coeficiente de correlación lineal está muy próximo a cero:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{16}{4,8 \times 48,99} = 0,068$$

EJERCICIO 32

Con datos de las variables y = gasto en hoteles de cinco estrellas y x = ingresos familiares, correspondientes a diez familias, se han obtenido las medias y las varianzas, $\bar{x} = 492$, $\bar{y} = 18{,}8$, $s_x^2 = 45.536$ y $s_y^2 = 144{,}76$ y el coeficiente de determinación de la regresión de y sobre x, $R^2 = 0{,}8942$.

32.1. Obtenga la covarianza de x e y e interprete el resultado obtenido.
32.2. Efectúe la regresión de y sobre x y explique el significado de los coeficientes estimados.
32.3. Calcule la varianza residual.
32.4. ¿Cuál es el campo de variación del coeficiente de determinación? Interprete su valor.

32.1. La covarianza puede obtenerse despejando en el coeficiente de correlación:

$r^2 = R^2$

$$r = \sqrt{R^2} = \sqrt{0{,}8942} = 0{,}9456$$

$$r = \frac{s_{xy}}{s_x s_y}$$

$$0{,}9456 = \frac{s_{xy}}{213{,}39 \times 12{,}03}$$

$$s_{xy} = 0{,}9456 \times 213{,}39 \times 12{,}03 = 2.427{,}43$$

La covarianza es mayor que cero, lo que **indica que entre el gasto en hoteles de cinco estrellas y los ingresos familiares existe una relación lineal de tipo directo**.

32.2. Estimación de la ecuación $y_t = \alpha + \beta x_t$:

$$b = \frac{s_{xy}}{s_x^2} = \frac{2.427,43}{45.536} = 0,053$$

$$a = 18,8 - 0,053 \times 492 = -7,276$$

$$\hat{y}_t = -7,276 + 0,053x_t$$

a = –7,276 es la ordenada en el origen de la recta de regresión estimada; por tanto, es el valor estimado del regresando, y, cuando el regresor, x, se anula.

Entonces, este coeficiente estimado indica que si los ingresos familiares fuesen iguales a cero, el gasto estimado en hoteles de cinco estrellas sería –7.276 unidades. Obviamente, esta interpretación carece de sentido, puesto que el mínimo valor que el gasto puede tomar es cero.

Si se iguala a cero el gasto estimado:

$$0 = -7,276 + 0,053x_t$$

$$x_t = \frac{7,276}{0,053} = 137,28$$

Luego se estima que el gasto es nulo para ingresos familiares de 137,28 unidades y sólo es positivo para ingresos superiores a éstos.

Es decir, que la ordenada en el origen de la recta de regresión estimada **indica que se estima que las familias con ingresos inferiores o iguales a 137,28 unidades no realizan ningún gasto en hoteles de cinco estrellas**.

b = 0,053 es la pendiente de la recta de regresión estimada; por tanto, es la variación que se estima que se produce en el regresando, y, cuando el regresor, x, experimenta una variación unitaria. En este caso, su valor es positivo, lo que indica que las variaciones de las dos variables tienen el mismo sentido.

Entonces, este coeficiente **indica que** para las familias con ingresos mayores que 137,28 unidades **si se produce un aumento (disminución) de una unidad en los ingresos familiares, se estima que el gasto en hoteles de cinco estrellas aumenta (disminuye) 0,053 unidades**.

32.3. Para obtener la varianza residual:

$$R^2 = 1 - \frac{s_e^2}{s_y^2} \Rightarrow s_e^2 = (1 - R^2) \times s_y^2 = (1 - 0{,}8942) \times 144{,}76 = 15{,}315$$

Luego la **varianza muestral del error es de 15,315 unidades del gasto en hoteles al cuadrado**.

32.4. En este modelo, estimado por mínimos cuadrados ordinarios y que tiene ordenada en el origen, **el coeficiente de determinación**, R^2, **toma valores comprendidos entre cero y la unidad**.

Para justificarlo, basta tener en cuenta que:

$$R^2 = 1 - \frac{s_e^2}{s_y^2}$$

donde s_e^2 y s_y^2 son valores positivos y la descomposición de la varianza garantiza que s_e^2 es menor o igual que s_y^2.

El coeficiente de determinación se obtiene, entonces, restando de la unidad un cociente mayor que cero y menor que uno, luego no puede ser superior a uno ni menor que cero.

Su valor está, en este caso, próximo a 0,90, de manera que **con esta ecuación se ha obtenido un ajuste aceptable, aunque el modelo deja sin explicar algo más del 10 por ciento de las variaciones del gasto en hoteles de cinco estrellas en esta muestra**.

EJERCICIO 33

Con datos de las variables x = renta familiar disponible (miles de euros) e y = miles de pernoctaciones registradas en establecimientos hoteleros en Galicia correspondientes al período comprendido entre 1992 y 1995, se han obtenido las varianzas, $s_x^2 = 31.918,75$ y $s_y^2 = 36,69$ y el coeficiente de correlación lineal de las variables, $r = 0,9871$.

33.1. Calcule e interprete la covarianza de las pernoctaciones y la renta.
33.2. ¿Qué información proporciona el coeficiente de correlación lineal? ¿Cuál es su campo de variación?
33.3. Obtenga el valor estimado de β en la ecuación $y_t = \alpha + \beta x_t$ e interprete el resultado obtenido.
33.4. Sabiendo que la suma de los cuadrados de los errores es 3,762 obtenga el valor de R^2, compruebe que es igual a r^2, y explique su significado.

33.1. El valor de la covarianza puede obtenerse despejando en r:

$$r = \frac{s_{xy}}{s_x s_y}$$

$$0,9871 = \frac{s_{xy}}{6,057 \times 178,66}$$

$$s_{xy} = 0,9871 \times 6,057 \times 178,66 = 1.068,18$$

La covarianza de las pernoctaciones y la renta es positiva. Por tanto, **indica la existencia de una relación lineal de tipo directo entre las variables**.

33.2. **El coeficiente de correlación lineal** toma valores comprendidos entre –1 y 1. Valores próximos a menos uno indican una fuerte asociación lineal inversa o negativa entre las variables, valores próximos a cero indican que la correlación es débil y valores próximos a uno indican una fuerte asociación lineal directa o positiva entre las variables.

En este caso, su valor está próximo a la unidad. Por tanto, **indica que la relación** lineal positiva entre las variables, **es muy intensa**.

33.3. $b = \frac{s_{xy}}{s_x^2} = \frac{1.068,18}{31.918,75} = 0,033$

La estimación de β, b, es la pendiente de la recta de regresión estimada. Indica la variación que se estima que se produce en el regresando, y, si el regresor, x, varía una unidad. En este caso, como el coeficiente es positivo, las variaciones de ambas variables tienen el mismo sentido, y las pernoctaciones están expresadas en miles y la renta en miles de euros.

Entonces, **b = 0,033 significa que** por cada aumento (disminución) de mil euros en la renta se estima que las pernoctaciones registradas aumentan (disminuyen) 0,03 miles, o lo que es lo mismo, **por cada euro que la renta familiar disponible aumenta (disminuye) se estima un aumento (disminución) de 0,03 pernoctaciones**.

33.4. La varianza residual es:

$$s_e^2 = \frac{SCE}{N} = \frac{3,762}{4} = 0,9405$$

Entonces:

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{0,9405}{36,69} = 0,974$$

El cuadrado del coeficiente de correlación es:

$r^2 = 0,9871^2 = 0,974$

Por tanto, efectivamente, **R^2 es igual a r^2**.

El coeficiente de determinación está, en este caso, muy próximo a la unidad, indicando que **el ajuste es satisfactorio. La ecuación explica algo más del 97 por ciento de las variaciones del número de pernoctaciones en esta muestra**.

EJERCICIO 34

Si la covarianza entre dos variables es nula, las variables son independientes. ¿Es esta afirmación correcta? Justifique su respuesta.

La afirmación es falsa. Si la covarianza entre dos variables es nula, entre ellas no existe relación lineal, pero pueden mantener una relación de dependencia de algún otro tipo.

Por ejemplo, si $y_t = x_t^2$, no hay relación lineal entre las variables x e y, puesto que la función que las relaciona no es lineal, sino que es un polinomio de grado 2. Sin embargo, y mantiene con x una relación de dependencia exacta; es decir, las variables no están linealmente relacionadas, pero no son independientes.

EJERCICIO 35

Para analizar la relación existente entre las variables y = duración de la estancia (días) en el viaje principal de vacaciones y x = ingresos familiares anuales (cientos de euros) se dispone de la siguiente información:

Familia	y_t	x_t
1	2	80
2	4	90
3	4	100
4	8	120
5	10	130
6	12	140
7	20	170

35.1. Efectúe la estimación mínimo–cuadrático ordinaria de la ecuación $y_t = \beta x_t$ y explique el significado del valor estimado de β.
35.2. Sabiendo que SCE es 6,8462 ¿podría indicar qué porcentaje de las variaciones de la duración de la estancia no son explicadas por los ingresos?

Tabla de cálculos intermedios

Familia	y_t	x_t	$x_t\, y_t$	x^2_t	y^2_t
1	2	80	160	6.400	4
2	4	90	360	8.100	16
3	4	100	400	10.000	16
4	8	120	960	14.400	64
5	10	130	1.300	16.900	100
6	12	140	1.680	19.600	144
7	20	170	3.400	28.900	400
	60		8.260	104.300	744

35.1. Estimación de la ecuación $y_t = \beta x_t$:

$$b = \frac{\sum_{t=1}^{N} y_t x_t}{\sum_{t=1}^{N} x_t^2} = \frac{8.260}{104.300} = 0{,}079$$

$$\hat{y}_t = 0{,}079 x_t$$

La estimación de β, b, es la pendiente de la recta de regresión estimada. Indica la variación que se estima que se produce en el regresando, y, si el regresor, x, varía una unidad. En este caso, como el coeficiente es positivo, las variaciones de ambas variables tienen el mismo sentido, y la duración de la estancia está expresada en días y los ingresos familiares en miles de pesetas.

Entonces, **b = 0,079 significa que por cada aumento (disminución) de cien euros en los ingresos familiares anuales se estima un aumento (disminución) de la duración de la estancia de 0,079 días**.

Se estima, por tanto, que para que la duración de la estancia se alargue un día:

$$1 \text{ día} = 0{,}079 \times \frac{1000}{79} \text{ días}$$

los ingresos familiares deberían aumentar:

$$100 \text{ euros} \times \frac{1000}{79} = 1.256{,}823 \text{ euros}$$

35.2. Las varianzas residual y del regresando son:

$$s_e^2 = \frac{SCE}{N} = \frac{6{,}8462}{7} = 0{,}978$$

$$s_y^2 = \frac{\sum_{t=1}^{N} y_t^2}{N} - \bar{y}^2 = \frac{744}{7} - 8{,}571^2 = 32{,}824$$

Entonces:

$$R^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{0{,}978}{32{,}824} = 0{,}970$$

Pero **este modelo no tiene ordenada en el origen** y, por tanto, no se puede descomponer la varianza. El coeficiente de determinación no es igual al cociente de varianzas del regresando estimado y el regresando, luego su valor no es la proporción de la varianza de y que representa la varianza de $\hat{y}$, que es el cociente que puede interpretarse (multiplicado por cien) como el porcentaje de variaciones del regresando que quedan explicadas por el modelo estimado.

Entonces, **el coeficiente de determinación no permite**, en este caso, **deducir cuál es el porcentaje de las variaciones de la duración de la estancia en esta muestra que no son explicadas por los ingresos**.

EJERCICIO 36

El coeficiente de determinación, ¿puede ser, en algún caso, mayor que uno? ¿y menor que cero? Justifique su respuesta.

El coeficiente de determinación en ningún caso es mayor que uno y sólo puede ser menor que cero si el modelo no tiene ordenada en el origen.

Se define como:

$$R^2 = 1 - \frac{s_e^2}{s_y^2}$$

En ningún caso es mayor que uno, porque el cociente s_e^2 / s_y^2 es positivo, de manera que R^2 es igual a uno menos una cantidad positiva y, por tanto, es menor que la unidad.

Si el modelo tiene ordenada en el origen no es negativo, porque la descomposición de la varianza garantiza que la varianza residual es como máximo igual a la varianza del regresando, de manera que el cociente entre las varianzas residual y del regresando es menor o igual que la unidad y R^2 es igual a uno menos una cantidad menor o igual que uno, por tanto, es mayor o igual que cero.

Si el modelo no tiene término independiente puede ser negativo, porque no se puede descomponer la varianza y s_e^2 puede ser menor o igual que s_y^2, de manera que el cociente s_e^2 / s_y^2 puede ser mayor que uno. En tal caso caso, R^2 sería igual a uno menos una cantidad mayor que uno; es decir, menor que cero.

EJERCICIO 37

Indique si los siguientes resultados son compatibles:

37.1.	$r > 0$	$b > 0$		
37.2.	$s_y^2 = 36{,}69$	$s_x = 178{,}66$	$s_{xy} = 1.068{,}21$	$s_e^2 = 0{,}9404$
37.3.	$\sum_{t=1}^{N} x_t y_t = 92{,}37$	$\sum_{t=1}^{N} x_t^2 = 13{,}896$	$\hat{y}_t = 6{,}647 x_t$	
37.4.	$s_{\hat{y}}^2 = 44{,}188$	$s_y^2 = 47{,}61$	$s_e^2 = 3{,}422$	$y_t = \alpha + \beta x_t$
37.5.	$\bar{x} = 360$	$\bar{y} = 18{,}8$	$a = -7{,}427$	$b = 0{,}053$

37.1. Los resultados son **compatibles**.

$$r = \frac{s_{xy}}{s_x s_y} > 0 \Rightarrow s_{xy} > 0$$

$$b = \frac{s_{xy}}{s_x^2} > 0 \Rightarrow s_{xy} > 0$$

37.2. Los resultados son **compatibles**.

$$s_{xy} = 1.068,21;\ s_x = 178,66 \text{ y } s_y^2 = 36,69 \Rightarrow r = \frac{1.068,21}{178,66 \times 6,057} = 0,987$$

$$s_e^2 = 0,9404 \text{ y } s_y^2 = 36,69 \Rightarrow R^2 = 1 - \frac{0,9404}{36,69} = 0,974 \Rightarrow r = \sqrt{0,974} = 0,987$$

37.3. Los resultados son **compatibles**.

$$\hat{y}_t = 6,647x_t \Rightarrow b = 6,647$$

$$\sum_{i=1}^{N} x_t y_t = 92,37 \text{ y } \sum_{i=1}^{N} x_t^2 = 13,896 \Rightarrow b = \frac{\sum_{i=1}^{N} x_t y_t}{\sum_{i=1}^{N} x_t^2} = \frac{92,37}{13,8956} = 6,647$$

37.4. Los resultados son **compatibles**.

$y_t = \alpha + \beta x_t \Rightarrow$ el modelo tiene ordenada en el origen; entonces:

$$s_{\hat{y}}^2 = 44,188 \text{ y } s_y^2 = 47,61 \Rightarrow R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{44,188}{47,61} = 0,928$$

$$s_e^2 = 3,422 \text{ y } s_y^2 = 47,61 \Rightarrow R^2 = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{3,422}{47,61} = 0,928$$

37.5. Los resultados son **incompatibles**.

a = –7,427 $\Rightarrow$ el modelo tiene ordenada en el origen.

$\bar{y}$ = 18,8

$\bar{x} = 360$, a = –7,427 y b = 0,053 $\Rightarrow$ $\bar{y} = -7{,}427 + 0{,}053 \times 360 = 11{,}65$

EJERCICIO 38

Se dispone de la siguiente información respecto a las variables y = millones de personas que llegan a España procedentes de los cinco países con mayor peso en la composición del turismo español y x = media ponderada de los PIB per cápita (miles de euros) de dichos países.

Países	y_t	x_t	$x_t y_t$	x_t^2	y_t^2
A	1,8	3,1	5,58	9,61	3,24
B	2,7	4,2	11,34	17,64	7,29
C	0,7	2,7	1,89	7,29	0,49
D	1,2	3,0	3,60	9,00	1,44
E	1,9	4,0	7,60	16,00	3,61
	8,3	17,0	30,01	59,54	16,07

38.1. Cabe esperar que entre las variables x e y exista una relación lineal intensa de tipo directo. Explique el significado de esta expresión y compruebe que se cumple utilizando la(s) medida(s) que considere adecuada(s).
38.2. Al efectuar la estimación del modelo sin ordenada en el origen que relaciona a y con x se ha obtenido un valor de la pendiente de la recta de regresión estimada igual a 0,5. ¿Es correcto este resultado? ¿Qué significa?
38.3. El coeficiente de determinación del modelo anterior resulta igual a 0,58. ¿Puede afirmarse, entonces, que las variaciones del PIB explican el 58 por ciento de las variaciones del número de turistas en esta muestra? Justifique su respuesta.

38.4. Suponiendo, ahora, que el modelo tiene término independiente, obtenga la ordenada en el origen de la recta de regresión estimada y explique su significado.
38.5. Calcule el coeficiente de determinación para el modelo con ordenada en el origen e interprete el resultado obtenido.

38.1. Significa que se supone que la función que relaciona al número de turistas con el PIB per cápita es lineal.

Además, si la relación que existe entre estas variables es de tipo directo, se espera que el número de turistas responda a las variaciones en el PIB per cápita con variaciones del mismo sentido.

Finalmente, si la relación es muy intensa, a valores pequeños (elevados) del PIB per cápita les corresponden valores pequeños (elevados) del número de turistas, de manera que la mayoría de los productos $(x_t - \bar{x})(y_t - \bar{y})$ son positivos y su suma es grande, de manera que es elevado el valor de la covarianza de las variables.

La covarianza está expresada en el producto de las unidades de medida de x e y, y no tiene límites inferior o superior para valorar su tamaño. Por eso, para analizar si es grande o pequeña se utiliza el coeficiente de correlación, que se define por cociente entre la covarianza de las variables y el producto de sus desviaciones típicas y es, por tanto, adimensional. En el caso de una relación lineal de tipo directo muy intensa el coeficiente de correlación debe resultar próximo a la unidad.

Calculamos, entonces, el coeficiente de correlación lineal, que debe estar próximo a la unidad. La covarianza y las desviaciones típicas de las variables son:

$$s_{xy} = \frac{\sum_{t=1}^{N} x_t y_t}{N} - \overline{xy} = \frac{30,01}{5} - 3,4 \times 1,66 = 0,358$$

$$s_x^2 = \frac{\sum_{t=1}^{N} x_t^2}{N} - \bar{x}^2 = \frac{59,54}{5} - 3,40^2 = 0,348 \quad \Rightarrow \quad s_x = 0,59$$

$$s_y^2 = \frac{\sum_{t=1}^{N} y_t^2}{N} - \bar{y}^2 = \frac{16,07}{5} - 1,66^2 = 0,458 \quad \Rightarrow \quad s_y = 0,68$$

Entonces:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{0,358}{0,59 \times 0,68} = 0,89$$

Efectivamente, **el coeficiente de correlación lineal está próximo a uno**, luego entre las variables x e y existe una relación lineal de tipo directo bastante intensa.

38.2. En el modelo sin ordenada en el origen, la pendiente de la recta de regresión estimada es:

$$b = \frac{\sum_{t=1}^{N} x_t y_t}{\sum_{t=1}^{N} x_t^2} = \frac{30,01}{59,54} = 0,5$$

Por tanto, **el resultado es correcto**. Su valor indica la variación estimada del regresando, y, si se produce en el regresor, x, una variación de una unidad. En este caso, como es positivo, las variaciones de las dos variables tienen el mismo sentido, lo cual es coherente con la existencia de una relación de tipo directo entre las variables, como indican la covarianza y el coeficiente de correlación.

Luego b = 0,5 es la variación estimada en los millones de personas que llegan a España procedentes de los cinco países a los que se refiere el análisis si la media ponderada del PIB per capita varía una unidad. Como el PIB está expresado en miles de euros, **por cada 1.000 euros que la media ponderada del PIB aumenta (disminuye) se estima que llegan a España** 0,5 millones de personas más (menos); es decir, **medio millón de personas más (menos)**.

38.3. **No es correcto**. En un modelo sin ordenada en el origen el coeficiente de determinación no es el cociente entre las varianzas del regresando estimado y el regresando, ya que no puede efectuarse la descomposición de la varianza.

Entonces, su valor (multiplicado por cien) no puede interpretarse como el porcentaje de variaciones del regresando en la muestra que son explicadas por el modelo estimado. En este caso, **R^2 sólo indica que no se ha obtenido un buen ajuste**, porque su valor no está próximo la unidad.

38.4. Para obtener la ordenada en el origen de la recta de regresión estimada debe efectuarse previamente el cálculo de la pendiente:

$$b = \frac{s_{xy}}{s_x^2} = \frac{0,358}{0,348} = 1,029$$

Entonces:

$a = \overline{y} - b\overline{x}$ = 1,66 – 1,029 x 3,4 = – 1,839

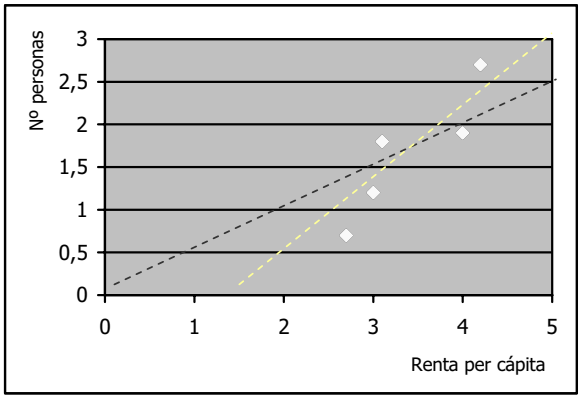
La ordenada en el origen de la recta de regresión estimada es el valor estimado del regresando, y, si se anula el regresor, x.

En este caso, como la media ponderada del PIB de los cinco países a los que se refiere el estudio es una variable no nula, **su valor carece de interés** desde el punto de vista económico.

38.5. $R^2 = r^2 = 0,89^2 = 0,79$

Su valor indica que, aunque **el ajuste** es algo mejor cuando el modelo incluye término independiente que cuando no lo incluye, **sigue sin ser satisfactorio**. **Las variaciones de la renta explican sólo el 79 por ciento de las variaciones del número de turistas en esta muestra**.

Algunas de las conclusiones a las que hemos llegado respecto a la relación entre estas variables pueden visualizarse en el diagrama de dispersión.



Efectivamente, la nube de puntos del diagrama indica la existencia de una relación aproximadamente lineal y directa entre las variables.

Las líneas de puntos añadidas en el gráfico son las rectas de regresión estimadas, con (blanca) y sin (negra) término independiente. Observando cómo se ajustan a los puntos del diagrama, puede verse que la que tiene término independiente representa mejor la nube de puntos que la que no lo tiene, aunque en ninguno de los dos casos la recta representa la información muestral correctamente.

EJERCICIO 39

Con datos correspondientes al período comprendido entre los años 1979 y 1985, la estimación mínimo–cuadrático ordinaria del modelo que relaciona las variables y = número de plazas registradas en establecimientos hoteleros españoles y x = número de viajeros que entran en España procedentes del extranjero, ha proporcionado el siguiente resultado:

$$\hat{y}_t = 344{,}67 + 11{,}50x_t$$
$$R^2 = 0{,}92$$

39.1. Coméntelo, explicando su significado.
39.2. Calcule e interprete el coeficiente de correlación lineal simple.

39.3. Demuestre que tanto si el modelo tiene ordenada en el origen, como si no la tiene se cumple la siguiente igualdad:

$$\sum_{t=1}^{N} e_t \hat{y}_t = 0$$

39.1. **a = 344,67** es la ordenada en el origen de la recta de regresión estimada. Indica el valor estimado del regresando, y, cuando el regresor, x, se anula.

En este caso, por tanto, si el número de viajeros es igual a cero, se estima que se registran 344,67 plazas. Pero el número de viajeros no es nulo; por tanto, **desde el punto de vista económico, este coeficiente no tiene ningún interés**.

b = 11.50 es la pendiente de la recta de regresión estimada. Indica la variación estimada del regresando, y, cuando en el regresor, x, se produce una variación unitaria. Como su valor es positivo, en este caso, las variaciones de las dos variables tienen el mismo sentido.

Entonces, si el número de viajeros varía una unidad, se estima que el número de plazas experimenta una variación del mismo sentido de 11,50 unidades; es decir, **por cada viajero adicional, se estima que se registran 11,50 plazas más**.

R^2 = 0,92 es el valor del coeficiente de determinación y, dado que está próximo a uno, **indica que con esta ecuación se ha obtenido un buen ajuste**. **Las variaciones en el número de visitantes explican el 92 por ciento de las variaciones del número de plazas registradas en esta muestra**.

39.2. Como $r^2 = R^2$:

$r = \sqrt{R^2} = \sqrt{0,92} = 0,96$

El coeficiente de correlación lineal simple toma, por tanto, un valor muy próximo a la unidad, que indica que **entre el número de plazas registradas y el número de viajeros extranjeros existe una relación lineal de tipo directo muy intensa**. Estas variables están fuertemente correlacionadas.

39.3. Para hacer esta demostración debe tenerse en cuenta que el proceso de estimación mínimo cuadrático ordinaria minimiza la suma de los cuadrados de los errores; es decir, iguala a cero las derivadas parciales de SCE respecto a los coeficientes estimados.

1. En el modelo con ordenada en el origen:

$$SCE = \sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{N} (y_t - (a + bx_t))^2$$

Entonces:

$$\frac{\delta SCE}{\delta a} = -2\sum_{t=1}^{N} (y_t - (a + bx_t)) = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} (y_t - \hat{y}_t) = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} e_t = 0$$

$$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N} (y_t - (a + bx_t))x_t = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} (y_t - \hat{y}_t)x_t = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} e_t x_t = 0$$

Y se deduce que:

$$\sum_{t=1}^{N} e_t \hat{y}_t = \sum_{t=1}^{N} e_t (a + bx_t) = \sum_{t=1}^{N} (ae_t + be_t x_t) = a\sum_{t=1}^{N} e_t + b\sum_{t=1}^{N} e_t x_t = 0$$

2. En el modelo sin ordenada en el origen:

$$SCE = \sum_{t=1}^{N} e_t^2 = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{N} (y_t - bx_t)^2$$

Entonces:

$$\frac{\delta SCE}{\delta b} = -2\sum_{t=1}^{N} (y_t - (bx_t))x_t = 0\_ \Rightarrow \quad \sum_{t=1}^{N} (y_t - \hat{y}_t)x_t = 0 \quad \Rightarrow \quad \sum_{t=1}^{N} e_t x_t = 0$$

Y se deduce que:

$$\sum_{t=1}^{N} e_t \hat{y}_t = \sum_{t=1}^{N} e_t (bx_t) = b\sum_{t=1}^{N} e_t x_t = 0$$

EJERCICIO 40

Con datos anuales de las variables y = número de viajeros (miles de personas) residentes en España, y x = PIB (millones de euros) español, correspondientes al período comprendido entre 1992 y 1995, se han calculado las varianzas, s_x^2 = 1.050.223,9 y s_y^2 = 29.441.161, y el coeficiente de correlación lineal, r = 0,9846.

40.1. Calcule la covarianza entre el número de viajeros y el PIB y explique su significado y el del coeficiente de correlación lineal.
40.2. Suponiendo que el modelo tiene término independiente, obtenga la pendiente de la recta de regresión estimada e interprete el resultado.
40.3. Calcule el coeficiente de determinación e indique qué información proporciona.

40.1. Dada la información disponible:

$$s_x = \sqrt{s_x^2} = \sqrt{1.050.223,9} = 1.024,8$$

$$s_y = \sqrt{s_y^2} = \sqrt{29.441.161} = 5.425,97$$

$$r = \frac{s_{xy}}{s_x \times s_y}$$

$$0,9846 = \frac{s_{xy}}{1.024,8 \times 5.425,97}$$

$$s_{xy} = 5.474.901,832$$

La covarianza es positiva, luego **indica la existencia de una relación lineal de tipo directo entre el número de viajeros y el PIB**.

El coeficiente de correlación lineal está próximo a uno, por tanto, **indica que dicha relación es, además, muy intensa**.

40.2. $$b = \frac{s_{xy}}{s_x^2} = \frac{5.474.901,832}{1.050.223,9} = 5,213$$

Su valor indica la variación estimada del regresando, y, si se produce una variación de una unidad en el regresor, x. Como es mayor que cero, las variaciones de las dos variables son del mismo sentido.

En este caso, **significa que** si el PIB español experimenta una variación de una unidad, se estima que el número de viajeros residentes en España varía 5,213 unidades. Como el PIB está expresado en millones de euros y los viajeros son miles de personas, **si aumenta (disminuye) el PIB un millón de euros, se estima que el número de viajeros aumenta (disminuye)** 5,213 miles de personas o **5.213 personas**.

40.3. Como $r^2 = R^2$:

$$R^2 = (0{,}9846)^2 = 0{,}969$$

Por tanto, **se ha obtenido un buen ajuste**. **Las variaciones del PIB explican casi el 97 por ciento de las variaciones del número de viajeros en esta muestra**.