Feliciano F. Ordóñez Fernández
Juan González Fernández

# ESTADÍSTICA DESCRIPTIVA *paso a paso*

EDICIONES PIRÁMIDE

COLECCIÓN «ECONOMÍA Y EMPRESA»

Director:
Miguel Santesmases Mestre
Catedrático de la Universidad de Alcalá

Diseño de cubierta: Anaí Miguel



Ediciones Pirámide se compromete con el medio ambiente reduciendo la huella de carbono de sus libros.

PAPEL DE FIBRA CERTIFICADO


Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
Depósito legal: M. 29.799-2020
ISBN: 978-84-368-4378-1
Printed in Spain

# Índice

# Introducción

El objeto de este libro es acercar a los alumnos una disciplina que por su temática es considerada en ocasiones como dura y difícil de comprender.

Es cierto que, como toda materia relacionada con los números, es considerada normalmente en su forma abstracta y, por tanto, difícil de trasmitir y comprender; sin embargo, queremos que el lector la vea como una disciplina práctica y sencilla de entender.

En estos años como profesores de esta materia en Ciencias Sociales nos hemos encontrado con un denominador común a la hora de explicar nuestra asignatura: su lenguaje específico. Es por ello que nuestro punto de partida y parte importante de este libro es desarrollar los conceptos y términos básicos relacionados con la estadística.

Nuestro primer debate ha sido cómo introducir estos conceptos de una forma lógica y continuada que permita la fácil adquisición del lenguaje estadístico. En un primer momento pensamos que sería fácil de entender el estilo que siguen los diccionarios, introduciendo de forma alfabética todos los términos y conceptos más importantes, pero ello seguramente conduciría irremediablemente a una mayor complejidad y pérdida de interés, ya que no tendrían un orden lógico a la hora de relacionarse con las diferentes partes de que consta esta disciplina. Finalmente hemos pensado introducir todos los conceptos a modo de glosario por bloques temáticos, siguiendo los contenidos que se desarrollan en casi todos los programas educativos. Así el lector podrá encontrar sin dificultad el concepto deseado, a la vez que irá adquiriendo

progresivamente un lenguaje que le permita comprender el discurso de la asignatura.

En este primer acercamiento a la estadística, nos centraremos en aquellos contenidos que tratan sobre su parte descriptiva, dejando para más adelante los contenidos probabilísticos e inferenciales.

Una vez descritos los conceptos y términos, cada tema irá acompañado de una serie de «ejercicios tipo» resueltos, que completarán un acercamiento a los temas propuestos desde su perspectiva conceptual al cálculo de los conceptos analizados.

## Objetivos

1. Establecer un glosario temático de los conceptos estadísticos para salvar el problema del lenguaje específico de esta disciplina.
2. Dotar al alumnado y profesorado de una herramienta para la adquisición de conceptos y ejercicios resueltos de estadística descriptiva.
3. Resaltar el sentido práctico de esta disciplina y la importancia que tiene ser rigurosos con las conclusiones en los resultados obtenidos.



# 1 La estadística como disciplina. Primeros conceptos necesarios para su comprensión

Es interesante plantearse dónde se ubica una disciplina como esta, ya que se tiende a incluirla en un gran número de ellas, baste con ver la multitud de estudios universitarios de diferentes materias en las que se imparte. Esto se ve agravado por la cantidad de definiciones en las que se ve reflejada, se puede ver como una ciencia en sí misma, ya que ha desarrollado sus propios modelos y métodos, o como Mood y Gybill (1978), para quien la estadística no es otra cosa que la tecnología del método científico[1].

Siguiendo este tipo de análisis, vemos que la Unesco las clasifica dentro de las Ciencias Sociales, mientras que FONDECYT la incluye en las matemáticas[2]. En nuestro día a día vemos que el ciudadano de a pie tiene una idea poco aproximada de qué es la estadística, ya que una gran parte la identifica con las matemáticas, mientras que otra parte se limita a no saber qué responder.

Nosotros hemos querido huir de estas controversias para darle una definición estructural que nos permita hacer una clasificación de la misma en tres partes: estadística descriptiva, probabilidad y estadística inferencial. No obstante, existen autores que refiriéndose de forma concreta a aplicaciones estadísticas en las Ciencias Sociales, dejan la probabilidad en un plano diferente, como de apoyo técnico e introducen como parte de la estadística el diseño experimental.

Esta perspectiva se enmarca en una definición de tipo práctica en la que **la estadística es la ciencia del análisis e interpretación de datos para la toma de decisiones:**

— Ciencia que se ocupa de la obtención de información y proporciona instrumentos para la toma de decisiones cuando prevalecen situaciones de incertidumbre.

— La rama del método científico que se ocupa de los datos obtenidos contando o midiendo las propiedades de determinados colectivos.

— Es la matriz de toda ciencia experimental y, por consiguiente, una rama del método científico, sino el método científico por excelencia —Kendall (1968)—.

[1] *Introducción a la Teoría Estadística*, 1978.

[2] Cit. en: Osvaldo Ferreiro P. y Pedro Fernández de la R. (1988). La estadística, una ciencia en la controversia. *Revista Universitaria*, 25.





Para llevar a cabo el proceso estadístico es necesario seguir varias fases:

1. **El diseño o planteamiento,** mediante el cual establecemos cuál es el objetivo a seguir. La población y muestra a estudiar, las características que nos interesa analizar y la representación de los datos obtenidos. Es importante saber el tiempo que tenemos y el coste de nuestro estudio.
2. **Recogida de los datos,** estos datos pueden ser **primarios** si es el propio equipo de trabajo (o individuo) quien facilita y recoge los datos, o **datos secundarios** o indirectos cuando se usan datos de otras investigaciones, ficheros, bancos de datos...
3. **Obtención de los resultados,** mediante el análisis de los datos recogidos, se trata de materializar las operaciones, cuadros, tablas o gráficos previstos en la fase de diseño para la obtención de resultados y facilitar su análisis.
4. **Interpretación de los resultados,** una vez que tenemos los resultados, se interpretan para facilitar a los destinatarios de los estudios tomar decisiones basándose en la información obtenida.

## 1.1. CONCEPTOS BÁSICOS

### 1.1.1. División de la estadística

La estadística como disciplina puede dividirse en tres grandes bloques: la estadística descriptiva (a la que se dedica este libro), la inferencial y la probabilidad.

**Estadística descriptiva.** Se refiere a la recolección, presentación, descripción, análisis e interpretación de una colección de datos y esencialmente consiste en resumir estos con uno o dos elementos de información (llamadas medidas descriptivas), que caracterizan a la totalidad de los mismos. En todo caso, la estadística descriptiva no pretende sacar conclusiones, sino poner de

manifiesto las principales características de una muestra de la población a estudio y, por tanto, es un proceso de abstracción, que busca organizar una colección de datos empleando métodos matemáticos y gráficos. También nos podemos referir a ella como estadística deductiva.

**Estadística inferencial.** En contraste con la estadística descriptiva, la estadística inferencial se refiere al proceso de obtener generalizaciones acerca de las propiedades del conjunto de la población partiendo de una parte representativa o muestra. Esto lleva implícitos una serie de riesgos: las muestras deben ser representativas de la población, la calidad de la información debe ser controlada y, por tanto, se tendrá que especificar el riesgo o probabilidad de los errores. La estadística inferencial busca obtener conclusiones que superen el análisis de los datos, buscando información de un colectivo mediante un metódico procedimiento del manejo de los datos de la muestra. Emplea modelos probabilísticos asociados a fenómenos aleatorios de las observaciones, obteniendo inferencias y predicciones sobre la población del conjunto de datos y su aplicación a una amplia gama de investigaciones de todas las ciencias. Tiene como limitación clara que las conclusiones obtenidas deben ser establecidas con una medida de su certidumbre, que a su vez viene dada por la probabilidad.

## 1.1.2. Población y muestra y muestreo

### Población

Siempre que nos enfrentemos a un proceso de investigación, un estudio o análisis, sea del tipo que sea, lo haremos basándonos en un fenómeno específico, el total de los casos que tiene ese fenómeno lo llamamos población. Denominamos población al conjunto de todos los elementos homogéneos que cumplen ciertas propiedades y en los cuales se desea estudiar un fenómeno determinado; se pueden denominar también colectivo o universo. De igual manera, se puede hablar de población finita cuando podemos conocer el número total de efectivos y de población infinita cuando no es posible conocer dicho dato. Pueden tener cualquier tipo de naturaleza, real o figurada: personas, edificios, hoteles, agencias, zonas... Cuando hablamos de población, estamos refiriéndonos a la totalidad de los individuos o datos a partir de los cuales se pretende realizar un estudio estadístico.





Este concepto se completa con lo que se denomina **tamaño** de una población, que es el número de elementos que configuran esa población. Este tamaño puede ser en ocasiones infinito, bien porque se trata de una operación que puede repetirse infinitas veces o bien porque se trata de un fenómeno tan grande que conviene tratarlo como infinito. No obstante, existen estudios que analizan uno o dos elementos o características con toda la población; a estos datos se las llama **censos**. La realización de estos censos conlleva mucho tiempo y trabajo, por lo que también se hacen estudios y análisis con una parte de la población, o muestra; cuando esto es así, hablamos de una **encuesta.**

El tamaño de la población viene representado por la letra $N$ (mayúscula). Las características poblacionales que deseamos investigar y que son desconocidas *a priori* se denominan **parámetros** y *suelen utilizar en su notación letras griegas*. Por ejemplo, la edad media, la proporción de personas que realizan una actividad, la proporción de jóvenes que eligen un tipo de bachillerato... No se debe equivocar con un **estadístico,** que es cualquier índice calculado para una muestra, *en este caso la notación utilizada es letras latinas*. Ejemplo: media poblacional «$\mu$», media de una muestra «$X$».

### Muestra

Como sabemos, casi nunca es posible obtener los datos sobre todos y cada uno de los elementos de una población, por ello cuando vamos a hacer un estudio lo realizaremos sobre un subconjunto de dicha población. A este subconjunto lo llamamos **muestra.** La muestra debe ser una representación de la población de estudio, este concepto de muestra lleva asociado el de **tamaño de muestra,** que es el número concreto de elementos que vamos a tomar del fenómeno estudiado, representado por la letra $N$ o $n$.

El interés de tener una buena muestra es que nos permite realizar inferencias (poder extender y generalizar nuestros resultados)

sobre el comportamiento de la población de estudio, con el ahorro de tiempo y dinero que supondría el estudio de poblaciones de grandes dimensiones. Por ello, el poder calcular el tamaño mínimo necesario para poder hacer este tipo de inferencias es una herramienta muy útil en cualquier estudio, ya que nos asegura la **validez** de la muestra y la **fiabilidad**[3] de los resultados, además de facilitarnos la tarea de la recogida de datos.

El problema se presenta a la hora de determinar si la muestra elegida tiene el tamaño adecuado y si es representativa de la población de estudio, es decir, cómo elegir la muestra significativa. Para ello existen diferentes técnicas **(muestreo),** y todas ellas tienen un común denominador: la selección debe ser al azar. Es decir, la selección de los sujetos de estudio es primordial si queremos asegurar que los resultados de nuestro estudio, análisis o investigación se puedan generalizar a la población objeto del estudio y no que sean válidos solamente para los elementos utilizados en el mismo **(muestra).** En cuanto al tamaño, existen métodos matemáticos para su medida[4].

Para ello la muestra debe responder a varias cuestiones:

1. Que la muestra sea representativa de la población que procede.
2. Que el tamaño sea el adecuado para poder generalizar los resultados.
3. Determinar claramente las características más representativas de la población objeto del estudio.
4. Que los elementos de la muestra sean elegidos al azar. Esto quiere decir que cualquier elemento y/o individuo de una población de estudio tiene la misma oportunidad de ser incluido en una muestra.
5. Todos los elementos de la muestra han de tener la misma probabilidad de pertenecer a la población.
6. La muestra así definida es una muestra no sesgada, ya que no hay ninguna restricción para que cualquier elemento

[3] Ambos conceptos se definen más adelante.
[4] Se analizará en el tema del cálculo de la muestra.





de la población pueda formar parte de la misma. No obstante, puede ocurrir que nuestra muestra sea sesgada; esto ocurre cuando ponemos una restricción sobre la inclusión de un determinado elemento en la muestra de estudio. Debido a esa restricción nuestra muestra puede que no sea representativa e introduce un error de muestreo sistemático. Debemos tener en cuenta que aun eligiendo nuestra muestra al azar, podemos incurrir en errores, pero la diferencia es que estos serán de tipo aleatorio, mientras que en la muestra sesgada se trata de errores sistemáticos. Para evitar este tipo de errores podemos asegurar la representatividad de nuestra muestra mediante alguna de las técnicas de muestreo existentes.

### Muestreo

Es la operación de seleccionar o elegir qué elementos o individuos de la población de estudio van a formar la muestra en la que se estudian uno o varios caracteres. El muestreo **debe ser** representativo, imagen lo más exacta posible de la población. Este muestreo puede ser aleatorio, mediante la selección al azar de todos sus elementos por medios matemáticos probabilísticos, o no aleatorio, dividiendo la población en función de una característica, y luego elegir el número de individuos que reproducirán la población general.

De entre todas las técnicas para la elección de una muestra significativa (representativa), destacan:

**Muestreo aleatorio simple (MAS).** Se trata de un procedimiento de muestreo (sin reemplazamiento) en el que se seleccionan $n$ *unidades* de las $N$ en la población, de forma que cualquier posible muestra del mismo tamaño tiene la misma probabilidad de ser elegida.

Se realizan $n$ selecciones independientes, de forma que en cada selección los individuos que no han sido elegidos tengan la misma probabilidad de serlo.

El procedimiento habitual consiste en numerar todos los elementos de la población y se seleccionan muestras del tamaño de-

seado utilizando una tabla de números aleatorios o un programa de ordenador que proporcione números aleatorios. Sería como usar una «mano inocente», que fuera sacando de una urna en la que están todos los elementos de la población los individuos de la muestra hasta configurarla por completo.

**Muestreo aleatorio sistemático.** Los elementos de la muestra se eligen sistemáticamente; en primer lugar, mediante algún criterio se ordena el colectivo y después se seleccionan elementos del mismo en función de su posición en la ordenación. Imaginemos que queremos analizar la población de bachillerato de Asturias, primero podemos utilizar como criterio el curso y el tipo de bachillerato elegido, y una vez establecido el número de unidades de bachillerato diferentes, y calculado el tamaño de la muestra, se decide que se analizarán aquellos que tengan los números en el aula 5 y 15.

**Muestreo aleatorio estratificado.** En primer lugar se busca una categoría homogénea dentro de la población a todos los elementos que la forman, y una vez establecida esa división se aplica el MAS en cada una de las categorías en las que hemos dividido nuestra población de estudio. Ejemplo: queremos estudiar una población en la que tengamos bien diferenciados unos intervalos de edad, dividimos nuestra población en tantos grupos como intervalos y luego dentro de cada grupo, utilizando el MAS, seleccionamos el número de elementos de nuestra muestra.

**Muestreo por conglomerados.** Se divide la población en grupos de acuerdo con una característica; por ejemplo, su proximidad geográfica o de otro tipo (conglomerados).

Cada grupo ha de ser heterogéneo y tener representadas todas las características de la población. Por ejemplo, los conglomerados en un estudio sobre la situación de las mujeres en una determinada zona rural pueden ser los municipios de la zona. Una vez dividida la población, se selecciona un grupo de los conglomerados al azar y se toma el conglomerado completo o una muestra del mismo. Se supone que cada conglomerado es representativo de la población que deseamos estudiar.





### 1.1.3. Individuo, elemento y dato

Individuo o unidad de investigación es cada uno de los elementos de una muestra o una población. Por ejemplo, una población sería el conjunto de visitantes del Museo del Jurásico de Asturias en el año 2014, cada uno de esos visitantes es un individuo o un elemento de la población. Una muestra, elegida aleatoriamente, podría ser la selección de los 5 visitantes que entraron en las instalaciones a partir de las 12:00 y a las 17:00, ambas horas elegidas al azar.

Llamamos **dato** a cada uno de los individuos, elemento, cosa o ente de cualquier naturaleza que forma una población o un universo de estudio determinado. Es decir, cada valor observado de una variable.

Los datos puede ser: de corte transversal, temporales o de panel.

- De corte transversal, en un momento de tiempo determinado y para distintos individuos medimos una o más variables.
- Temporal, cuando medimos una o más variables en intervalos de tiempo regular para un solo individuo. Se llaman también series temporales.
- De panel, una combinación de las dos anteriores: diferentes variables, en distintos individuos durante intervalos de tiempo regular.

### 1.1.4. Aproximación de cifras

Siguiendo con los posibles errores de medida que podemos cometer a la hora de analizar datos, debemos tener en cuenta el efecto del redondeo. Es cierto que actualmente trabajamos con herramientas estadísticas de gran precisión que pueden hacer cálculos con grandes números, pero en ocasiones para poder interpretar y transmitir la información y que pueda ser utilizada en la toma de decisiones es conveniente redondear los resultados obtenidos. En general, los resultados se muestran con 3, 4 o 5 cifras

importantes, por ello a la hora de redondear un valor es necesario eliminar alguna cifra; esto debe hacerse de manera que el error sea lo menor posible, aproximando por arriba o debajo, previamente debemos tener en cuenta el criterio de redondeo, ya que se debe tener en cuenta si se redondea a un número entero, o décimas, centésimas, etc. Existen dos criterios básicos:

— Si tenemos un número con una carga decimal del 0 al 4, este se elimina. Ejemplo: 23,3 redondeamos a 23. Si por el contario es un 9, 8, 7, 6 o 5 seguido de otras cifras, se aumenta en una unidad la cifra anterior, 23,8 pasaría a 24. Otros ejemplos: 32,234 quedaría 32,23 o 32,2; mientras que 32,276 sería 32,28 o 32,3, en función del criterio tomado si se ajusta a centésimas o décimas.

— Si el número a eliminar es un 5, solo o seguido de otras cifras no significativas, el error es el mismo si se aumenta en una cifra que si se elimina. En estos casos el criterio que se toma es redondear haciendo que la cifra anterior sea par, de esta manera cuando operamos con muchos números los errores se compensan unos a otros. Ejemplo: 45,5 pasaría a 46, mientras que 45,25 pasaría a 45,2.

*Truncamiento:* es un método de aproximación en el que se ha cortado el número en la cifra decimal deseada, sin tener en cuenta los dígitos posteriores, por ejemplo, 3,076923 = 3,076.

## 1.1.5. Variable o variable estadística

Si existe un concepto importante y a la vez complejo en estadística es el de variable, ya que la sola definición no abarca todas las notas características que se formulan alrededor de este concepto. Se entiende por **variable estadística** una característica o atributo que nos permite generalizar la descripción de los individuos de una muestra y, por tanto, de una población.

Se puede decir que una variable es un fenómeno observable que puede adoptar diferentes valores excluyentes entre sí para su análisis. Estos valores pueden ser finitos o infinitos, y dependiendo





de las características que puede adoptar, podemos distinguir múltiples acepciones del concepto (Kelinger, 1964; Edwards, 1968; Rodrigues, 1975; Arnau, 1978; Anderson y Borokowski, 1978; Álvaro y Garrido, 1995).

La clasificación y forma que puede adoptar la variable dependerá del criterio que estemos utilizando:

1. *Si se puede medir o no.* Como hemos dicho anteriormente, si es medible el fenómeno observado se denomina variable y sus valores son sus medidas. Edad, altura, peso... Mientras que si no son mensurables, estamos ante los denominados **atributos;** estos pueden ser ordenables o no ordenables. La transformación de una categoría o modalidad en una variable propiamente dicha se establece mediante una **relación biunívoca** entre las **modalidades** o **categorías** del atributo y un valor numérico aleatorio.
2. *En función del tipo de* ***escala***[5] ***nominal, orden, intervalo o razón.*** Se dividen en dos tipos: **cualitativas** (escala nominal) o **cuantitativas** (escalas de orden, intervalo y razón). Las cuantitativas a su vez pueden ser: **continuas** o **discretas,** mientras que las cualitativas siempre son discretas.
3. *En función del tiempo.* Pueden ser **variables temporales** o **atemporales.**
4. *En función del estudio que estemos realizando,* las variables se pueden comportar como: **variables dependientes** o **independientes.**
5. *En función de las características que se asocian a los valores que puede tomar,* las variables pueden ser **aleatorias,** cuando el resultado no se puede predecir *a priori.*

### Variables cualitativas

Se refieren a cualidad o cualidades de un fenómeno observable. Denotan **cualidad,** que es cada una de las circunstancias o caracteres naturales o adquiridos que distinguen a los objetos,

[5] Mirar la definición de escalas en este glosario (p. 26).

personas, organismos vivos y fenómenos[6]. Se trata de una propiedad o atributo. Cuando a las propiedades **(modalidades)** de los atributos les otorgamos un valor numérico aleatorio (y solo uno), pasan a denominarse variables cualitativas. En este tipo de variable se utiliza una escala nominal para medirla, se puede decir que damos nombre (nominamos) con valores numéricos las características de la variable. También son llamadas variables categóricas, representadas en una escala nominal, y pueden estar representadas numéricamente, aunque esos números sean una mera traducción de sus nombres, una codificación. Existen dos tipos de variables cualitativas: las dicotómicas o binarias (un ejemplo de este tipo de variables es el sexo) y las policotómicas (ejemplos pueden ser la raza, el tipo de alimentación, medios de transporte, etc.).

### Variables cuantitativas

Cuando estudiamos o investigamos un fenómeno observable cuyas características son numéricas, medibles, dichos fenómenos se denominan variables, concretamente variables cuantitativas. Para su medición, pueden utilizar escalas de orden, de intervalo o razón.

**Cuantitativo** es un adjetivo que está vinculado a la cantidad. Este concepto hace referencia a una cuantía, una magnitud, una porción o un número de cosas. Lo cuantitativo, por tanto, presenta información sobre una cierta cantidad. Este tipo de variables puede tomar dos formas diferentes en función del tipo de valores que tenga, denominadas continuas o discretas.

### Variables discretas

Denominamos variable **discreta** a aquella que entre dos valores consecutivos puede tomar a lo sumo un número finito de valores; dicho de otro modo, cuando entre dos valores consecutivos no es posible observar un valor intermedio o aquellas que se pueden describir con un número natural o con un número finito o

[6] Definición de tipo etimológico que aparece recogida en los diccionarios.





infinito numerable de valores diferentes. Un ejemplo es el número de estrellas de un hotel: se tiene una, dos, tres... pero no es posible tener una y media, Son ejemplos de este tipo de variables el número de habitantes de un país, el número de turistas en una ciudad o el número de objetos que se producen en una fábrica. Se puede afirmar que prácticamente la totalidad de las variables cualitativas son discretas, ya que sus características se asocian con un número real y solo uno, no teniendo en cuenta valores intermedios, ya que las características (modalidades) son únicas y excluyentes entre sí.

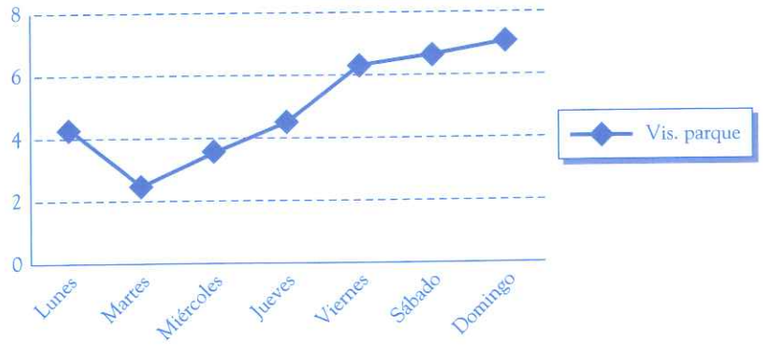
### Variables continuas

Denominamos variable **continua** a aquella en la cual siempre es posible encontrar un valor intermedio entre dos valores adyacentes cualesquiera de la variable, o bien aquella que puede tomar infinitos valores de un intervalo (teóricamente). Un ejemplo: la edad, nivel de ingresos, gasto de un turista; si la variable es «altura medida en centímetros de un grupo de personas», entre dos consecutivas siempre puede haber infinitas medidas.

### Variables temporales y atemporales

Denominamos **variables temporales o históricas** a aquellas en que reflejamos sus distintos valores en «momentos de tiempo» y que adoptan una forma de serie; la clave en este tipo de variables es el concepto de serie, ya que se debe ver como una nube de valores. Un ejemplo: la serie anual de personas alojadas en un hotel, la serie semanal o mensual de visitantes a un parque temático...

Denominamos **variables atemporales o de corte transversal** a aquellas que nos muestran un momento o período concreto y más o menos largo de tiempo. Siguiendo los ejemplos anteriores, el número de personas alojadas en el año 2015 en el mes de mayo en un alojamiento, o el número de visitantes de un parque temático en el mes de abril de 2015.




Si consideramos esta gráfica como la serie de valores que obtiene la variable «visitantes de un parque temático» en la semana primera de mayo, se trata de una variable temporal.

Sin embargo, si digo que la primera semana de mayo los visitantes de un parque temático han sido 3.490 personas, será una variable atemporal o de corte transversal.

### Variables dependientes e independientes

Cuando queremos investigar, nuestro objetivo suele ser explicar o analizar los cambios que se observan en una variable en relación con los datos obtenidos en otra u otras variables. A la variable cuyos cambios queremos explicar o analizar la denominamos **variable dependiente.**

A la variable o variables que se utilizan para explicar dichos cambios la denominamos **independiente.** Es decir, la independiente es la que influye en otra, la dependiente. En condiciones normales se suele decir que las variables independientes son las elegidas y/o manipuladas por el investigador para ver los cambios que estas producen en la dependiente, objeto de nuestro estudio.

## 1.1.6. Escalas de medida (medición)

Hablamos de **medición** cuando asignamos un valor numérico a un fenómeno siguiendo unas determinadas reglas. Por ejemplo,





decimos que medimos el color de los ojos cuando asignamos un valor numérico a cada uno de los diferentes colores: 1 al azul, 2 si es verde, 3 si es marrón...

A los instrumentos utilizados para este proceso los denominamos **escala de medida.** Como veremos, este concepto condiciona en parte el tipo de variable y los análisis que podemos hacer, ya que en función de las reglas de medida utilizadas, diferentes escalas posibles, los valores de la variable difieren indicando orden, dirección, distancia o simplemente que una característica es distinta a la otra. Esto ha hecho que se determine una tipología de escalas de medida: **nominales, ordinales, de intervalo y razón.**

### Escala nominal

Esta fórmula de medida solo nos permite establecer una diferencia entre una característica u otra de un fenómeno observado (objetos, sucesos, personas, atributos, fenómenos), una relación de igualdad o desigualdad entre aquello que estamos midiendo. No establece ninguna otra relación entre los valores. En el fondo se trata de otorgar un símbolo que puede ser sustituido por cualquier otro sin alterar el resultado de la medición; se trata de una clasificación de un nombre. Ello impide operar matemáticamente con los números, aunque, según Stevens (1959), estas variables así creadas pueden ser analizadas con los estadísticos de frecuencias (proporciones, moda y coeficiente de contingencia). El SPSS, sistema de análisis estadístico, establece como herramientas de análisis de variables cualitativas en escala nominal: **distribución de frecuencias** y **moda.** En cuanto a la representación, sería el **diagrama de barras.**

### Escala ordinal

Este tipo de escala nos permite establecer relaciones de orden entre los fenómenos que estamos midiendo, además de la igualdad/desigualdad. Es decir, si uno es mayor o menor que otro en función de su magnitud. No obstante, no se pueden hacer operaciones con los números (valores), ya que la distancia entre cada uno de ellos no tiene por qué ser la misma. Un ejemplo: en una carrera las diferencias entre el primero, el segundo y el tercero no suelen ser las mismas.

Este tipo de valores nos permite clasificar en un grupo de personas que realizan una actividad, por ejemplo, el número de errores cometidos, pero no tiene por ello que haber una pauta entre los puestos, es decir, el primero puede cometer «*n*» errores, el segundo «*p*» y el tercero «*m*», pero entre «*n*» y «*p*» no tiene por qué haber el mismo número de errores que entre «*p*» y «*m*». Nuevamente según Stevens, la estructura matemática es una función creciente; podemos utilizar los estadísticos media, percentil y correlaciones ordinales. El SPSS[7] establece como herramientas para variables cualitativas ordinales: **valores máximos y mínimos, la mediana, los cuartiles, percentiles y el rango intercuartílico.**

### Escala de intervalo

Este tipo de escala incluye las propiedades de las anteriores igualdad/desigualdad y orden; la diferencia está en que los intervalos entre los valores de la escala son iguales, lo que nos permite hacer operaciones de suma y resta, además de multiplicaciones y divisiones de las sumas y restas que hay entre los valores. No obstante, carecen del valor 0 absoluto al tratarse de mediciones continuas, en las que el valor 0 no es la ausencia o falta del fenómeno observado, sino un punto arbitrario elegido por el investigador; un ejemplo es el caso de la temperatura, en la que el valor 0 no significa que no exista.

No obstante, no podemos hacer operaciones de división ni multiplicación directamente con los valores de la escala, en el caso anterior no podemos decir que 20 grados es la mitad de 40 grados, ni que 40 sea el doble de 20 grados. Stevens afirma que con este tipo de mediciones podemos emplear la media, desviación típica, correlaciones, etc.

### Escala de razón

Se trata del nivel más alto de medición, ya que en este tipo de escala existe el llamado 0 absoluto: la ausencia del fenómeno medido. Esto nos permite la realización de cualquier tipo de algoritmo

[7] Uno de los paquetes estadísticos más comúnmente utilizados en Ciencias Sociales.





matemático con los valores de la escala. En las Ciencias Sociales es difícil encontrar este tipo de valores, son del tipo longitud, peso, intensidad... Dadas las características de este tipo de escala, podemos hacer cualquier análisis estadístico.

El SPSS engloba las herramientas para el análisis de variables cuantitativas de intervalo o razón: **la media, el rango, la varianza, la desviación típica, el coeficiente de variación, la asimetría y la curtosis.** Y como gráficos, añade el **histograma** como el más representativo.

# 2 Medidas de la variable. Cómo se deben presentar los datos estadísticos de una variable para su correcta interpretación

Los datos obtenidos tras las observaciones realizadas en una muestra o población sobre una variable que queremos estudiar forman lo que llamamos serie estadística.

Estas series, para poder ser analizadas y comprendidas (ser útiles para la toma de decisiones), deben ser presentadas de forma ordenada y clasificada, deben seguir unas normas y tener un mismo lenguaje, de tal manera que cualquier persona pueda ser capaz de comprenderlas y llegar a las mismas conclusiones con los datos que se presentan.

Para que esto sea posible construimos las **tablas estadísticas,** que nos permiten presentar los resultados obtenidos y nos dan una primera visión del comportamiento de la variable que estamos analizando.

Uno de los conceptos más importantes en esta fase del estudio de una variable es la **distribución de frecuencias,** que consiste en observar, clasificar y ordenar las repeticiones de ciertos valores de la variable. Se puede definir como el conjunto de valores que toma una variable de forma ordenada, y acompañada de sus frecuencias absolutas.

Por todo ello, es importante guardar una notación estadística consensuada, que nos ayude a seguir cualquier estudio de una variable; la notación más habitual es la siguiente:

- «$X$» (mayúscula): la variable o característica que estamos estudiando.
- «$x_i$» (minúscula, normalmente): es el valor que toma la variable o característica $X$ en el sujeto/individuo.
- «$k$»: el número de valores distintos que toma una variable.
- «$n_i$»: número de veces o frecuencia con la que aparece un valor determinado $x_i$.
- «$N$»: número total de unidades en las que hacemos nuestro estudio o medición, o disponemos de datos.
- «$n$»: se utiliza normalmente minúscula para hacer referencia al total de los datos; no obstante, se utiliza «$N$» para el total de datos referidos a una población y «$n$» cuando los datos son de la muestra de dicha población.





Ejemplo. Queremos estudiar el tiempo que se tarda en realizar una tarea en nuestra empresa; dicha tarea es realizada por 10 personas de forma habitual. Los datos que hemos recogido son los siguientes:

- Sujeto 1 = 6´; Sujeto 2 = 7´; Sujeto 3 = 5´; Sujeto 4 = 6´; Sujeto 5 = 5´; Sujeto 6 = 7´; Sujeto 7 = 5´; Sujeto 8 = 6´; Sujeto 9 = 6´; Sujeto 10 = 5´ (el tiempo en segundos).
- «$X$»: la variable de estudio: es el tiempo de realización de la tarea.
- «$N$»: 10, el número de sujetos del estudio.
- «$k$»: los posibles valores de la variable, en nuestro caso se trata de una variable continua, por tanto puede tomar infinitos valores $\{1, 2, 3, 4, ..., k\}$.
- «$x_i$»: los valores de la variable $x_1 = 1$; $x_2 = 2$; $x_3 = 3$; ... $x_i = k$.
- «$n_i$»: las veces que aparece cada uno de los valores de la variable.

Nuestra tabla estadística de frecuencias quedaría de la siguiente forma:

| $x_i$ | $n_i$ |
|---|---|
| 5 | 4 |
| 6 | 4 |
| 7 | 2 |
| | $N = 10$ |

## 2.1. DEFINICIONES EN DISTRIBUCIONES DE FRECUENCIAS UNIDIMENSIONALES (UNA SOLA VARIABLE)

### 2.1.1. Frecuencia absoluta

La frecuencia absoluta representa el número de veces que un valor dado se repite en un conjunto de datos, es decir, cuántas

veces aparece un dato concreto entre todos los valores que toma la variable estadística. Se suele notar como $f$. También se puede notar como $n_i$.

En nuestro ejemplo sería:

$$n_5 = 4; \quad n_6 = 4 \quad \text{y} \quad n_7 = 2$$

Es decir, en nuestro estudio la frecuencia absoluta de efectuar la tarea en 5 segundos es 4.

### 2.1.2. Frecuencias relativas

Las frecuencias relativas representan la proporción de individuos de una muestra dada que presentan una característica determinada y se calcula como el cociente entre la frecuencia absoluta y la cantidad de datos observados.

Normalmente se expresa en tantos por ciento.

En nuestro ejemplo sería: el valor

$$f_5 = 4/10 = 0{,}4 \times 100 = 40\,\%$$

Es decir, el 40 % de nuestros trabajadores realizan la tarea estudiada en 5 segundos.

### 2.1.3. Frecuencias acumuladas

Si consideramos una frecuencia acumulada como la suma de las frecuencias de todos los datos anteriores al analizarlo, podemos definir frecuencia absoluta acumulada como el número de datos inferiores o iguales al considerado y frecuencia relativa como la proporción de datos inferiores o iguales al considerado. Estas dos frecuencias tendrán valor en la medida en que se pretenda establecer una escala ordinal entre las categorías de la variable.

El concepto «acumulado/a» en estadística hace referencia en un determinado valor a la frecuencia absoluta de ese valor de la variable más la suma de los valores anteriores. La notación de la **frecuencia absoluta acumulada** suele ser $N_i$ (en mayúscula), mientras que la **frecuencia relativa acumulada** es $F_i$. Este tipo de valores tiene gran interés, ya que se utiliza para algunos cálculos estadísticos más avanzados.





### 2.1.4. Tabla de frecuencias

La información aportada por cada una de las diferentes frecuencias consideradas puede no ser totalmente útil *per se*, por lo que se acostumbra a ordenar y resumir la información en tablas, que presenten ordenados en forma creciente de valor los datos, y que permitan hacer una análisis más pormenorizado del problema.

Siguiendo nuestro ejemplo, a la hora de representar los datos en la tabla de frecuencias, quedaría de la siguiente forma:

| $x_i$ | $n_i$ | $N_i$ | $f_i$ | $F_i$ |
|---|---|---|---|---|
| 5 | 4 | 4 | 40 % | 40 % |
| 6 | 4 | 8 | 40 % | 80 % |
| 7 | 2 | 10 | 20 % | 100 % |
| | $N = 10$ | | | |

## 2.2. TRATAMIENTO DE LOS VALORES DE LA VARIABLE

### Intervalos de clase

Cuando analizamos una variable continua o una variable discreta que toma una gran cantidad de valores muy cercanos entre sí es operativamente muy útil agrupar los datos observados en grupos que llamamos intervalos de clase, de tal manera que cada intervalo equivaldría a una modalidad o característica de la variable estudiada.

La amplitud del intervalo o de la clase vendrá dada por la diferencia entre los valores mayor y menor de los datos que forman dicho intervalo. En cada uno de estos intervalos a sus valo-

res extremos (superior e inferior) los llamamos extremos del intervalo, o extremos de la clase, y al valor o punto medio de ese intervalo lo llamaremos marca del intervalo, o marca de la clase, pudiendo emplear estos valores como representantes para posteriores cálculos.

Gráficamente los intervalos se representan entre paréntesis o entre corchetes. Los paréntesis indican que los valores cercanos a ellos no están incluidos, mientras que en los corchetes los valores cercanos sí están incluidos. Así hablaremos, respectivamente, de intervalos abiertos o cerrados o abiertos por un extremo y cerrados por el otro. Los intervalos también serán susceptibles de poder resumirse en una tabla de frecuencias.

Así pues, sumaremos las frecuencias de todos los datos incluidos en cada intervalo para obtener la frecuencia del intervalo o frecuencia de la clase y a continuación podremos calcular el resto de frecuencias. La tabla estadística en este caso tendrá en su primera columna o fila los valores de los extremos del intervalo de la clase y en las restantes columnas o filas se dispondrán los datos de forma similar a como se hace en las tablas de valores discretos.

La decisión de agrupar los datos en intervalos debe tener en cuenta algunas consideraciones más o menos subjetivas, en función del tipo de análisis que se quiera realizar, como el número de intervalos, la amplitud (igual o diferente) de los intervalos y el extremo inferior del primer intervalo.

El número de intervalos dependerá del tamaño de la muestra y de la dispersión de los datos, por lo que un mayor tamaño muestral y/o dispersión implicará mayor número de intervalos. La recomendación más habitual es construir tantos intervalos como el número entero más próximo a la raíz cuadrada de la frecuencia absoluta, no superando en ningún caso los 20 intervalos. En lo que se refiere a la amplitud del intervalo, asumiendo que salvo indicación metodológica en contra, siempre tendrán la misma, se calculará redondeando (por exceso) el cociente entre el recorrido de la variable y el número de intervalos que se van a construir. El valor del extremo inferior del primer intervalo, usualmente, se elige como un valor un poco menor que el mínimo de los valores presentes en la muestra.





## 2.3. TIPOS DE DISTRIBUCIONES DE FRECUENCIAS (TIPO I, II Y III)

Según el número de observaciones y según el recorrido de la variable estadística, podemos clasificar las tablas de frecuencia en tres categorías:

*a)* Tablas tipo I.
*b)* Tablas tipo II.
*c)* Tablas tipo III.

**Tablas tipo I:** cuando el tamaño de la muestra y el recorrido de la variable son pequeños, como el caso de nuestro ejemplo, tenemos una muestra de diez personas. En este caso no hay que hacer nada especial, simplemente anotarlas de manera ordenada en filas o columnas.

Ejemplo: 6; 7; 5; 6; 5; 7; 5; 6; 6; 5

**Tablas tipo II:** cuando el tamaño de la muestra es grande y el recorrido de la variable es pequeño, por lo que hay valores de la variable que se repiten. Por ejemplo: si queremos saber el tamaño de las familias de los trabajadores de una empresa (40 familias) obtenemos la siguiente tabla:

**40 FAMILIAS**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 3 | 2 | 2 | 1 | 2 | 4 | 2 | 3 | 1 |
| 2 | 3 | 2 | 1 | 2 | 1 | 3 | 4 | 2 | 2 |
| 2 | 2 | 1 | 2 | 1 | 2 | 1 | 3 | 2 | 2 |
| 3 | 2 | 3 | 1 | 2 | 4 | 2 | 1 | 4 | 1 |

Podemos observar que la variable toma valores comprendidos entre 1 y 4, por lo que precisaremos una tabla en la que resumamos estos datos, quedando la siguiente tabla:

| Tamaño familiar | $n_i$ |
|---|---|
| 1 | 10 |
| 2 | 19 |
| 3 | 7 |
| 4 | 4 |
| | $N = 40$ |

**Tablas tipo III:** cuando el tamaño de la muestra y el recorrido de la variable son grandes, por lo que será necesario agrupar en intervalos los valores de la variable. Por ejemplo: si a un grupo de 24 personas les preguntamos el dinero que piensan gastar en sus vacaciones, nos encontramos con los siguientes datos:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 175 | 180 | 325 | 1.680 | 605 | 785 | 1.595 | 2.300 | 2.500 | 1.200 | 120 |
| 675 | 500 | 375 | 1.500 | 205 | 985 | 1.850 | 125 | 315 | 425 | 560 | 1.200 |

Evidentemente la variable estadística tiene un **recorrido** muy grande, por lo que si queremos hacer una tabla con estos datos tendremos que tomar intervalos.

### Rango recorrido o amplitud total

En una distribución, si ordenamos previamente los valores de menor a mayor, se define como la diferencia entre el mayor y el menor valor de la distribución. Se nombra $R$ y se obtiene mediante:

$$R_x = x_r - x_1 = \text{máx}\,\{x_i\} - \text{mín}\,\{x_i\} \quad \text{para} \quad 1 \leqslant i \leqslant r$$

o lo que es lo mismo:

$$R_x = x_n - x_i$$

En nuestro ejemplo: 2.500 – 120 = 2.380.





Para decidir la amplitud de los intervalos necesitaremos decidir cuántos intervalos queremos. Normalmente se suele trabajar con no más de 10 o 12 intervalos.

$$\textbf{Amplitud} = 2.380/10 = 238$$

Por lo que tomaremos intervalos de amplitud 250.
Debemos tener en cuenta las siguientes consideraciones:

- Tomar pocos intervalos implica que la «pérdida de información» sea mayor.
- Los intervalos serán siempre cerrados por la izquierda y abiertos por la derecha $[L_{i-1} - L_i)$.
- Procuraremos que en la decisión de intervalos los valores observados no coincidan con los valores de los extremos del intervalo y, si esto ocurre, que no sea en más de un 5 % del total de observaciones.

Con estas recomendaciones tendremos la siguiente tabla:

| $[L_{i-1} - L_i)$ | $n_i$ |
|---|---|
| [0-250) | 5 |
| [250-500) | 5 |
| [500-750) | 4 |
| [750-1.000) | 2 |
| [1.000-1.250) | 2 |
| [1.250-1.500) | 0 |
| [1.500-1.750) | 3 |
| [1.750-2.000) | 1 |
| [2.000-2.250) | 0 |
| [2.250-2.500) | 1 |
| [2.500] | 1 |

## 2.4. MEDIDAS O ESTADÍSTICOS EN ANÁLISIS UNIVARIADOS

**Medidas descriptivas.** Tras la elaboración de las tablas de frecuencias el siguiente paso es el análisis de los datos recogidos en esas tablas, con una serie de valores descriptivos que resumen aspectos fundamentales de la muestra. Estas medidas se pueden clasificar en tres (o cuatro) grandes grupos. Así, hablaremos de medidas de posición, de centralización y de dispersión (y de forma) y cada una de ellas medirá alguna característica de la muestra, como pueden ser la simetría, la agrupación en torno a valores centrales, etc.

**Medidas de centralización.** Permiten describir cómo se organizan los datos de una muestra en torno a valores cercanos a un valor central y son, muy a menudo, representantes del conjunto de la muestra, aunque es probable que ninguno de los valores coincida con valores muestrales. Nos dan una idea rápida sobre la simetría de los datos respecto a un valor central y nos permiten tener una idea previa sobre la forma de la gráfica que define la distribución de los datos estudiados. Estas medidas son la media (aritmética, truncada, recortada, cuadrática, geométrica y armónica), la mediana y la moda.

**Medidas de posición.** Estas medidas nos permiten analizar cómo se distribuyen los datos muestrales si hacemos grupos con el mismo número de datos, para lo cual se divide el conjunto ordenado de datos en partes iguales o intervalos con el mismo número de datos, llamado cada uno cuantil. Los más utilizados son los cuartiles, los deciles y los percentiles, derivados de dividir el conjunto de datos en 4, 10 o 100 grupos.

**Medidas de dispersión.** Estas medidas intentan describir la dispersión o concentración del conjunto de datos con respecto a un valor central. La dispersión sería equivalente a la desviación o grado de variabilidad de un conjunto de observaciones. Las más utilizadas son el rango o recorrido, el rango o recorrido intercuartílico, la desviación media, la varianza, la desviación típica o estándar y el coeficiente de variación.

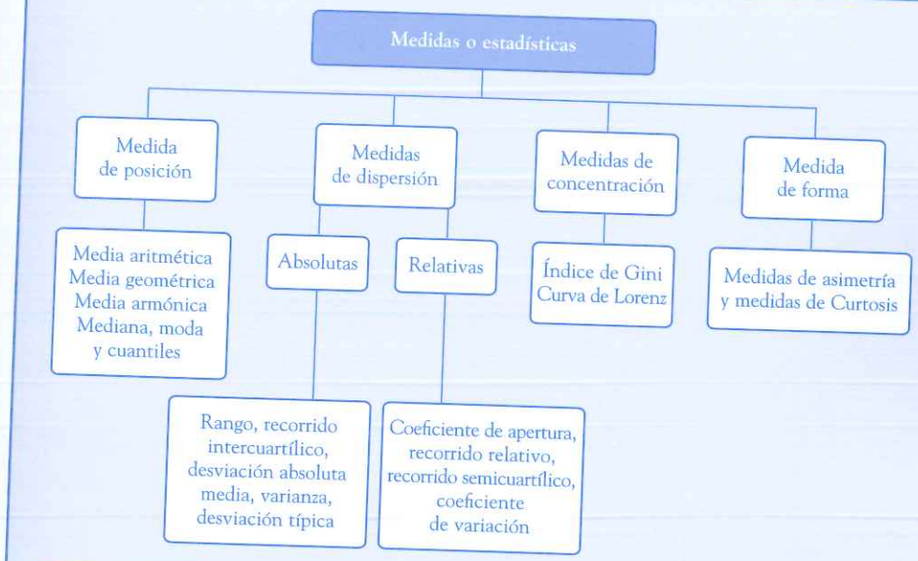




**Medidas de forma.** Como su nombre indica, nos muestran el aspecto que tiene la distribución de frecuencias, de tal forma que hablaremos de simetría (o asimetría) cuando la forma de la distribución es igual a ambos lados de un valor central y de curtosis o apuntamiento, que indica el grado de aplastamiento de una curva.

**ESQUEMA 2.1**

**ESQUEMA ADAPTADO DE J. SANTOS, A. MUÑOZ ALAMILLOS Y A. MUÑOZ MARTÍNEZ (2007)**



# 3 Representación gráfica de las distribuciones

La representación gráfica de variables estadísticas es una manera muy útil de describir datos correspondientes a una distribución estadística, ya que permite, de un solo golpe de vista, obtener una visión general, sin necesidad de realizar un completo análisis condensando de manera inteligible toda la información. Siguiendo la categorización de las variables, podríamos realizar la siguiente clasificación, teniendo en cuenta que a la hora de elegir el diagrama más útil hay que tener en cuenta otras consideraciones y no solo el tipo de variable:

— Representación gráfica de variables cualitativas:

- Diagrama de barras.
- Diagrama de sectores.
- Pictograma.
- Cartograma.

— Representación gráfica de variables cuantitativas continuas:

- Diagrama de barras.
- Histograma.
- Polígono de frecuencias.
- Polígono de frecuencias acumuladas.

— Representación gráfica de variables cuantitativas discretas:

- Diagrama de barras.
- Polígono de frecuencias acumuladas.
- Diagrama de cajas y bigotes.
- Diagrama de tallo y hojas.
- Series temporales.

## 3.1. REPRESENTACIÓN GRÁFICA DE VARIABLES CUALITATIVAS

### 3.1.1. Diagrama de barras

Se emplea de forma habitual para variables discretas en distribuciones de frecuencias con datos sin agrupar. En este tipo de



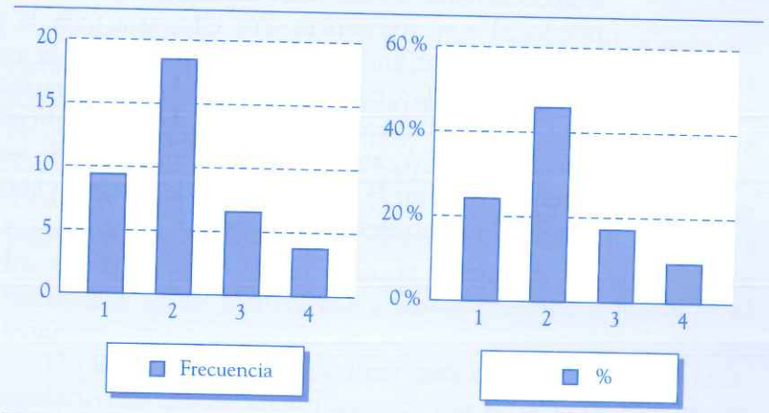


gráfico se representan tantas barras como valores tome la variable a estudiar y en una altura proporcional a la frecuencia con que aparece en la distribución. Para construirlo se emplea un eje de coordenadas cartesianas, de tal manera que en el eje de abscisas (eje de las X) se representan los valores o modalidades que toma la variable y en el eje de ordenadas (eje de las Y) las frecuencias, de tal manera que la altura de la barra será proporcional a la frecuencia elegida. Son gráficos fácilmente manipulables y hay que ser cuidadoso con las escalas empleadas y con los puntos de origen. En el diagrama de barras es relativamente fácil y visual comparar todas las categorías entre sí.

Siguiendo con el ejemplo del tamaño familiar de los trabajadores de una empresa (cap. 2), en el gráfico de barras, podemos observar las diferencias de frecuencias de los valores de la variable y poder sacar alguna conclusión inicial. En los diagramas de barras a menudo en el eje «y» (vertical/ordenadas) podemos encontrar en lugar de frecuencias el porcentaje de la variable.

**FIGURA 3.1**

**NÚMERO DE MIEMBROS FAMILIARES DE LOS TRABAJADORES DE UNA EMPRESA**



| Tamaño familiar | $n_i$ | % |
|---|---|---|
| 1 | 10 | 25 |
| 2 | 19 | 47 |
| 3 | 7 | 18 |
| 4 | 4 | 10 |
| | $N = 40$ | 100 |

## 3.1.2. Diagrama de sectores

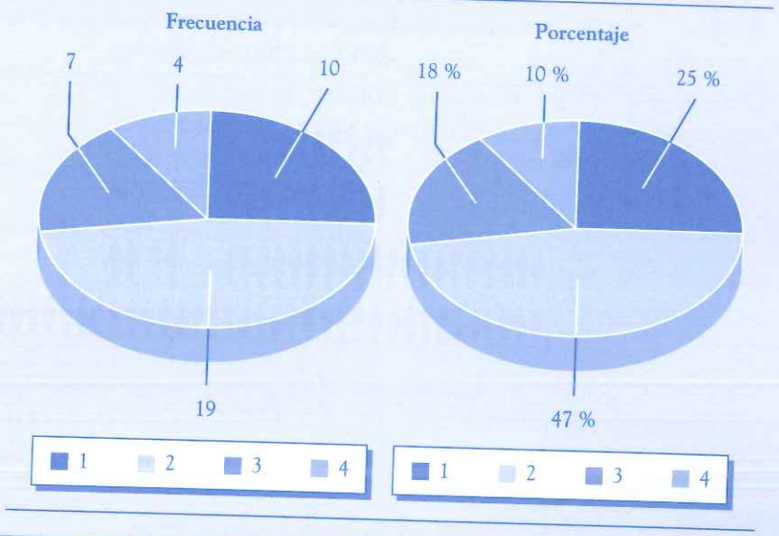
La idea de este tipo de gráfico es esencialmente la misma que la del gráfico de barras y consiste en repartir de manera proporcional a la frecuencia con que aparecen las diferentes modalidades de una variable la superficie de un círculo. En este tipo de gráficos solamente se representa una serie.

Para su construcción se reparte el círculo en partes proporcionales a las frecuencias relativas de los valores de la variable, correspondiéndole a cada valor un sector circular de amplitud proporcional a su frecuencia. Por ello, si a toda la población le corresponde una amplitud de 360°, para hallar la amplitud de un valor o categoría $x_i$ cuya frecuencia relativa es $f_i$ multiplicamos 360 por $f_i$ y el resultado será la amplitud del sector circular correspondiente al valor o categoría $x_i$. En el diagrama de sectores es muy sencillo visualizar la relación de cada categoría con el total. Tomando el mismo ejemplo anterior:





FIGURA 3.2

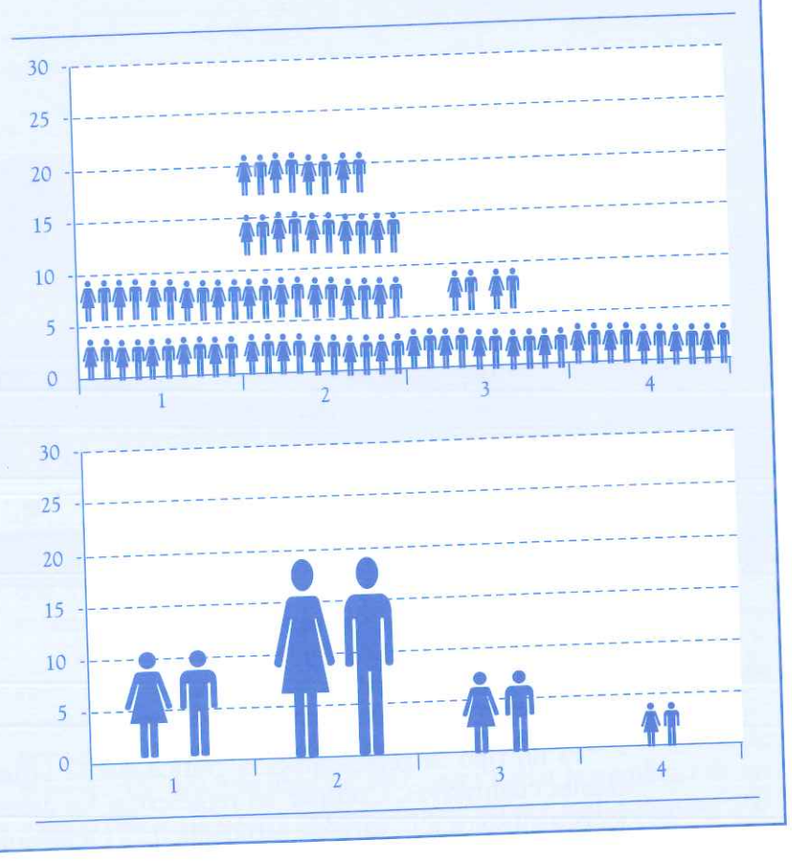



## 3.1.3. Pictograma

Es un tipo de representación gráfica que se utiliza para variables cualitativas. Consiste en representar los datos mediante iconos alusivos a la variable estudiada. Los pictogramas pueden resultar muy expresivos, pero muy poco precisos. Básicamente hay dos clases de pictogramas: aquellos en los que se utiliza un dibujo para representar la variable estadística y esta se repite tantas veces como haga falta (frecuencia absoluta), y otros en los que el icono utilizado varía de tamaño dependiendo de su frecuencia, de tal manera que a mayor frecuencia, mayor es el icono.

El pictograma se puede utilizar para expresar atributos, utilizando iconos que se identifiquen con la variable de estudio; en nuestro caso un pictograma que pueda identificarse con la unidad familiar.

**FIGURA 3.3**




## 3.1.4. Cartogramas

Los **cartogramas** son un tipo de gráficos que representan variables cualitativas/atributos, con las diferentes modalidades en marcos geográficos donde han sido medidas dichas variables. Los cartogramas no son mapas, ya que no representan el espacio geográfico, sino cambios en el tamaño de los elementos dependiendo de un atributo concreto. Ello hace que algunos cartogramas parezcan muy similares al espacio geográfico que estamos representando, o bien que no se



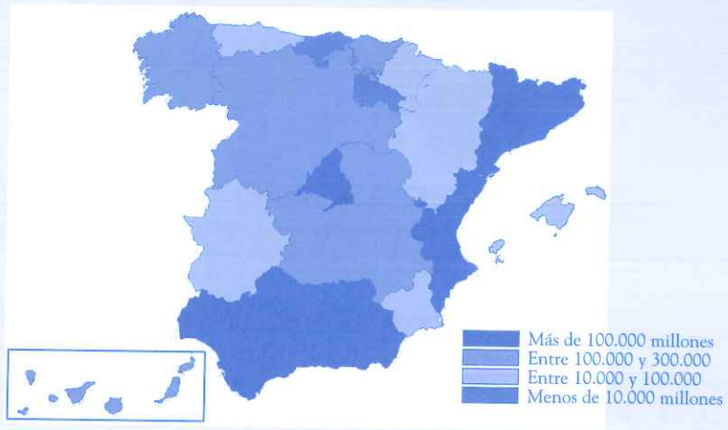


parezcan en nada. En función de esta característica, podemos encontrarnos con tres tipos diferentes de cartogramas:

1. Con contigüidad: cuando el mapa se parece al espacio geográfico convencional.
2. Sin contigüidad: cuando aunque se represente con una imagen relativamente semejante al espacio geográfico, no tiene por qué existir conectividad en todos los espacios geográficos.
3. Cartogramas de Dorling: estos no mantienen ni la forma ni los contornos de los elementos, normalmente se modifican usando círculos cuyo tamaño es proporcional a la frecuencia.

En este primer ejemplo se observa un cartograma con contigüidad, en el que se representa el gasto en prestación por desempleo de España[1]. Los colores representan las modalidades de la variable.

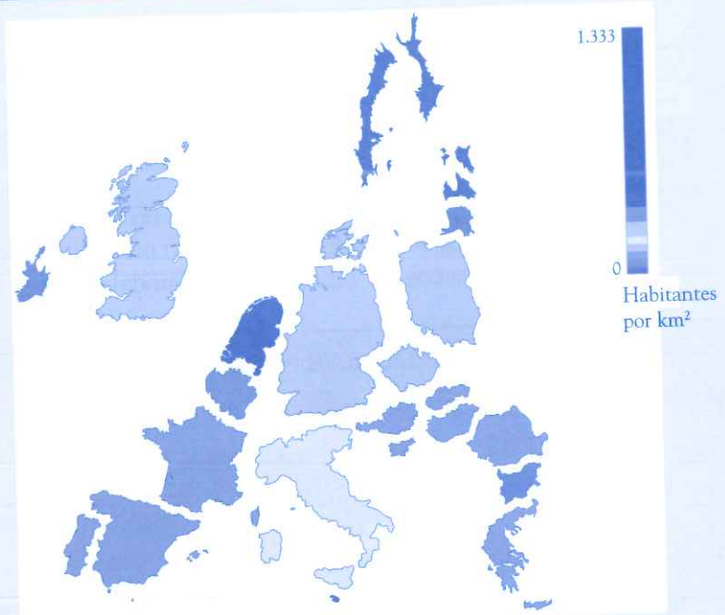
**GASTO TOTAL EN PRESTACIÓN POR DESEMPLEO** **FIGURA 3.4**




[1] Imágenes recuperadas de http://analisisydecision.es/wp-content/uploads/2010/06/mapa_espana_excel_2.JPG.

En este caso nos encontramos con un cartograma sin contigüidad; en ocasiones es difícil reconocer el área que se determina en función de las medidas de la variable. En el caso concreto, tenemos la distribución de la población de la UE[2].

**FIGURA 3.5**



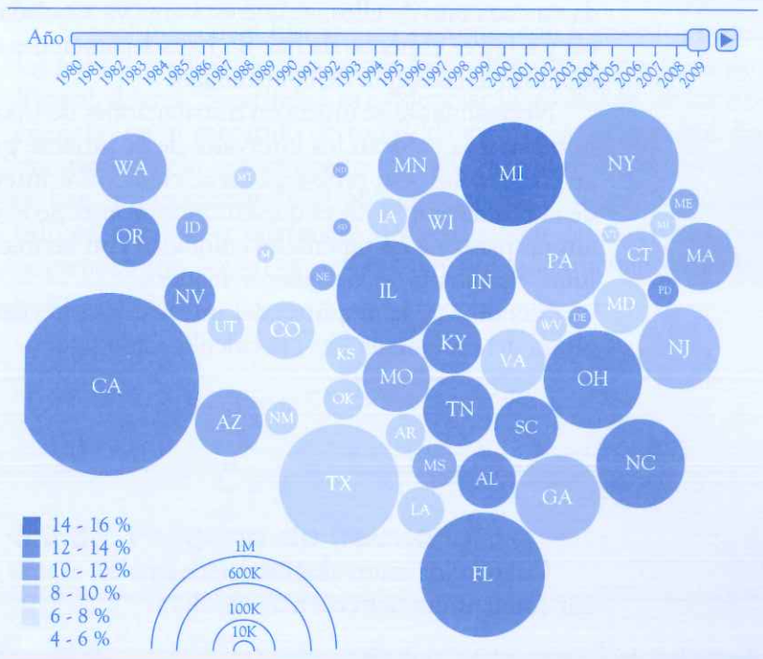

Este cartograma Dorling representa las tasas de desempleo de EE. UU. desde el año 1980 hasta 2009[3]. En él se observa que no hay una forma definida ni representan espacios geográficos. Los círculos muestran en tamaño las frecuencias de las variables por estados.

[2] https://www.pngfuel.com/.
[3] Imagen recuperada de: Data Driven, via Chart Porn, junio 2010.





**FIGURA 3.6**




## 3.2. REPRESENTACIÓN GRÁFICA DE VARIABLES CUANTITATIVAS CONTINUAS

### 3.2.1. Histograma

Este tipo de gráfico se suele comparar al gráfico de barras, con la salvedad de que en el histograma las barras se juntan, ya que está pensado para variables cuantitativas continuas, en las que los datos están agrupados.

Se parte de un eje de coordenadas cartesianas y se construyen rectángulos de base equivalente a la amplitud de los intervalos considerados y de altura proporcional a la frecuencia absoluta o

relativa. Hay que tener en cuenta que en un histograma lo más importante no es la altura de los rectángulos, sino el área encerrada en cada uno de ellos, ya que esta área va asociada a la frecuencia y a la vez el hecho de que no haya separaciones entre ellos va asociado al concepto de continuidad.

Normalmente se utiliza en distribuciones de tipo III; en el eje horizontal se colocan los intervalos de la variable y en el vertical las frecuencias. Un problema surge cuando los intervalos no son de la misma amplitud, ya que entonces es preciso el cálculo de las alturas para que las superficies coincidan con las frecuencias absolutas. Se calcula de la siguiente manera:

Siendo «$a_i$» la amplitud del intervalo, su frecuencia «$n_i$» y la altura del rectángulo «$h_i$», se calcula como sigue:

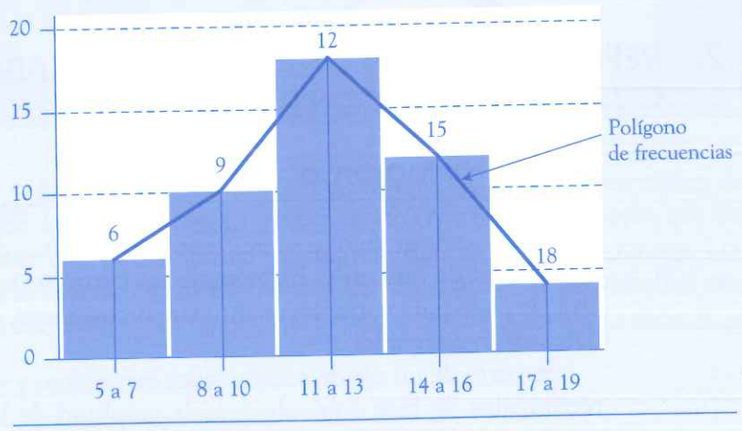
$$a_i = L_i - L_{i-1}$$

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

Cuando los intervalos son todos iguales, es más cómodo colocar como altura la frecuencia absoluta.

**FIGURA 3.7**

**HISTOGRAMA**








### 3.2.2. Polígono de frecuencias

Si partimos de un histograma y unimos con segmentos rectos los puntos medios de cada rectángulo, obtenemos una línea poligonal abierta. Su utilidad es evidenciar la evolución de las frecuencias en el recorrido de todos los valores de la variable. En nuestro ejemplo se han puesto los valores medios del intervalo y se han unido formando el polígono de frecuencias. Se suelen utilizar, para hacer referencia al punto medio del intervalo, dos nomenclaturas: «$x_i$» o «$m_i$». También se denomina marca de clase y se halla:

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

### 3.2.3. Polígono de frecuencias acumuladas

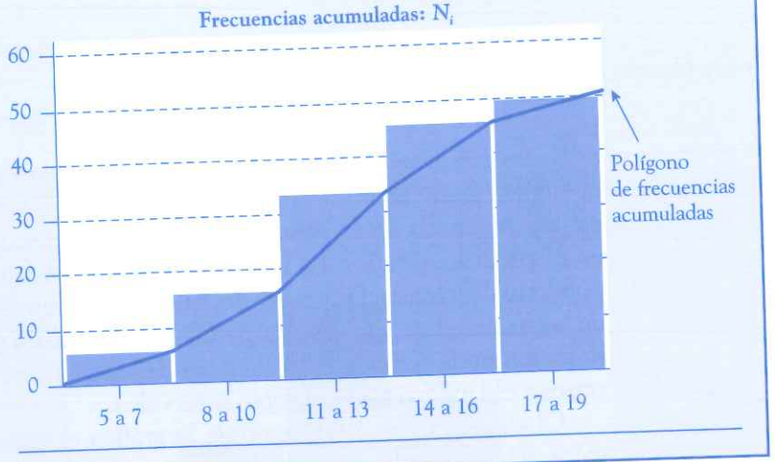
Como su propio nombre indica, este tipo de gráfico se basa en el empleo de frecuencias acumuladas, absolutas o relativas, es decir, a partir de las frecuencias acumuladas, cuyo cálculo se ha visto en el capítulo anterior. La forma de elaborarlo es similar a la del polígono de frecuencias y parte del histograma, pero con el empleo de las frecuencias acumuladas, de tal manera que en cada intervalo incluiremos la suma de frecuencias de todos los intervalos inferiores a él (incluyendo el propio intervalo).

Obviamente, en el extremo de la gráfica es donde se dará el valor más alto de la variable, por lo que su utilidad radica en conocer la evolución de la variable en la distribución. Se construye uniendo los valores inicial y final de cada intervalo de frecuencias acumuladas.

**TABLA DE DATOS EJEMPLO GRÁFICOS**

| $L_{i-1} - L_i$ | $n_i$ | $N_i$ | $m_i/X_i$ |
|---|---|---|---|
| 5-7 | 6 | 6 | 6 |
| 8-10 | 10 | 16 | 9 |
| 11-13 | 18 | 34 | 12 |
| 14-16 | 12 | 46 | 15 |
| 17-19 | 4 | 50 | 18 |

**FIGURA 3.8** HISTOGRAMA DE FRECUENCIAS ACUMULADAS. MISMOS DATOS QUE EN EL GRÁFICO DE HISTOGRAMA DE FR.




## 3.2.4. Pirámide poblacional

Un caso especial de histograma es la pirámide poblacional; en realidad se trata de dos histogramas colocados verticalmente, en ella se representan habitualmente dos variables: el sexo y la edad.



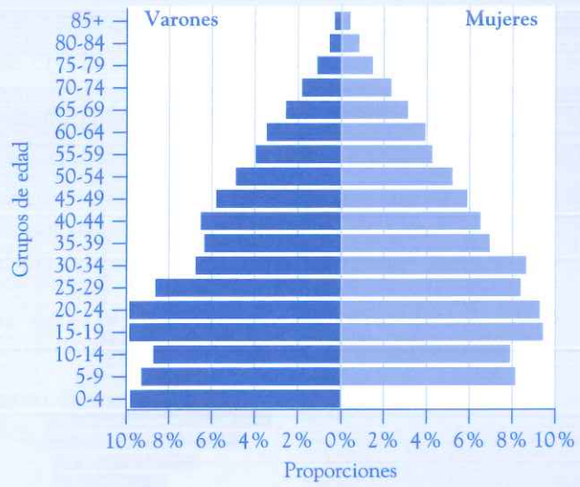


En cada lado de la pirámide se colocan los valores de la variable sexo, es decir, hombres y mujeres, y se hacen coincidir las barras por franja de edad. Se trata de una representación muy fácil de ver y muy intuitiva a la hora de interpretar y comparar con otras pirámides.

Los distintos tipos de pirámides de población son: **progresiva,** de base ancha y cima pequeña, **regresiva,** de base más estrecha que el centro y cima relativamente ancha, y **desequilibrada,** cuando existe una desproporción tanto en lo que respecta a la composición según sexo como en lo que respecta a las edades y casi siempre se produce una combinación de las dos posibilidades.

Los ejemplos que siguen ilustran este tipo de gráficos[4]:

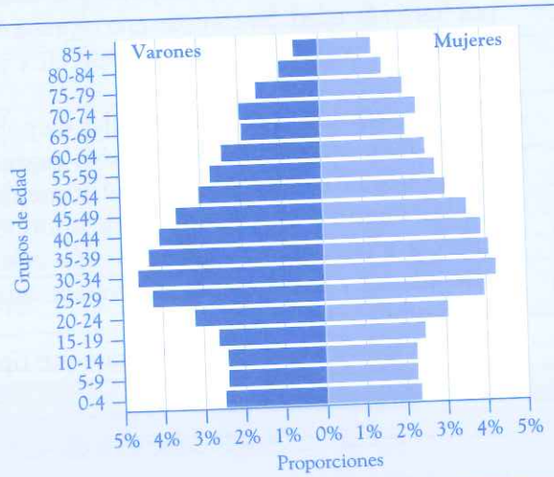
**FIGURA 3.9** PIRÁMIDE DE POBLACIÓN EN ESPAÑA, AÑO 1950




FUENTE: Instituto Nacional de Estadística. Censo de 1950.

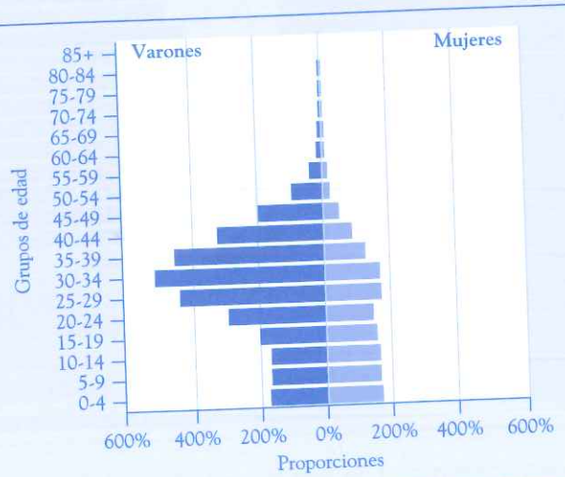
[4] Imagen 1 y 2 recuperadas: Instituto Nacional de Estadística. Censo 1 de enero 1960 y 2007.Imagen 3: ONU 2004.

**FIGURA 3.10** PIRÁMIDE DE POBLACIÓN EN ESPAÑA, AÑO 2007




FUENTE: Instituto Nacional de Estadística. Censo a 1 de enero de 2007.

**FIGURA 3.11** PIRÁMIDE DE POBLACIÓN, EMIRATOS ÁRABES UNIDOS, 2005




FUENTE: ONU (World Population Prospects: The 2004 Revision).





Como se puede ver, en la tabla de los Emiratos Árabes existe un gran desequilibrio entre la población masculina y femenina, además de una concentración muy alta de varones en edades productivas. Las dos primeras muestran la evolución de la población española.

## 3.2.5. Diagrama de tallo y hojas

Se puede considerar una variante del diagrama de barras, pero empleando un dígito como marca; es una técnica de recuento y ordenación de datos. Tiene la peculiaridad de aunar la visualización global de la distribución con la individualidad de los datos numéricos. Para su elaboración se separa la cifra de las unidades de cada dato, que constituirán las hojas, de las demás cifras, que serán el tallo. Se ordenan ascendentemente en una columna los tallos y en cada tallo se colocan a su derecha o izquierda, en fila, sus hojas o unidades correspondientes, también en orden creciente. Cada tallo, que se representa una sola vez, define una clase y el número de hojas representa la frecuencia de cada clase.

EJEMPLO. Representar la siguiente distribución de frecuencias: tiempos empleados en atención al cliente por el personal de tierra en un aeropuerto (en segundos):

32 23 25 33 41 20 44 36 27 55 38 40 36 26 37 21 30 34 29 51

Primero se toman los tallos, en nuestro ejemplo 3, 2, 4 y 5, que ordenados 2, 3, 4 y 5, seguido añadimos cada hoja a su tallo:

| Tallo | Hojas |
|---|---|
| 2 | 3 5 0 7 6 1 9 |
| 3 | 2 3 6 8 6 7 0 4 |
| 4 | 1 4 0 |
| 5 | 5 1 |

Reordenamos las hojas y queda:

| Tallo | Hojas |
|---|---|
| 2 | 0 1 3 5 6 7 9 |
| 3 | 0 2 3 4 6 6 7 8 |
| 4 | 0 1 4 |
| 5 | 1 5 |

## 3.2.6. Diagrama de caja y bigotes

A veces se llama diagrama de la mediana en recuadro, ya que representa los siguientes cinco valores característicos de una distribución estadística: mínimo (mín.); primer cuartil ($Q_1$); mediana (med.) o segundo cuartil ($Q_2$); tercer cuartil ($Q_3$), y máximo (máx.). Con ello da una idea mejor que otros gráficos sobre el sesgo y la dispersión de esa distribución. Para su construcción elaboramos un rectángulo (caja) en el que los lados estrechos representan los cuartiles $Q_1$ y $Q_3$, respectivamente, quedando entre ellos el 50 % de todos los datos; en el interior de esta caja una línea simboliza la mediana y de la caja parten dos segmentos laterales (bigotes), cuyos extremos corresponden a los valores mínimo y máximo, que se encuentran a menos de 1,5 veces el recorrido intercuartílico (*RIQ*). Si algún dato se encuentra a más de este último valor (1,5 × *RIQ*), se considera que es una observación atípica *(outlier)* y se representa individualmente fuera de estos límites por puntos. Este gráfico es muy útil para representar las diferencias entre grupos (Segovia y Rico Romero, 2009).

**Tomemos el ejemplo de los tiempos empleados en atención al cliente de un aeropuerto[5]:**

32 23 25 33 41 20 44 36 27 55 38 40 36 26 37 21 30 34 29 51

[5] Estos valores se desarrollarán en el capítulo de medidas de posición.

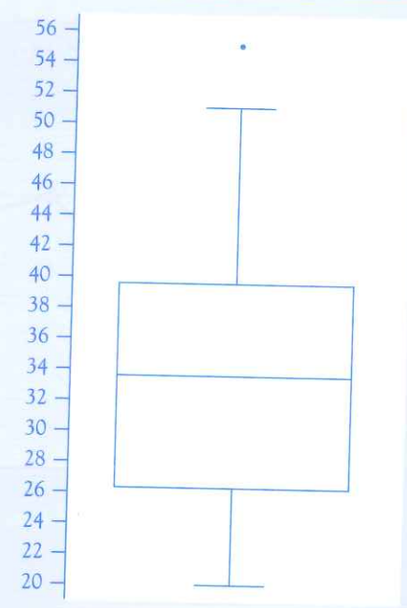




— Tamaño de la población: 20.
— Mediana: 33,5.
— Menor valor: 20.
— Mayor valor: 55.
— Primer cuartil: 26,25.
— Tercer cuartil: 39,5.
— Rango intercuartílico: 13,25.

El lado inferior del rectángulo representa el primer cuartil, y el lado superior, el tercer cuartil, el 50 % de los datos, tal como hemos explicado con anterioridad. En consecuencia, la altura de la caja representa el rango intercuartílico. La línea horizontal a través de la caja es la mediana.

**FIGURA 3.12**



Las líneas verticales que sobresalen de la caja, «bigotes», se extienden, respectivamente, hasta el mínimo y el máximo del conjunto de datos, siempre que estos valores no difieran de la media más de una vez y media el rango intercuartílico. Los extremos de los bigotes están marcados por dos líneas horizontales cortas. Los valores, indicados por puntos, respectivamente, por debajo y por encima de los bigotes inferior y superior se consideran valores atípicos.

### 3.2.7. Gráficos temporales (series temporales)

Representan una variable situando en el eje de las abscisas los períodos temporales en los que se midió la variable y en el eje de las ordenadas los valores que han tomado esas variables en ese recorrido temporal. Estos gráficos permiten observar la evolución

**FIGURA 3.13** TITULADOS EN TURISMO EN ASTURIAS. CURSOS 2000/01-2010/11


FUENTE: tabla de datos de elaboración propia.





general y las desviaciones significativas, es decir, obtendremos un patrón de frecuencias que denominaremos tendencia. Esta tendencia puede ser de tipo estacionario o constante; en este caso su gráfica será una línea recta paralela al eje de las abscisas, o puede ser variable creciente, decreciente lineal o curva. En esta representación, que en realidad es un polígono de frecuencias, debemos hacer notar que en el eje de las abscisas (X) se representan intervalos temporales.

# 4 Medidas de centralización de una variable

4.1. Estudio y análisis de la media de una variable. Media aritmética simple.
4.2. Media ponderada.
4.3. Media geométrica.
4.4. Media armónica.
4.5. La moda ($M_o$).
4.6. El uso de la mediana ($M_e$).
4.7. Medidas de posición no central. Los cuantiles.

A la hora de afrontar el estudio de una variable hemos de establecer cuáles son sus **medidas descriptivas.** Es decir, cuáles son los mínimos valores de la variable necesarios para su correcta interpretación, saber cuáles utilizar y por qué se trata de comenzar a resumir los datos.

Tras la elaboración de las tablas de frecuencias y ver los datos en un gráfico, el siguiente paso es el análisis de los datos recogidos en esas tablas, con una serie de valores descriptivos que resumen aspectos fundamentales de la muestra.

Estas medidas se pueden clasificar en tres (o cuatro) grandes grupos[1]. Así, hablaremos de medidas de posición (centralización y de dispersión), concentración y de forma. Cada una de ellas medirá alguna característica de la muestra, como pueden ser la simetría, la agrupación en torno a valores centrales, etc.

Podemos definir las **medidas de posición** como aquellas medidas que nos permiten describir cómo se distribuyen los datos muestrales, es decir, cómo se organizan los datos de un estudio. Para ello es necesario el cálculo de algunos estadísticos que nos den una información general e inicial sobre la muestra; los principales son las medidas de centralización (tendencia central) y dispersión.

Siguiendo el esquema propuesto, en este capítulo cuarto vamos a analizar las medidas que denominamos de centralización.

Las **medidas de centralización, promedios o reducción de datos** permiten describir cómo se organizan los datos de una muestra en torno a valores cercanos a un valor central y son, muy a menudo, representantes del conjunto de la muestra, aunque es probable que ninguno de estos coincida con los valores recogidos (valores muestrales). Estas medidas dan una idea rápida sobre la simetría de los datos respecto a un valor central y nos permiten tener una idea previa sobre la forma de la gráfica que define la distribución de los datos estudiados. Estas son la media (aritmética, truncada, recortada, cuadrática, geométrica y armónica), la mediana, la moda y los cuantiles. Estos últimos estadísticos permiten ver el comportamiento de la variable en partes o grupos con el mismo número de datos, para lo cual se divide el conjunto ordenado de datos en partes iguales o intervalos con el mismo número de datos, llamado cada uno cuantil. Los más utilizados son los cuartiles, los deciles y los percentiles, derivados de dividir el conjunto de datos en 4, 10 o 100 grupos.

[1] Esquema adaptado de J. Santos, A. Muñoz Alamillos y A. Muñoz Martínez (2007).





## 4.1. ESTUDIO Y ANÁLISIS DE LA MEDIA DE UNA VARIABLE. MEDIA ARITMÉTICA SIMPLE

Es una medida de centralización también llamada promedio; se puede definir como el cociente entre la suma de todos los valores que toma la variable en una distribución ($x_i$) y el número de observaciones ($N$). Se suele representar por $\bar{x}$.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{N} = \frac{\sum_{i=1}^{i=n} X_i}{N}$$

El símbolo $\Sigma$ significa sumatorio, y la notación $\sum_{i=1}^{i=n} X_i$ se lee como el sumatorio de todos los valores de $x$ desde el primero ($i = 1$) hasta el último ($i = n$).

En este caso estamos ante una variable de tipo I, por ello el cálculo es muy sencillo, pero normalmente trabajaremos con variables de tipo II y tipo III, en las cuales es necesario tener en cuenta la frecuencia de los valores de las variables. En ocasiones se denomina «media aritmética ponderada por las frecuencias» y se emplea la siguiente fórmula:

$$\bar{X} = \frac{X_1 \cdot n_1 + X_2 \cdot n_2 + X_3 \cdot n_3 + \cdots + X_n \cdot n_n}{N} = \frac{\sum_{i=1}^{i=n} X_i \cdot n_1}{N}$$

Hay que recordar que en las distribuciones de tipo III, los valores de ($x_i$) vienen dados por la marca del intervalo.

EJEMPLO. Queremos saber la percepción de la limpieza de 25 visitantes a nuestras instalaciones deportivas:

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

Cálculo de la media en distribuciones tipo I:

$$\bar{X} = \frac{4+3+3+5+2+3+4+2+3+5+6+7+8+3+6+7+4+6+4+3+8+7+7+6+9}{25} =$$

$$= \frac{125}{25} = 5$$

### 4.1.1. Cálculo de la media en distribuciones tipo II

| $X_i$ | $n_i$ | $X_i \cdot N_i$ |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 2 | 4 |
| 3 | 6 | 18 |
| 4 | 4 | 16 |
| 5 | 2 | 10 |
| 6 | 4 | 24 |
| 7 | 4 | 28 |
| 8 | 2 | 16 |
| 9 | 1 | 9 |
| **Suma** | **25** | **125** |





$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i \cdot n_i}{N} = \frac{125}{25} = 5$$

### 4.1.2. Cálculo de la media en distribuciones tipo III

| $L_{i-1} - L_i$ | $M_i = X_i$ | $N_i$ | $X_i \cdot N_i$ |
|---|---|---|---|
| 1-3 | 2 | 8 | 16 |
| 4-6 | 5 | 10 | 50 |
| 7-9 | 8 | 7 | 56 |
| Suma | | 25 | 122* |

* Aunque no es muy acertado calcular este tipo de distribuciones de tipo I como si fueran de tipo III, hemos querido hacerlo para que el lector vea la facilidad de su cálculo; vemos que no coincide exactamente, aunque la media es aproximadamente 4,9.

### 4.1.3. Características de la media

1. Este valor no tiene por qué corresponderse con ninguno de los valores de la distribución, pero da información de un valor centrado que representa a todos los de la distribución.
2. Es una medida única y tiene más valor si se acompaña de una medida de dispersión.
3. Puede verse condicionada por valores atípicos, de forma que valores altos o bajos pueden tener mucho peso en su valor.
4. En el caso de variables continuas agrupadas en intervalos, su valor depende del diseño de los intervalos.

### 4.1.4. Propiedades de la media

1. La suma de las desviaciones de todos los valores respecto a su media aritmética es 0.

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

2. Si multiplicamos o dividimos todas las observaciones por un mismo número, lo que se conoce como cambio de escala, la media queda multiplicada o dividida por ese número.
3. Si le sumamos a todas las observaciones un mismo número, lo que se conoce como cambio de origen, la media cambiará en dicha cantidad. Como consecuencia, si se aplican estas dos propiedades (cambio de origen $O_t$ y cambio de escala C), obtenemos un resultado que sirve para simplificar cálculos de valores muy elevados.

### 4.1.5. Ventajas e inconvenientes del uso de la media

**Principales ventajas**

1. Es un concepto conocido para la mayoría de personas y es claro e intuitivo.
2. Es calculable en todas las variables, siempre que nuestras observaciones sean cuantitativas.
3. Para el cálculo se utilizan todos los valores de la distribución.
4. Es única para cada distribución de frecuencias.
5. Al ser el *centro de gravedad* de la distribución, representa todos los valores observados.

Es útil para llevar a cabo procedimientos estadísticos, como la comparación de medias de varios conjuntos de datos.





**Principales inconvenientes**

1. Es un valor sensible a los valores extremos, con lo que si tenemos distribuciones con una gran dispersión de datos puede llegar a perder su significado. Un ejemplo es la famosa anécdota del pollo: dos personas se disponen a comer un pollo, si una persona come un pollo y otra no come pollo, como media, se habrán comido medio pollo cada uno.
2. Que no se puede calcular cuando los parámetros son cualitativos.
3. Podemos tener dificultades para su cálculo en distribuciones tipo III en las que existan intervalos abiertos; en estos casos es necesario estimar una marca de clase para calcular la media y esta varía en función de la marca de clase elegida.

## 4.2. MEDIA PONDERADA

La media ponderada es muy útil cuando en una colección de datos u observaciones cada uno tiene un diferente peso específico. Se calcula como el cociente entre la suma de los productos de cada valor observado multiplicado por su peso específico entre la suma de los diferentes pesos específicos.

**Media ponderada por coeficientes**

En ocasiones es necesario introducir un coeficiente de ponderación que dé mayor peso a algunos valores de la variable; estas ponderaciones suelen representarse por $(w_i)$, quedando la fórmula de la media de la siguiente manera:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i \cdot N_i \cdot W_i}{\sum_{i=1}^{n} W_i \cdot N_i}$$

Ejemplo. Queremos seleccionar para nuestro departamento de eventos un comercial internacional, para ello hemos realizado cuatro pruebas en nuestro proceso de selección a dos candidatos, cada una de las pruebas con un peso específico diferente. Los datos obtenidos son los siguientes:

| Pruebas | Resultados obtenidos | | Coeficiente de ponderación |
|---|---|---|---|
| | Candidato 1 | Candidato 1 | |
| Idioma | 5 | 5 | 2 |
| Comunicación | 10 | 5 | 1 |
| Experiencia | 5 | 8 | 3 |
| Formación | 8 | 8 | 1 |

La media aritmética de los candidatos es $C_1 = 7$; $C_2 = 6{,}5$. Si nos guiáramos por este tipo de media elegiríamos al candidato número 1, mientras que con la media ponderada:

$$C_1 = \frac{10 + 10 + 15 + 8}{4} = 10{,}75$$

$$C_2 = \frac{10 + 5 + 24 + 8}{4} = 11{,}75$$

Con los datos obtenidos ponderados deberíamos elegir al candidato número 2.

### 4.2.1. Otras medidas de la media aritmética

**Media truncada.** Para evitar la sensibilidad a las medidas extremas o atípicas se calcula la media con un determinado porcentaje central de datos. Así, una media truncada del 10% es aquella que elimina de su cálculo el 10% de los datos superiores e inferiores.





**Media recortada.** Consiste en la media aritmética de una modificación de los datos originales: un porcentaje central permanece sin modificar, cada uno de los datos de los menores excluidos se sustituye por el menor de los datos del porcentaje central no modificado y cada uno de los datos mayores excluidos se reemplaza por el mayor de los datos del porcentaje central no modificado (Cao y cols., 2001).

**Media cuadrática.** Notada como Q, es la raíz cuadrada positiva de la media aritmética de los cuadrados de las observaciones. Es interesante emplearla cuando hay valores con signo negativo, como pueden ser los errores de diferentes medidas.

## 4.3. MEDIA GEOMÉTRICA

Notada como G, se calcula como la raíz enésima del producto de los *n* valores observados. La media geométrica de una distribución de frecuencias con *N* observaciones es la raíz de índice *N* del producto de todas las observaciones elevado a sus respectivas frecuencias.

Suele usarse para promediar porcentajes promedios o índices y también para valores con tendencia a crecer exponencialmente.

Su formulación es la siguiente[2]:

— Distribuciones tipo I:

$$G = \sqrt[n]{(x_1 \cdot x_2 \cdot \dots \cdot x_n)} = \sqrt[n]{\prod_{i=1}^{i=r} x_i^{n_i}}$$

— Distribuciones tipo II y III:

$$G = \sqrt[n]{(x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_r^{n_r})} = \sqrt[n]{\prod_{i=1}^{i=r} x_i^{r}}$$

[2] La notación $\prod$ significa que se trata del producto de todos los valores de la variable.

- Este estadístico solo se puede utilizar si no hay observaciones nulas, ya que si no, tomaría el valor 0.
- Tampoco cuando exista algún valor negativo, ya que nos podría dar valores irracionales que invaliden el valor estadístico.

### 4.3.1. Ventajas e inconvenientes del uso de la media geométrica

**Principales ventajas**

1. En su cálculo intervienen todos los valores de la distribución.
2. Es menos sensible que la media aritmética cuando la distribución tiene valores extremos.
3. Es más representativa cuando la distribución evoluciona de forma acumulativa o multiplicativa.
4. Su valor es objetivo y único, cuando existe.

**Principales inconvenientes**

1. Su significado estadístico es menos intuitivo que la media aritmética.
2. Mayor complicación de los cálculos.
3. Su indefinición (da números imaginarios) cuando tiene valores negativos y su valor nulo cuando una observación toma ese valor.

Existe una advertencia: dada su formulación, deberemos utilizar logaritmos o programas informáticos para su cálculo. Además, la media geométrica tiene la siguiente propiedad:

Significa que el logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

$$\log G = \frac{1}{N}\sum_{i=1}^{n} n_i \log x_i$$





Es evidente que para calcular G, una vez utilizada esta fórmula, se debe hacer el antilogaritmo del resultado.

Pongamos un ejemplo sencillo del uso de esta media geométrica.

Ejemplo (distribuciones tipo I). Una agencia de viajes ha experimentado desde su apertura en 2007 un incremento del 10 % en 2008, 20 % en 2009 y en 2010 tiene una previsión de crecimiento del 25 %. ¿Cuál es el alcance del crecimiento en ese trienio?

| Año | Crecimiento |
|---|---|
| 2007 | 100 |
| 2008 | 110 (10 %) |
| 2009 | 132 (20 % sobre 110) |
| 2010 | 165 (25 % sobre 132) |

$$G = \sqrt[n]{(x_1 \cdot x_2 \cdot \cdots \cdot x_n)} = \sqrt[n]{\prod_{i=1}^{i=r} x_i^{n_i}} = \sqrt[3]{110 \cdot 132 \cdot 165} = 133{,}8$$

La media aritmética será:

$$\frac{110 + 132 + 165}{3} = 135{,}66$$

Si nos fijamos, la diferencia es significativa en cuanto que estamos trabajando con porcentajes y dependiendo del volumen un 1,86 % puede ser una gran diferencia. En este caso se cumple una relación: $G \leqslant \bar{X}$.

Ejemplo (distribuciones tipo II y III). En los años 2009, 2010 y 2011 varios establecimientos de una cadena hotelera han experimentado un crecimiento, obteniendo los siguientes datos:

| Crecimiento | Núm. de hoteles |
|---|---|
| 10% | 4 |
| 15% | 6 |
| 20% | 10 |

$$G = \sqrt[n]{(x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_r^{n_r})} = \sqrt[n]{\prod_{i=1}^{i=r} x_i^r} = \sqrt[20]{10^4 \cdot 15^6 \cdot 20^{10}} = 15{,}97$$

Cuando trabajamos con cifras que se hacen tan grandes, es preferible usar la fórmula G logarítmica.

$$\log G = \frac{1}{N} \sum_{i=1}^{r} n_i \log x_i = \frac{1}{20} \cdot (4 \cdot \log 10 + 6 \cdot \log 15 + 10 \cdot \log 20) =$$

$$= 1{,}2$$

Aplicamos el antilogaritmo:

$$G = \text{Antilogaritmo}^{3}\ 1{,}20 = 15{,}85$$

Si hacemos la media aritmética, el resultado sería:

$$\frac{(10 \times 4) + (15 \times 6) + (20 \times 10)}{20} = 16{,}5$$

que es una valor susceptiblemente mayor. Se mantiene la regla de $G \leq X$.

## 4.4. MEDIA ARMÓNICA

Notada como $H$, se calcula dividiendo el número de observaciones por la suma del inverso de cada valor observado. Tiene

[3] El antilogaritmo en las calculadoras aparece como: $(10^x)$.





como principal característica que se deja influir menos por valores extremos y, por tanto, goza de una mayor robustez.

— **En distribuciones tipo I:**

$$H = \frac{N}{\sum_{i=1}^{n} \frac{1}{n_i}}$$

— **En distribuciones tipo II:**

$$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_r}{x_r}} = \frac{N}{\sum_{i=1}^{n} \frac{n_1}{x_i}}$$

Su utilización es bastante poco frecuente y solo debe emplearse cuando la variable está medida en unidades relativas, por ejemplo km/h..., es decir, para promediar velocidades, tiempos, rendimientos...

### 4.4.1. Ventajas e inconvenientes del uso de la media armónica

**Principales ventajas**

1. Está definida de forma objetiva y es única.
2. Para su cálculo tiene en cuenta todos los valores de la distribución.
3. Es más representativa que otras medidas en los casos de obtener promedios de velocidad, rendimientos, productividades...
4. Los valores extremos tienen una menor influencia que en la media aritmética.

**Principales inconvenientes**

1. Matemáticamente solo se puede calcular si no hay observaciones iguales a 0, ya que en este caso nos aparecería un cociente indeterminado del tipo $n_i/0 = \infty$.
2. Cuando la variable toma algunos valores muy pequeños, puede carecer de significado; en estos casos sus inversos pueden aumentar casi hasta el infinito, eliminando el efecto del resto de los valores.

### 4.4.2. Relación entre las medias, armónica, geométrica y aritmética

La relación entre las medias, armónica, geométrica y aritmética, se cumple siempre que la media armónica, si existe, es menor que la media geométrica (si también existe) y que esta a su vez es menor que la media aritmética.

$$H \leq G \leq \bar{X}$$

## 4.5. LA MODA ($M_o$)

Es una medida de centralización que indica el dato que aparece con mayor frecuencia en una distribución; **suele designarse por $M_o$ y se define como el valor de la variable que presenta mayor frecuencia absoluta.** Se puede dar la situación de que haya más de un dato con la máxima frecuencia, por lo que la distribución será bimodal o trimodal, o multimodal.

Cuando hay más de una moda, se diferencia entre moda o modas absolutas y modas relativas. Es relativa cuando su frecuencia absoluta no es superada por la de los valores contiguos.

Su cálculo varía según el tipo de distribución en que nos encontremos:

— **Distribución tipo I:** no se puede hablar de moda, ya que las frecuencias son todas unitarias.





— **Distribución tipo II:** para obtener la moda en las distribuciones de tipo II basta con observar la columna $n_i$.

— **Distribución tipo III:** se pueden dar dos supuestos:

Que los intervalos sean de la misma amplitud. La moda absoluta se sitúa en el intervalo que presente mayor frecuencia absoluta y las relativas en aquellos que superen la frecuencia absoluta de los intervalos contiguos. Para calcular el valor exacto hay que aplicar esta expresión (recordando que $c_i$ es la amplitud del intervalo):

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

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

EJEMPLO. Queremos saber la moda de visitantes al Museo del Prado durante una exposición itinerante del siglo XIX. Los resultados en los 15 días que ha durado la exposición se han agrupado de 5 en 5.

| $L_{i-1} - L_i$ | $n_i$ |
|---|---|
| 0-5 | 20.000 |
| 5-10 | 35.000 |
| 10-15 | 5.000 |
| | 60.000 |

Aplicamos la fórmula y tenemos:

$$5 + \frac{5.000}{20.000 + 5.000} \cdot 5 = 6$$

Existe un pequeño problema en este tipo de distribuciones cuando los intervalos no tienen la misma amplitud.

Para solucionar este problema debemos obtener una ratio de densidad de frecuencias. Se trata de dividir la frecuencia por la amplitud del intervalo. Se suele datar de la siguiente forma:

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

La fórmula de la moda en este tipo de distribuciones se calcula:

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

EJEMPLO. Hemos recogido una muestra al azar de nuestro local para tratar de establecer una nueva tarifa de precios en función de la edad; queremos saber la moda y a partir de ella realizar nuestra promoción

| $L_{i-1} - L_i$ | $n_i$ | $c_i$ | $h = n_i/c_i$ |
|---|---|---|---|
| 0-25 | 125 | 25 | 5 |
| 25-30 | 50 | 5 | 10 |
| 30-35 | 35 | 5 | 7 |
| 40-45 | 25 | 5 | 5 |
| 45-55 | 80 | 10 | 8 |
| 55-60 | 40 | 5 | 8 |
| | 355 | | |

Aplicando la fórmula, obtenemos el siguiente resultado:

$$25 + \frac{7}{5+7} \cdot 5 = 27,91$$





Debemos, por tanto, comenzar nuestra promoción teniendo en cuenta que la moda es de 28 años.

**Propiedades de la moda**

- Es la medida representativa en distribuciones con datos que no se pueden ordenar.
- Se usa con variables cualitativas al no necesitar cálculos aritméticos.
- Es un valor muy sensible, al ser independiente de la mayor parte de los datos.
- No siempre se sitúa en el centro de la distribución.

### 4.5.1. Ventajas e inconvenientes del uso de la moda

**Principales ventajas**

1. Es un estadístico que puede obtenerse en todas las distribuciones, ya que siempre se puede hallar su valor, la categoría o valor que más se repite, tanto en variables cuantitativas como cualitativas.
2. Su cálculo es sencillo.
3. Tiene una fácil interpretación, ya que nos da el valor de la variable o la modalidad de los atributos que más se repiten.

**Principal inconveniente**

Su mayor inconveniente es que no intervienen todos los valores de la distribución, centrándonos en la mayor frecuencia absoluta de un valor de la variable o de la modalidad de los atributos.

## 4.6. EL USO DE LA MEDIANA ($M_e$)

Es una medida de centralización y se define como el valor que deja a la mitad de las observaciones por debajo de ella y a la mitad

por encima o, dicho de otra manera, es el valor que ocupa la posición intermedia cuando una distribución se ordena numéricamente creciente o decreciente. Si el número de datos de la distribución es impar, la mediana es el valor que ocupa la posición intermedia y si el número de datos es par se calcula haciendo la media aritmética de los dos valores centrales.

Es una medida robusta que no se deja influir por valores extremos en la distribución. Intuitivamente, si las medidas anteriores las vemos como puntos de los cuales equidistan el resto de valores, la mediana es una barrera que divide toda la nube de puntos en dos partes con el mismo peso.

El lugar que ocupa se determina dividiendo el número de valores entre dos ($n/2$).

El objeto de la mediana es superar dos aspectos de la media: la escasa representatividad de la media cuando hay mucha dispersión en la distribución de una variable y cuando nos enfrentamos a una distribución de una variable cualitativa en la que la media no suele tener mucho sentido.

El primer problema se ve superado fácilmente con la mediana, ya que realiza un equilibrio de las frecuencias de las variables y no de los valores, mientras que para superar el segundo problema usamos la moda.

Como hemos dicho con anterioridad, se trata de una barrera física, por tanto su medida se ve afectada por el número de frecuencias, es decir, su cálculo varía si se trata de distribuciones pares o impares.

### 4.6.1. Cálculo de la mediana en distribuciones tipo I

**Cuando los valores son impares,** existe un término central, que será el valor de la mediana; la mediana coincide con un valor de la variable.

$$M_e = X \cdot \frac{n+1}{2}$$





Ejemplo. Hemos medido el número de errores que cometen siete personas al realizar una tarea compleja y los resultados son: 5, 8, 9, 7, 6, 3, 2; como hemos dicho en la definición, los valores deben estar ordenados, por ello el primer paso es ordenar los valores: 2, 3, 5, 6, 7, 8, 9; luego buscamos el término central:

$$\frac{7+1}{2} = 4$$

Ya sabemos la posición, vamos a los valores y buscamos el que ocupa el lugar 4. En nuestro caso es: 2, 3, 5, **6**, 7, 8, 9; el valor 6 de la variable, que deja tanto a la izquierda como a la derecha el mismo número de valores (tres). La diferencia de interpretación con respecto a la media sería:

La media de errores cometidos es de 5,7; mientras que si usamos la mediana nos diría que la mitad de las personas cometen más de seis errores al realizar esta tarea.

#### Cálculo de la mediana cuando la distribución es par

En este caso no hay un término central, sino que este se encuentra entre dos valores de la variable; por ello su cálculo varía:

$$M_e = \frac{x_i + X_{i+1}}{2}$$

Ejemplo. Hemos pedido a diez clientes de un establecimiento que valoren el grado de satisfacción (0-50), obteniendo los siguientes resultados:

15, 22, 40, 35, 27, 42, 28, 45, 20, 30

Ordenamos:

15, 20, 22, 27, 28, 30, 35, 40, 42, 45

Aplicamos N/2, obtenemos dos términos centrales el 5.º y 6.º; que corresponden a dos valores 28 y 30, aplicando la fórmula:

$$M_e = \frac{28 + 30}{2} = 29$$

es decir, que la mitad de nuestros clientes puntúan más de 29. Qué ocurre cuando obtenemos un valor decimal (intermedio) y nuestra variable no admite este tipo de valores[4]. Imaginemos que queremos saber el número de cliente diarios de un establecimiento y obtenemos 28,5 como mediana. Para su interpretación tendremos que decir que hay dos medianas conjuntas: tomaremos los dos valores centrales y diremos que la mitad de los días hay X y la otra mitad $X_{i+1}$.

### 4.6.2. Cálculo de la mediana en distribuciones tipo II

Para su cálculo el primer paso que debemos realizar es ordenar los valores y trabajar con la frecuencia acumulada ($N_i$), obteniendo en concreto el valor N/2.

Tal como hemos hecho en el caso de distribuciones tipo I, distinguimos dos casos para calcular la mediana en distribuciones tipo II:

1. Cuando existe un valor de $N_i$ que sea igual a N/2; la mediana entonces será la media aritmética entre el valor de su $X_i$ y del siguiente $X_{i+1}$, siempre y cuando la variable admita valores intermedios; si no es así, convenimos en que la mediana serán los dos valores conjuntos.
2. Cuando no existe un $N_i$ que iguale a N/2, la mediana corresponde al primer valor que supere al de N/2.

[4] Recordar que el objeto de la estadística no es un cálculo matemático, sino la interpretación de unos datos para la toma de decisiones.





Ejemplo 1. Hemos preguntado a 200 turistas de una localidad el número de horas promedio que han utilizado en visitas culturales, con el siguiente resultado:

| $x_i$ | $n_i$ | $N_i$ |
|---|---|---|
| 4 | 2 | 2 |
| 5 | 4 | 6 |
| 6 | 8 | 14 |
| 8 | 12 | 26 |
| 9 | 10 | 36 |
| **12** | 64 | **100** |
| **16** | 31 | 131 |
| 20 | 37 | 168 |
| 25 | 15 | 183 |
| 30 | 17 | 200 |
| | 200 | |

El valor de N/2 = 100.

Que coincide con un valor de la frecuencia acumulada; por tanto, tomamos, para calcular la mediana, el valor $X_i = 12$ y el valor $X_{i+1} = 16$.

Hallamos su media:

$$M = \frac{12 + 16}{2} = 14$$

Podemos concluir que más del 50 % pasan un promedio de 14 horas o más realizando visitas culturales en esa localidad.

Ejemplo 2. Hemos medido el grado de satisfacción (de 0-10) en usuarios de un aeropuerto español relativos al tiempo de espera

en sus enlaces internacionales en un día al azar, obteniendo los siguientes resultados:

5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4

Recordar que hay que trabajar con las frecuencias acumuladas.

| $x_i$ | $n_i$ | $N_i$ |
|---|---|---|
| 2 | 2 | 2 |
| 3 | 2 | 4 |
| 4 | 5 | 9 |
| **5** | 6 | **15** |
| 6 | 2 | 17 |
| 8 | 3 | 20 |
| | 20 | |

$N/2 = 10$

El primer valor que supera 10 es $N_i = 15$, que corresponde al valor $X_i = 5$; por tanto, la mediana es 5.

Esto significa que más de la mitad de los usuarios ha valorado satisfactoriamente su tiempo de espera en el aeropuerto, por encima del 5.

### 4.6.3. Cálculo de la mediana en distribuciones tipo III

Actuamos como siempre, solo que esta vez trabajamos con intervalos. Hallamos el valor $N/2$, y una vez establecido buscamos en la columna de frecuencias acumuladas un valor que iguale o supere al obtenido. En el intervalo que eso ocurre (intervalo mediano), si nos encontramos con un intervalo del tipo $L_{i-1} - L_i$, la





$M_e = L_{i-1} - L_i + m$, ahora hay que calcular ese valor $m$. Para ello tenemos dos opciones:

1. Mediante un razonamiento:

   Aproximar la mediana: $M_e = L_{i-1} + m$; para calcular $m$ es necesario tener en cuenta el valor de la Fr acumulada hasta el intervalo anterior, y restarlo al $N/2 - N_{i-1}$, al intervalo mediano le corresponde una frecuencia distribuida por toda su amplitud, es por ello por lo que hay que corregir esa diferencia mediante: $c_i/n_i$; por tanto:

$$m = \left(\frac{N}{2} - N_{i-1}\right) \cdot \frac{c_i}{n_i}$$

2. Aplicando la fórmula que se deduce de lo anterior:

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

EJEMPLO. Queremos evaluar los ingresos medios de los clientes de una cadena hotelera durante un año. Hemos analizado 1.000 clientes. Obteniendo los siguientes resultados:

| $L_{i-1} - L_i$ | $n_i$ | $N_i$ |
|---|---|---|
| 0-20.000 | 25 | 50 |
| 20.001-21.000 | 50 | 75 |
| 21.001-22.000 | 40 | 125 |
| 22.001-23.000 | 270 | 395 |
| 23.001-24.000 | 475 | 870 |
| 24.001-25.000 | 35 | 905 |
| 25.001-26.000 | 15 | 920 |
| 26.001-27.000 | 10 | 930 |
| 27.001- | 70 | 1.000 |
| | 1.000 | |

$$N/2 = 500$$

que corresponde con el intervalo 23.001-24.000, ya que acumula más de 500 clientes, concretamente 870. Se ve que hay un exceso de 370 clientes sobre el valor $N/2$. Con lo que debemos ajustar nuestros datos.

Si aplicamos la fórmula, obtenemos:

$$M_e = 23.001 + \frac{500 - 395}{475} \cdot 1.000 \cong 23.222$$

La conclusión es que tenemos más del 50 % de clientes que tienen unos ingresos por encima de los 23.222 euros.

Presenta como propiedades:

1. Que no se ve afectada por la mayor o menor dispersión de la distribución estadística y porque su valor puede variar en función de que los datos estén agrupados en intervalos.
2. Tiene bastante interés emplear la media junto con la mediana, y cuanta mayor sea su diferencia, más heterogéneos serán los datos.

### 4.6.4. Ventajas e inconvenientes del uso de la mediana

**Principales ventajas**

1. Es la medida más representativa en el caso de variables cualitativas o atributos.
2. Su cálculo es sencillo.
3. Tiene una fácil interpretación.
4. No es sensible a los valores extremos de la distribución.

**Principal inconveniente**

En su determinación no se tienen en cuenta todos los valores de la variable, aunque en algunos momentos es conveniente esta





desventaja, ya que permite el cálculo cuando no se conocen los extremos, pero sí la frecuencia, como en nuestro ejemplo.

## 4.7. MEDIDAS DE POSICIÓN NO CENTRAL. LOS CUANTILES

Hemos visto cómo la mediana permite establecer un corte en la mitad de los datos obtenidos en una distribución 50 % posteriores y 50 % anteriores. Siguiendo esta idea, los cuantiles dividen la distribución de frecuencias en partes iguales, siendo las más utilizadas:

1. Cuartiles: se trata de cada uno de los tres valores que dividen al conjunto ordenado de datos (nube de datos) en cuatro partes iguales, es decir, $Q_1$, $Q_2$ y $Q_3$. Al dividir la distribución en dos partes iguales, obtenemos la mediana, que se corresponde con $Q_2$, quedando la mitad de la distribución a la derecha de dicho cuantil y la otra mitad a la izquierda. Si cada una de estas partes las dividimos en dos, los valores correspondientes serán respectivamente $Q_1$ y $Q_3$. Cada uno de los cuatro intervalos generados contiene el 25 % de datos de la distribución.
2. Quintiles: se trata de los cuatro valores que dividen la nube de datos en cinco partes. Corresponde en proporción al 20 %, 40 %, 60 %, 80 %.
3. Deciles: son los cuantiles de orden 1/10 y dividen al conjunto de los datos de una distribución en 10 partes de igual frecuencia. Corresponden al 10 %, 20 %, 30 %, ..., 90 %. También se pueden definir como cada uno de los 9 valores que dividen la distribución en 10 partes iguales.
4. Percentiles: son los 99 valores que dividen en 100 partes iguales la nube de datos. Corresponderían al 1 %, 2 %, ..., 99 %. También se puede definir como el menor valor $P\alpha$, que es mayor o igual que el $\alpha$ % de los valores muestrales.

De manera más rigurosa llamaremos cuantil $r$-ésimo de orden $q$, y representaremos por $Q_{r/q}$ a aquel valor de la variable tal que, una vez ordenados crecientemente los valores de la distribución, dejan a su izquierda (incluido él) los $r/q$ partes de la distribución, al menos, y a su derecha (incluido él) las $(q-r)/q$ partes restantes, al menos.

Dicho de otra manera, si disponemos de $N$ valores, el cuantil $r$-ésimo dejará a su izquierda, al menos, $(r/q)N$ y a su derecha, al menos, los restantes

$$N - \frac{r}{q}N$$

## 4.7.1. Cálculo de los cuantiles

Para calcular los cuantiles se sigue un proceso semejante a la mediana: primero debemos localizar los puntos, si $N$ es el número de datos de una distribución, los cuantiles se calculan mediante la expresión: $rN/q$, en la que $r$ indica el cuantil correspondiente y $q$ el número de intervalos.

Si $q = 4$, tenemos cuartiles tomando $r$ los valores 1, 2 y 3; si $q = 5$, tenemos quintiles $r$ tomando los valores 1, 2, 3 y 4; si $q = 10$ deciles, $r$ tomará los valores 1, 2, ..., 9; y si $q = 100$ centiles, siendo los valores de $r = 1, 2, 3, ..., 99$.

Para el cálculo de los centiles tenemos que valorar los mismos casos que con la mediana; como ya hemos dicho, trabajamos con frecuencias, por tanto todos estos valores en las distribuciones tipo I no tienen sentido, ya que la frecuencia es uno.

En **distribuciones tipo II (sin agrupar),** primero ordenamos los valores de $X_i$ y hallamos la frecuencia acumulada $N_i$, aplicamos la fórmula $rN/q$ y obtenemos los valores $Q_1, Q_2, ..., Q_r$. Si la cantidad obtenida coincide con el valor de una frecuencia absoluta acumulada, se toma como valor de «Q» la media entre el valor que acumula la frecuencia obtenida y el siguiente valor:

$$Q = \frac{X_i + X_{i+1}}{2}$$





sin embargo, si la cantidad obtenida no coincide con ningún valor de la frecuencia absoluta acumulada, tomamos el valor de la variable superior que contenga esa frecuencia:

$$N_{i-1} < (\%) \cdot n < N_i \approx Q = x_i$$

**En distribuciones tipo III** (agrupadas) primero hallamos el intervalo donde se encuentra el cuantil (sean cuartiles, quintiles, deciles o percentiles) de forma semejante al caso anterior, y una vez localizado el intervalo aplicamos la fórmula[5]:

$$Q_{r/q} = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i}$$

EJEMPLO. **Distribuciones no agrupadas.**

Hemos analizado los precios medios de habitaciones que ofrece una localidad, encontrando los siguientes resultados (véase tabla) para posteriormente presentar nuestros resultados en la reunión del ayuntamiento; queremos en nuestra presentación que se vean los datos acumulados en porcentajes fácilmente entendibles; para tal fin hemos decidido hallar: los cuartiles, el primer decil y el percentil 80.

**Primero calculamos los cuartiles: $Q_1$, $Q_2$ y $Q_3$**

| $x_i$ | $n_i$ | $N_i$ |
|---|---|---|
| 50 | 30 | 30 |
| 100 | 70 | 100 |
| 150 | 50 | 150 |
| 200 | 30 | 180 |
| 250 | 20 | 200 |
| | 200 | |

[5] Hay que señalar que el lector puede encontrar otra fórmula en diferentes libros, pero el resultado es el mismo, y hemos querido usar una fórmula que se ajuste al contenido desarrollado en todos los temas anteriores.

$Q_1$: primero calculamos el lugar:

$$\frac{rN}{q} = \frac{1}{4}200 = 50$$

Como no coincide con ningún valor de frecuencias, tomamos el superior:

$$N_{i-1} < (\%) \cdot n < N_i; \quad 30 < 50 < 100; \quad Q_1 = X_i = 100$$

$Q_2$:

$$\frac{2}{4}200 = 100$$

que coincide con un valor de la frecuencia absoluta acumulada, por lo que:

$$Q_2 = \frac{X_i + X_{i+1}}{2} = \frac{100 + 150}{2} = 125$$

$Q_3$:

$$\frac{3}{4}200 = 150; \quad Q_3 = \frac{X_i + X_{i+1}}{2} = \frac{150 + 200}{2} = 175$$

Podemos sacar varias conclusiones: que el 50 % de los precios es inferior a 125 € o bien que el 75 % no supera los 175 €.

**Primer decil: $Q_1$**

$$\frac{1}{10}200 = 20; \quad 20 < 30$$





en este caso no hay inferior, por lo que tomamos la frecuencia que supera 30 y que corresponde a $X_i = 50$; este valor por sí solo no nos aporta muchos datos, pero si tenemos todos los deciles podríamos interpretar la tabla de precios de una forma sencilla.

**Octogésimo percentil: $Q_{80}$**

$Q_{80}$:

$$\frac{80}{100}200 = 160$$

no corresponde a ningún valor de frecuencias acumuladas. Tomamos entonces 180, que supera el valor obtenido, por lo que $Q_{80} = X_i = 200$. Concluimos que el 80 % de los precios de las habitaciones es menor o igual a 200 €.

EJEMPLO. **Distribuciones agrupadas.**

El mismo tipo de estudio ha sido realizado en otra localidad, pero en este caso han agrupado los precios en intervalos. Hallar el primer cuartil, quinto decil y percentil 90. Obteniendo los siguientes datos:

| $L_{i-1} - L_i$ | $n_i$ | $N_i$ |
|---|---|---|
| [0-100) | 120 | 120 |
| [100-200) | 190 | 310 |
| [200-300) | 190 | 500 |
| | 500 | |

**Cuartil $Q_1$:**

$$\frac{1}{4}500 = 125$$

por lo que estará en el intervalo [100-200). Aplicamos la fórmula:

$$\frac{Q_r}{q} = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i}$$

$$Q_1 = 100 + \frac{125 - 120}{190} \cdot 100 = 102{,}63 \approx 103$$

**Decil $Q_5$:**

$$\frac{5}{10} 500 = 250;\ \text{intervalo [100-200)}$$

aplicando la fórmula obtenemos:

$$100 + \frac{250 - 120}{190} \cdot 100 = 168{,}4 \approx 168$$

**Percentil $Q_{90}$:**

$$\frac{90}{100} 500 = 450;\ \text{intervalo [200-300);}$$

$$200 + \frac{450 - 310}{190} \cdot 100 = 273{,}68 \approx 274$$



# 5 Medidas de dispersión, concentración y forma

Existen situaciones en las que debemos tomar decisiones en función de los datos que hemos tomado y las medidas que hemos estudiado hasta el momento, aunque importantes, en ocasiones son insuficientes para tener una idea de la realidad del tema y pueden hacernos cometer errores.

En estos casos debemos estudiar otro grupo de medidas estadísticas que deben completar la información aportada por las medidas de posición; nos referimos a las medidas de dispersión, estas nos facilitan la cercanía o lejanía que existe entre nuestros datos. Debemos tener en cuenta que a mayor dispersión de los datos, las medidas de posición son menos eficaces para la toma de decisiones, ya que pierden representatividad.

Un ejemplo sencillo puede ser la distribución de notas de conocimiento de varios temas en una asignatura; tenemos 9, 2, 7, la media que se obtiene es un 6, sin embargo sabemos que esta persona en uno de los temas tiene un gran vacío de conocimientos, no sería lo mismo si hubiera obtenido 6, 5 y 7, la media es la misma, 6, pero en este caso tendría un nivel de conocimiento óptimo en todos los temas.

## 5.1. MEDIDAS DE DISPERSIÓN

A modo de definición, las **medidas de dispersión** o **variabilidad** intentan describir la dispersión o concentración del conjunto de datos con respecto a un valor central. La dispersión sería equivalente a la desviación o grado de variabilidad de un conjunto de observaciones. Asimismo, se pueden definir como el grado de separación de los valores de la muestra en función de las medidas de centralización utilizadas. Las medidas de dispersión completan la información y permiten evitar importantes errores.

Las más utilizadas son el rango o recorrido, el rango o recorrido intercuartílico, la desviación media, la varianza, la desviación típica o estándar, denominadas medidas de dispersión absolutas. El coeficiente de apertura, recorrido relativo y coeficiente de variación serán medidas de dispersión relativas[1].

[1] Siguiendo el esquema general que aparece en el capítulo 2, todas ellas pueden clasificarse en dos grupos: medidas de dispersión absolutas y relativas. La diferencia entre ambos grupos estriba en que los primeros surgen directamente de la distribución y por tanto están expresadas en las unidades que hemos medido las variables, mientras que las segundas suelen ser cálculos de segundo orden o relativas *(con este término nos referimos a que suelen ser calculadas en función de otras medidas estadísticas, en este caso de las medidas de dispersión absolutas que dependen directamente de los datos de la variable)* y, por tanto, tienen la ventaja de expresarse en medidas independientes a las utilizadas en los datos de la variable, con lo que su interpretación es más sencilla.





### 5.1.1. Recorrido o rango (amplitud total)

Cuando en una distribución ordenamos los valores de la variable de menor a mayor, la diferencia entre el mayor y el menor valor de la variable será el rango de esa variable. Se denota con $R$ y se formula:

$$R_x = x_r - x_1;\ \text{máx}\ \{x_i\} - \text{mín}\ \{x_i\} \quad \text{para } 1 \leqslant i \leqslant r$$

o lo que es lo mismo:

$$R_x = x_n - x_1$$

EJEMPLO. Queremos ofrecer un incentivo a los trabajadores de una empresa; esta empresa está dividida en dos sectores, hemos tomado sus edades para poder determinar si pueden recibir el mismo tipo de incentivo o no.

Sector X = 24, 26, 28, 30, 32; suman = 140, la media es 28

Sector Y = 22, 24, 20, 29, 45; suman = 140, la media es 28

Si la persona responsable de dar el incentivo tiene en cuenta solo el valor medio es muy posible que no consiga obtener el beneficio deseado, ya que los dos grupos son diferentes entre sí. $R_x = 32 - 24 = 8$; $R_y = 45 - 22 = 23$, de lo que podemos deducir que la distribución del sector X es menor que la del sector Y.

### 5.1.2. Coeficiente de apertura

Se define como el cociente/relación entre valor mayor y menor de una distribución.

$$C_{ap} = \frac{X_n}{X_i}$$

En nuestro ejemplo:

$$C_{apx} = \frac{32}{24} = 1{,}33$$

$$C_{apy} = \frac{45}{22} = 2{,}04$$

podemos decir que [X] tiene menor apertura que [Y].

Estas dos medidas de dispersión tienen el inconveniente de que dependen de los valores extremos y a veces las distribuciones pueden tener unos valores extremos alejados entre sí, pero el resto de la distribución sigue un esquema más concentrado (o no), esto puede ser problemático a la hora de establecer conclusiones; para paliar este tipo de influencia de los valores extremos podemos emplear las siguientes cuatro medidas.

## 5.1.3. Recorrido intercuartílico

Es la diferencia entre el tercer y primer cuartil de una distribución. Al utilizar los cuantiles como medida para su cálculo evitamos el peso que tienen los valores extremos. $R_i = Q_3 - Q_1$.

Siguiendo con el ejemplo anterior, tenemos[2]:

$X = \frac{rN}{q}$ $Q_1$ está en el valor $1 \cdot \frac{5}{4} = 1{,}25$, que corresponde al valor 2 $x_i = 26$

$Q_3$ está en el valor $3 \cdot \frac{5}{4} = 3{,}75$, que corresponde al valor 4 $x_i = 30$

Por tanto, $R_{ix} = 30 - 26 = 4$.

$$X = \frac{rN}{q} \quad Q_1 = 22;\ Q_3 = 29$$

[2] Aplicamos la fórmula del cálculo de los cuantiles, capítulo 4.





Por tanto, $R_{iy} = 29 - 22 = 7$, siendo el recorrido mayor que X.

## 5.1.4. Rango entre percentiles

El concepto es el mismo, ya que se trata de una relación entre cuantiles, solo que en este caso se usa el percentil. Se halla mediante la diferencia ente el percentil 90 y el percentil 10. $R_i = P_{90} - P_{10}$.

## 5.1.5. Recorrido relativo

Para hallar esta medida no utilizamos los datos directamente, sino que se usan estadísticos ya calculados, es por ello por lo que se encuadra dentro de las denominadas medidas de dispersión relativas. Se trata del cociente entre el *recorrido* y la *media aritmética*, y representa el número de veces que el recorrido de la variable contiene a la media.

$$RR_x = \frac{R}{\bar{X}}$$

Siguiendo nuestro ejemplo:

$$RR_x = \frac{8}{28} = 0{,}28; \quad RR_y = \frac{23}{28} = 0{,}82$$

## 5.1.6. Recorrido semintercuartílico

Estamos ante otra medida considerada de dispersión relativa, en este caso para su cálculo necesitamos haber calculado previamente los cuartiles $Q_1$ y $Q_3$. Se halla dividiendo el recorrido intercuartílico y la suma del primer y tercer cuartil.

$$R_{si} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

## 5.1.7. Desviación media

La desviación media es una medida que pretende establecer la media aritmética de todos los valores de la variable con respecto a una medida de posición central, que puede ser la media o la mediana.

A la hora de poner en práctica esta medida debemos tener en cuenta una de las propiedades de la media: cada valor de la variable tiene una desviación[3] con respecto a la media ($d_x$); la suma de todas estas desviaciones multiplicada por todas sus frecuencias es igual a 0. Para salvar esta igualdad podemos hacer varias operaciones, la primera es introducir los valores absolutos, llamada **desviación media** o **desviación absoluta media:**

$$D\bar{x} = \frac{n\sum_{i=1}^{n}|x_i - \bar{x}|}{N}$$

Otra forma de obtener este valor es utilizando, en lugar de la media como valor central, la mediana; el planteamiento y la idea es la misma, esta puede ser una buena medida en aquellas variables en las que el uso de la media no es el más adecuado o su cálculo no es posible. **Desviación mediana:**

$$D_{Me} = \frac{\sum_{i=1}^{n}|x_i - M_e|\, n_i}{N}$$

## 5.1.8. La varianza

Es una medida muy importante, ya que de ella surgen una serie de modelos estadísticos: los denominados análisis de varianza (utilizados sobre todo en la estadística inferencial). Se trata de una medida que permite analizar la información de la variable desde un punto de vista diferente y poder comparar los datos. Es la media aritmética de los cuadrados de las desviaciones de los datos respecto a la media aritmética; por tanto, este estadístico nos da el cuadrado de la unidad de medida de la variable: si la variable se mide por ejemplo en metros la varianza será en metros cuadrados; esto dificulta su interpretación, aunque permite su cálculo, ya que si no la media del valor de cada uno de los datos con respecto a la media aritmética de la distribución sería 0. Es por ello que aun siendo una medida muy importante a la hora de interpretar los datos, se hace la raíz cuadrada de este estadístico, dando como resultado la desviación típica.

Tiene dos notaciones $S^2$ o $\sigma^2$, normalmente se usan de manera indistinta; no obstante, se mantiene que la nomenclatura latina es para las distribuciones muestrales, mientras que la nomenclatura griega es para las distribuciones poblacionales.

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

Si desarrollamos el cuadrado de la fórmula podemos obtener otra formulación; la ventaja frente a la anterior es que permite utilizar los valores de la variable en lugar de las diferencias entre los valores y la media, que en ocasiones pueden dificultar su cálculo.

$$S^2 = \sigma_x^2 = \frac{\sum x_1^2 \cdot n_1}{N} - \bar{x}^2$$

La varianza siempre es un valor no negativo, que puede ser igual o mayor que 0.

## 5.1.9. La desviación típica

**Desviación típica** o **estándar:** si a la varianza le extraemos la raíz cuadrada positiva obtendremos la desviación estándar, notada

[3] Llamamos desviación a la diferencia entre el valor de la variable y la media aritmética.

como $\sigma$ o S y que se expresa en las mismas unidades de medida que los datos originales. Se puede interpretar como la variabilidad de unos datos y al comparar la desviación típica con los datos, cuanto más pequeña sea esta, mayor será la concentración de los datos alrededor de la media.

Habitualmente una distribución estadística se suele indicar con su media y la desviación típica. Esto es así, ya que al hacer la raíz cuadrada de la varianza las unidades de medida de la media y la desviación típica son las mismas.

Su cálculo se puede obtener de tres formas:

$$S_x = \sigma_x = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x}) \frac{n_i}{N}}$$

$$S_x = \sigma_x = \sqrt{\frac{\sum x_i^2 \cdot n_1}{N} - \overline{x}^2}$$

$$S_x = \sigma_x = \sqrt{S^2}$$

La desviación típica siempre es mayor o igual a 0; se trata de la medida de dispersión más pequeña, y además de estas propiedades tiene la propiedad de que si se le suma una constante a todos los valores de la variable, esta no varía. Por ende, si se multiplican todos los valores de la variable por una constante, la desviación queda multiplicada por el valor absoluto de dicha constante.

## 5.1.10. El coeficiente de variación de Pearson

El objetivo de la estadística no es el cálculo de estadísticos, sino la organización de los datos para transformarlos en información y posteriormente en conocimiento, siendo los estadísticos las herramientas para tal fin.





Las medidas descritas en este capítulo, de dispersión, son indicadores que pueden causar problemas a la hora de comparar variables y dar lugar a conclusiones erróneas, ya que pueden operar con diferentes unidades de medida; es por ello por lo que a la hora de tomar decisiones basándose en nuestros datos sea necesario recurrir a medidas de dispersión relativas.

El coeficiente de variación de Pearson es una de las más significativas: representa el número de veces que la desviación típica contiene a la media, y **nos permite afirmar el grado de representatividad de la media en una distribución.** Su cálculo es muy sencillo, ya que se trata del cociente entre la desviación típica y la media. La nomenclatura que se suele utilizar es V o $C_v$ y puede expresarse en tantos por ciento.

$$V = Cv = \frac{S}{\overline{X}} = \frac{\sigma}{\overline{x}}$$

En ocasiones se utiliza en tantos por ciento $\left(\frac{\sigma}{\overline{x}} \cdot 100\right)$.

Su interpretación:

1. Si el $C_{vx} < C_{vy}$, significa que la media de $\overline{X}$ representa mejor la distribución que la media de $\overline{Y}$.
2. Si el $C_v = 0$, la representatividad de la media es máxima.
3. Por consenso se dice que un $C_v$ tiene una representación óptima cuando es igual o menor que 0,3.

## 5.1.11. Tipificación de una variable y el teorema de Tchebychev

El objetivo de la tipificación es transformar cualquier distribución normal en una distribución normal (0, 1). Para lograrlo debemos eliminar la influencia de la media y la desviación típica convirtiendo nuestra distribución en una cuyos valores son $\overline{X} = 0$,

y $S = 1$; con ello se consigue que podamos comparar distribuciones estadísticas.

A la hora de comparar distribuciones con variables tipificadas lo que determinamos no es un valor concreto, sino la probabilidad de una variable en un intervalo de la misma.

La transformación tiene un cálculo muy sencillo:

$$Z = \frac{X - \bar{X}}{S} = \frac{X - \mu}{\sigma}$$

los valores que toma la distribución de la $N(0,1)$ se representan habitualmente tabulados para diferentes valores de Z, de esta forma se evita el tener que calcular el valor de la integral que lo define.

Ejemplo. Consideramos que el tiempo medio dedicado a la atención al cliente en un mostrador de un aeropuerto es de 3 minutos y la desviación típica es de 5 minutos; queremos saber la probabilidad, es decir, la proporción de clientes que son atendidos en menos de 9 minutos; nuestra distribución tiene una forma $N(3,5)$.

**Paso 1. Tipificación**

$$P(X < 9) = P\left(\frac{x-3}{5} < \frac{9-3}{5}\right) = P\left(Z < \frac{6}{5}\right) = P(Z < 1{,}2)$$

**Paso 2**

$$\frac{x-3}{5}$$

es una variable tipificada del tipo $N(0,1)$, cuyos valores de probabilidad los podemos buscar en la tabla de distribución normal





$N(0,1)$[4]. Miramos el valor en la tabla, fila 1,2 columna 0,0. $Z = 0{,}8849$.

$$P(X < 9) = P(Z < 1{,}2) = 88{,}49\,\%$$

*A la hora de calcular Z, las reglas básicas serán:*

**Caso 1.** La probabilidad Z ⩽ que el valor:

$P(Z \leqslant x_1)$ la solución $P(Z \leqslant X_1) = F(x_1)$
$P(Z \leqslant 1{,}2) = 0{,}8849$ (caso del ejemplo)

**Caso 2.** La probabilidad Z ⩾ que el valor:

$P(Z \geqslant x_1)$ la solución $P(Z \geqslant x_1) = 1 - F(x_1)$
$P(Z \geqslant 1{,}2) = 1 - F(1{,}2) = 1 - 0{,}8849 = 0{,}1151$

**Caso 3.** Cuando el valor calculado es negativo:

$P(Z \leqslant -x_1)$ la solución $P(Z \leqslant -x_1) = F(-x_1) = 1 - F(x_1)$
$P(Z \leqslant -1{,}7) = 1 - F(1{,}7) = 1 - 0{,}9554 = 0{,}0446$

**Caso 4.** Z está comprendida entre dos valores positivos:

$P(x_1 \leqslant Z \leqslant x_2)$ la solución $P(x_1 \leqslant Z \leqslant x_2) = F(x_2) - F(x_1)$
$P(1{,}2 \leqslant Z \leqslant 2{,}1) = F(2{,}1) - F(1{,}2) = 0{,}9821 - 0{,}8849 = 0{,}0972$

**Caso 5.** Z está comprendida entre dos valores negativos:

$P(-x_1 \leqslant Z \leqslant -x_2)$ la solución
$P(-x_1 \leqslant Z \leqslant -x_2) = P(Z \leqslant -x_2) - P(Z \leqslant -x_1) =$
$= 1 - P(Z \leqslant x_2) - 1 - P(Z \leqslant x_1) = P(Z \leqslant x_1) - P(Z \leqslant x_2)$
$P(-2{,}1 \leqslant Z \leqslant -1{,}2) = F(2{,}1) - F(1{,}2) = 0{,}9821 - 0{,}8849 = 0{,}0972$

[4] http://www.vaxasoftware.com/doc_edu/mat/dnormal.pdf. VAXA Software, recursos gratuitos de matemáticas.

**Caso 6.** $Z$ está entre dos valores de diferente signo:

$P(-x_1 \leq Z \leq x_2)$ la solución $P(-x_1 \leq Z \leq x_2) = F(x_2) - F(-x_1)$

$$P(-2 \leq Z \leq 3) = F(3) - 1 + F(2) = F(3) + F(2) - 1 =$$
$$= (0{,}9987 + 0{,}9772) - 1 = 0{,}9759$$

Como la distribución normal es simétrica, los valores $F(x)$ correspondientes a valores $x$ negativos se obtienen por simetría, si $F(1{,}2) = 0{,}8849$; $F(-1{,}2) = 0{,}1151$; $F(x) + F(-x) = 1$.

### Teorema de Tchebychev

Este teorema establece que la fracción de área entre los valores simétricos en torno a la media está relacionada con la desviación típica; se trata de una proporción basada en que cualquier distribución que esté a menos de $k$ desviaciones de la media:

$$P(K\sigma - \mu \leq K\sigma + \mu) \geq 1 - \frac{1}{K^2}; \quad \text{siempre que } k > 1$$

Es decir, que si sabemos la media y la desviación típica de cualquier distribución podemos calcular (con cierto margen de error) la cantidad de observaciones/valores que hay en una porción de la variable. El teorema establece que a dos desviaciones típicas de la media siempre se encontrará al menos el 75 % de los datos. Este teorema puede aplicarse a cualquier tipo de distribución, no solo las distribuciones normales.

Entre:

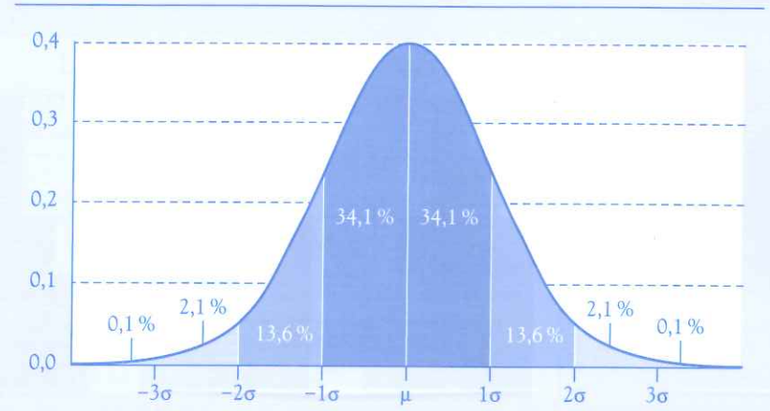
1. $(1\mu - \sigma \leq 1\mu + \sigma)$ está el 68 % de la distribución.
2. $(2\mu - \sigma \leq 2\mu + \sigma)$ está el 75 % de la distribución.
3. $(3\mu - \sigma \leq 3\mu + \sigma)$ está el 89 % de la distribución.

Esta puede mejorar al 95 % o al 99 % en el caso de que la distribución sea normal (figura 5.1).





**FIGURA 5.1**




FUENTE: Tomada de: https://habrastorage.org/getpro/geektimes/post_images/b12/21f/885/b1221f8851808d590d44c18c2752ad49.png.

## 5.2. MEDIDAS DE FORMA: ASIMETRÍA Y CURTOSIS

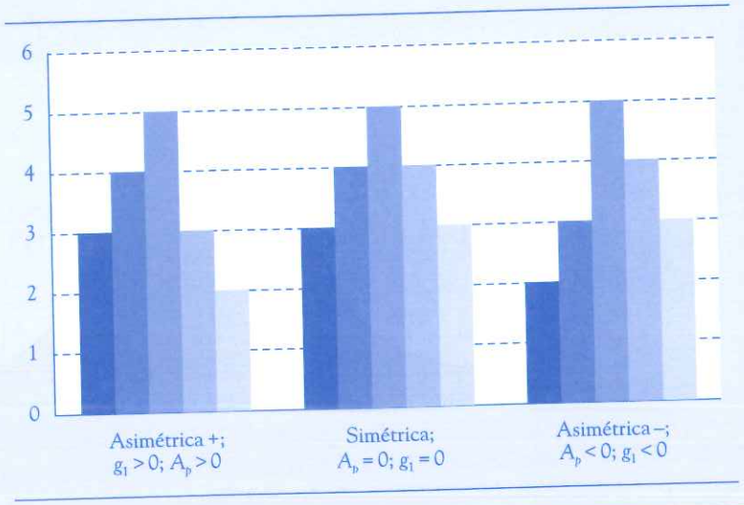
Este apartado nos ayuda a conocer mejor cómo es la distribución de los datos de una variable, en este caso concreto qué información nos aporta la forma que tiene su representación gráfica. Existen dos medidas para tal fin: las medidas de asimetría y curtosis (apuntalamiento).

### 5.2.1. Asimetría

Se dice que una distribución estadística es simétrica cuando su representación gráfica, diagramas de barras, histograma o polígono de frecuencias lo es; dicho de otra forma, si respecto a la media o a la mediana tenemos la misma cantidad de frecuencias. Afirmamos que es asimétrica cuando esta representación está desplazada en mayor o menor forma a la izquierda o a la derecha.

Si el desplazamiento es hacia la derecha, diremos que tiene asimetría positiva, mientras que si lo hace a la izquierda tendrá asi-

**FIGURA 5.2**




metría negativa. Debemos tener en cuenta que si hay asimetría por la derecha, existe una mayor concentración de datos en la izquierda, mientras que si la asimetría es por la izquierda, la mayor concentración de datos será en la derecha.

En el análisis de una variable la realidad nos dice que es improbable que encontremos distribuciones simétricas perfectas, es por ello por lo que se afirma que una distribución es simétrica cuando su coeficiente de simetría ($As$) se aproxima al valor 0.

Para calcular el coeficiente de asimetría de una variable existen varios procedimientos; no obstante, dos son los más utilizados:

1. Coeficiente de asimetría de Pearson ($Ap$).
2. Coeficiente de asimetría de Fisher ($g_1$).

El coeficiente de asimetría de Pearson tiene dos formas: una a partir de la moda y que es aplicable en aquellas distribuciones con una moda y con una distribución en forma de campana, y la otra derivada de esta en función de la media y mediana.





$$As = Ap = \frac{\bar{X} - M_o}{S}$$

La relación entre la media mediana y moda de una distribución nos permite obtener una forma aproximada del coeficiente de asimetría de Pearson:

$$\bar{X} - M_o = 3(\bar{X} - M_e); \qquad As = Ap = \frac{3(\bar{X} - M_e)}{S}$$

La interpretación de los resultados:

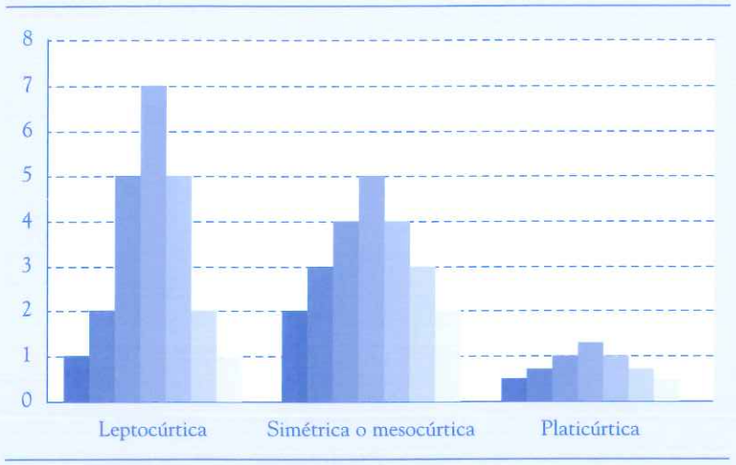
- Si $Ap > 0$ o bien hemos utilizado la fórmula de Fisher $g_1 > 0$, la distribución será asimétrica positiva (hacia la derecha).
- Si $Ap = 0$ o $g_1 = 0$, la distribución es simétrica.
- Si $Ap < 0$ o $g_1 < 0$, la distribución será asimétrica negativa (hacia la izquierda).

## 5.2.2. Curtosis o apuntalamiento

Las medidas de apuntalamiento o curtosis tratan de establecer la mayor o menor cantidad de datos en torno a la media (en algunos libros se dice en torno a la moda).

Si tomamos como referencia una distribución normal o tipo (campana de Gauss) en la cual la mayor concentración de datos se encuentra en torno a la media, siendo pocos los que corresponden a los valores extremos, una distribución muy apuntada será aquella que tiene un mayor índice de frecuencias muy cerca del punto medio, denominándose leptocúrtica; mientras que si las frecuencias en torno al punto medio no se diferencian mucho del resto de valores, nos encontraremos ante una distribución platicúrtica.

**FIGURA 5.3**




Al igual que en la anterior medida de forma, es improbable que encontremos distribuciones totalmente simétricas; es por ello por lo que nuestra interpretación se basa en los resultados obtenidos mediante su cálculo matemático; una de ellas es la fórmula de la curtosis de Fisher $g_2$ (es la más utilizada, aunque podemos encontrar más: como el coeficiente de apuntamiento percentílico 10-90).

Decimos que una distribución es simétrica cuando sus valores son cercanos a 0; leptocúrtica cuando $g_2 > 0$, mientras que es considerada palticúrtica cuando $g_2 < 0$.

$$g_2 = \frac{\frac{1}{n}\sum (x_i - x)^4 - n^i}{nS^4} - 3$$





## 5.3. MEDIDAS DE CONCENTRACIÓN: ÍNDICE DE GINI Y CURVA DE LORENZ

Las medidas de concentración muestran cómo se reparten los valores de una variable, si esta tiene un reparto equitativo o desigual; por tanto, nos dice el grado de distribución de la variable.

El concepto de concentración no debe engañarnos, no es algo contrario a las medidas de dispersión, ya que estas medidas lo que nos dicen es dónde se acumula una gran cantidad en determinados individuos. Hemos de añadir que suelen ser utilizadas para analizar los aspectos retributivos como los salarios, rentas, riqueza de una zona, etc.

Los valores de la concentración varían desde la concentración máxima, que es cuando uno recibe el total y los demás nada. Sería un reparto no equitativo. Y la concentración mínima, que es cuando todos los valores de la variable están repartidos por igual. Reparto equitativo. Las medidas más importantes de concentración son el índice de Gini y la curva de Lorenz. La primera es un coeficiente de concentración, mientras que la segunda es una curva, un gráfico. Hay que recordar que se trata de valores asociados a datos económicos, de riqueza.

### 5.3.1. Índice de Gini

El cálculo del coeficiente de concentración Gini se representa por $IG$ o $I_{co}$. Y su fórmula es:

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

Para entender cómo se calcula este índice debemos saber paso a paso el significado de sus valores y su construcción.

Partimos como siempre de saber qué queremos determinar, analizar, es decir, definir la variable a medir. Pensemos en estudiar el grado de concentración de las rentas de una empresa, tomaremos como siempre por un lado el valor de la variable ($x_i$) y la frecuencia de cada valor de la variable ($n_i$).

La riqueza del grupo, es decir, la riqueza total de la empresa quedará medida por el producto de ($x_i \times n_i$). Al tratarse de una medida de concentración, debemos tener en cuenta sus frecuencias acumuladas. Nuestro objetivo es buscar un índice que nos permita comparar cualquier tipo de distribución, por ello hay que anular el efecto de la medida/escala de los datos, por ello utilizamos en el cálculo la frecuencia relativa tanto para ($q_i$) como para ($p_i$), quedando su formulación así:

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

siendo $N_i$ la frecuencia acumulada y $N$ el total de la frecuencia, con lo que tenemos la frecuencia acumulada relativa.

Lo mismo haremos para el cálculo de la riqueza total de todo el personal de la empresa $q_i$: tomaremos los ($x_j \times n_j$) acumulados y los dividiremos por el sumatorio de los valores no acumulados ($x_i \times n_i$). Su formulación queda expresada así:

$$q_i = \frac{x_j \cdot n_j}{\sum x_i \cdot n_i}$$

Ambos valores pueden darse en tantos por ciento.

Como puede observarse en la fórmula del índice de Gini, para su cálculo se usan los $n - 1$, ya que se alcanza el 100 % de los trabajadores y el 100 % de la riqueza.

Nuestro ejemplo puede plasmarse de la siguiente forma. La empresa «XXXXXXXX» desea estudiar la concentración de su masa salarial, cuya distribución de salarios es la siguiente:





| $L_{i-1} - L_i$ | $n_i$ | $N_i$ | $p = N_i/N$ | $x_i$; $m$ | $x_i \times n_i$ | Acumulado $x_i \times n_i = x_j \times n_j$ | $q_i$ | $p_i - q_i$ |
|---|---|---|---|---|---|---|---|---|
| 600-1.000 | 200 | 200 | 0,40 | 800 | 160.000 | 160.000 | 0,2133 | 0,1867 |
| 1.001-1.500 | 50 | 250 | 0,50 | 1.250 | 62.500 | 222.500 | 0,2966 | 0,2034 |
| 1.501-2.000 | 70 | 320 | 0,64 | 1.750 | 122.500 | 345.000 | 0,4600 | 0,1800 |
| 2.001-2.500 | 180 | 500 | ~~1,00~~ | 2.250 | 405.000 | 750.000 | 1 | 0 |
| Sumatorios | $N = 500$ | | | | $\Sigma(x_i \cdot n_i) =$ $= 750.000$ | | | $\sum_{i=1}^{n-1}(p_i - q_i) =$ $= 0,5701$ |
| $\sum_{i=1}^{n-1} p_i$ | | 1,54 | | | | | | |

El valor de *IG* es:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = 0,3701$$

Si tuviéramos que comparar esta empresa con otras no habría ningún problema, ya que al tratarse de un índice, carece de media y por tanto se puede comparar independientemente de la medida utilizada. La interpretación también es fácil: cuanto más cerca esté del valor 0, más equitativo será el reparto de la riqueza, mientras cuanto más cerca esté del valor 1, esta se concentrará en un valor.

## 5.3.2. Curva de Lorenz

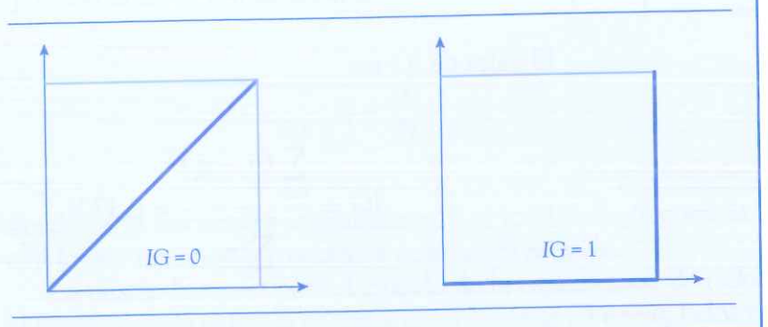
La curva de Lorenz es la gráfica de la función de los porcentajes de $p_i$ y $q_i$. Es una forma gráfica de mostrar el grado de dispersión o concentración.

El gráfico es siempre un cuadrado, y la gráfica una curva que se une al cuadrado por los valores (0, 0) a (100, 100), y siempre por debajo de la diagonal.

Por consenso se toman los valores menores a 0,5 como homogéneos y los mayores corresponderían a distribuciones con un alto índice de concentración.

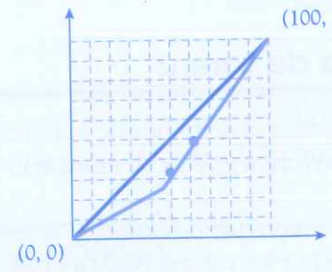
Si la variable tuviera una forma equitativa perfecta, su representación sería esa diagonal, mientras que el caso contario correspondería al eje horizontal ($p_i$). Su interpretación es muy sencilla: cuanto más cerca esté la curva de la distribución estudiada de la diagonal, menor concentración habrá, es decir, más equitativa será; mientras que si se acerca a los ejes de la parte inferior del cuadrado, mayor concentración tendrá.

**FIGURA 5.4**




Nuestro ejemplo quedaría así:

**FIGURA 5.5**




| $p_i$ (%) | $q_i$ (%) |
|---|---|
| 40 | 21,33 |
| 50 | 29,66 |
| 64 | 46,00 |



# 6 El estudio de las distribuciones bidimensionales. La covarianza, regresión lineal, bondad de ajuste y el coeficiente de correlación de Pearson

Llegados a este punto, es importante poder estudiar el efecto de más de una variable en una población determinada, poder analizar de forma conjunta varios caracteres y ver el grado de relación o dependencia que tienen entre sí. En este capítulo vamos a estudiar una serie de estadísticos que nos permitirán determinar el efecto de una variable sobre otra, cómo es su relación y la fuerza de la misma.

Para el estudio conjunto de dos variables el primer paso que debemos dar es la elaboración y tratamiento de los datos para su posterior interpretación; ello se hace mediante las llamadas **tablas de contingencia.** La clave para su interpretación y construcción es darnos cuenta de que se trata de una representación en la que los datos son compartidos por dos o más variables, y que la potencia de su uso se debe a que nos permite establecer si existe algún tipo de relación; por sí sola la tabla de contingencia no determina cómo es esa relación ni si existe causalidad entre ambas variables, para ello debemos utilizar una serie de estadísticos que se presentarán en este capítulo.

Los estadísticos que veremos a continuación son importantes para las Ciencias Sociales, ya que en este tipo de Ciencias es habitual interesarse por el efecto que tienen unas variables sobre otras: la edad, el sexo, el nivel cultural, el ocio, etc. Se trata de análisis multidimensionales en los cuales es importante ver las relaciones conjuntas: la edad y sus gustos, el sexo y la elección de destinos, el CI y el éxito o fracaso escolar... Para poder trabajar conjuntamente debemos comenzar por poder construir la tabla de frecuencias conjuntas; estamos ante una tabla de correlación o tabla de contingencia (tabla de doble entrada).

## 6.1. TABLAS DE CONTINGENCIA

Como ya hemos dicho, la tabla de contingencia muestra de forma conjunta la distribución de los datos de dos variables relacionadas.

Para su construcción debemos en primer lugar ver el tipo de variables que tenemos, ya que su representación variará en función de si se trata de distribuciones de tipo I, II o III. Este tipo de clasificación ya ha sido explicado en capítulos anteriores, recordaremos que las distribuciones tipo I son aquellas en las que la frecuencia de cada uno de los valores de las variables es uno; en la tabla de contingencia ocurre lo mismo, se trata de pares de valores de





ambas variables y que solamente aparecen una vez. Lo normal es encontrarnos con distribuciones tipo II, en las que los pares de valores aparecen en más de una ocasión, se pueden representar mediante tres columnas, las dos primeras con los valores de la variables y la tercera con la frecuencia conjunta; este tipo de representación puede ser difícil de comprender, por lo que se usan las tablas de contingencia. Cuando los valores de las variables se agrupan en intervalos, hablamos de distribuciones de tipo III.

Veamos unos ejemplos.

En una empresa de ocio y aventura han querido analizar la elección de sus diferentes productos ofertados y el sexo; para ello han tomado los datos de ocupación en los últimos tres meses. Los productos ofertados son: descenso en canoa, barranquismo, senderismo, escalada y rutas a caballo. Y los resultados obtenidos son los siguientes: la variable X (sexo) y la variable Y (oferta). La tabla de contingencia quedaría de la siguiente forma:

| X \ Y | Canoa (1) | Barranquismo (2) | Senderismo (3) | R. caballo (4) | Escalada (5) | Total |
|---|---|---|---|---|---|---|
| Mujer (1) | 125 | 35 | 60 | 80 | 25 | 325 |
| Hombre (2) | 130 | 50 | 75 | 70 | 55 | 380 |
| | 255 | 85 | 135 | 150 | 80 | 705 |

Otra forma de representar los datos sería:

| $X_i$ | $Y_j$ | $n_{ij}$ |
|---|---|---|
| 1 | 1 | 125 |
| 1 | 2 | 35 |
| 1 | 3 | 60 |
| 1 | 4 | 80 |
| 1 | 5 | 25 |
| 2 | 1 | 130 |
| 2 | 2 | 50 |
| 2 | 3 | 75 |
| 2 | 4 | 70 |
| 2 | 5 | 55 |
| | | 705 |

Como se puede ver, en este tipo de representación es más difícil ver los datos, la columna tercera muestra las frecuencias conjuntas de las dos variables y se representa por $n_{ij}$ (en la tabla de contingencia son los valores marcados en color claro). Es muy importante a la hora de analizar los datos conjuntos de dos o más variables saber usar ambas representaciones, ya que en ocasiones puede facilitar los cálculos.

La tabla de contingencia tiene otra ventaja añadida: nos referimos a la posibilidad de ofrecer fácilmente las **distribuciones marginales de las variables.** Se trata del número de veces que aparece un valor de una de las variables independientemente del valor de la otra. La forma más común de calcular las distribuciones marginales es ir sumando los valores por filas y columnas en las tablas de contingencia; en nuestro ejemplo son los valores de color oscuro, siendo las filas los valores de la distribución marginal de X y las columnas los valores de la distribución marginal de Y. Estos valores marginales son necesarios para el cálculo de los estadísticos analizados en los anteriores capítulos de análisis de una variable: media de X e Y, varianza $S_x^2$ y $S_y^2$, así como la desviación típica $S_x$ y $S_y$.

En nuestro ejemplo serían:

**Distribuciones marginales de *X***

| | $\sum_{j=1}^{k} n_{ij}$ |
|---|---|
| 1 | 325 |
| 2 | 380 |
| | $n_{ij} = 705$ |

**Distribuciones marginales de *Y***

| $Y_j$ | $\sum_{i=1}^{k} n_{ij}$ |
|---|---|
| 1 | 255 |
| 2 | 85 |
| 3 | 135 |
| 4 | 150 |
| 5 | 80 |
| | $n_{ij} = 705$ |

Para el cálculo de las medidas de posición y dispersión utilizamos estas tablas de distribuciones marginales y aplicamos las fórmulas que ya conocemos.





Aunque en el ejemplo que tenemos no tiene sentido calcular la media de las variables (ya que se trata de variables cualitativas discretas), la fórmula sería:

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

$$\overline{y} = \sum_{j=1}^{n} \frac{x_j \cdot n_j}{N}$$

Para el cálculo de la varianza podemos utilizar la fórmula conocida o bien utilizar las fórmulas derivadas de los momentos. La equivalencia es la siguiente:

$$S_x^2 = \alpha_{20} - \alpha_{10}^2$$

sabiendo que $\alpha_{10}$ es la media, y:

$$\alpha_{20} = \sum \frac{x_i^2 \cdot n_i}{N}$$

quedando la fórmula de la siguiente forma:

$$\sum \frac{x_i^2 \cdot n_i}{N} - \overline{x}^2$$

Lo mismo para el cálculo de la varianza de Y. Mediante la raíz cuadrada de cada una de la varianzas obtendremos la desviación típica de X e Y.

## 6.2. LA COVARIANZA

Desde un principio hemos venido diciendo que lo importante es poder saber si existe algún tipo de relación entre las variables

que estamos estudiando; aunque en la representación de la tabla de contingencia ya vemos la relación, la covarianza nos muestra con exactitud el tipo de relación que tienen.

La covarianza se calcula:

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

O bien la fórmula respecto al momento:

$$S_{xy} = \frac{\sum\sum x_i y_j n_{ij}}{N} - \bar{x}\,\bar{y}$$

En nuestro ejemplo se calcularía así:

| $X_i$ | $Y_j$ | $n_{ij}$ | $x_i \cdot y_j \cdot n_{ij}$ |
|---|---|---|---|
| 1 | 1 | 125 | 125 |
| 1 | 2 | 35 | 70 |
| 1 | 3 | 60 | 180 |
| 1 | 4 | 80 | 320 |
| 1 | 5 | 25 | 125 |
| 2 | 1 | 130 | 260 |
| 2 | 2 | 50 | 200 |
| 2 | 3 | 75 | 450 |
| 2 | 4 | 70 | 560 |
| 2 | 5 | 55 | 550 |
| | | 705 | 2.840 |

$$S_{xy} = \frac{2.840}{750} - 1{,}539 \cdot 2{,}7659 = 4{,}0283 - 4{,}2567 = -0{,}2284$$





### Interpretación de los resultados

Si $S_{xy} > 0$, existe una dependencia directa y positiva: cuanto más crece $X$, más crece $Y$.

Si $S_{xy} = 0$, las variables no tienen relación lineal entre ellas, aunque pueden tener otro tipo de relación.

Si $S_{xy} < 0$, existe una dependencia inversa o negativa: al aumentar una, disminuye la otra.

En nuestro caso la relación es negativa, lo que equivaldría a tener una relación inversa; no obstante, el valor calculado es prácticamente igual a 0, por lo que se puede decir que no existe una relación lineal entre ser mujer u hombre y elegir un tipo de actividad en nuestro establecimiento.

## 6.3. LA REGRESIÓN LINEAL

La regresión es un estadístico que trata de mostrar si existe una función que represente de forma sencilla la nube de puntos que presenta el conjunto de dos variables. Es decir, dada la representación de una distribución bidimensional, se puede observar que los datos de dicha distribución suelen formar relaciones que pueden ser descritas por una función matemática conocida; en otros casos no hay ningún tipo de formación que responda a una función matemática y se dice que las variables son independientes o incorreladas. El objetivo final de la regresión es determinar cómo se relacionan y establecer una función matemática que pueda predecir su comportamiento.

Se trata de poder predecir el valor de la variable dependiente (endógena o explicada) mediante una función de la variable independiente (exógena o explicativa). Existen diferentes tipos de funciones de ajuste: parabólico, cuadrático, logarítmico e hiperbólico; no obstante, el que más se utiliza es el ajuste lineal (la función es una recta) por su sencillez y poder predictivo, de ahí que en la mayoría de la bibliografía se hable de *regresión lineal*.

El objetivo es encontrar una función de la recta que mejor se ajuste a la nube de puntos, es decir, aquella que establezca el me-

nor error posible. Su representación es: $y = a + bx$, cuando hacemos $y$ sobre $x$; o bien $x = a + by$, cuando hacemos $x$ sobre $y$. Como hemos apuntado, se elige en función de la variable dependiente que estemos estudiando, en el primer caso $x$ es la variable independiente e $y$ la dependiente.

Debemos asumir la idea de que no existen distribuciones exactas. Es decir, es prácticamente improbable que una recta pase por todos los puntos de la distribución. Lo que hacemos es elegir aquella que expresa un menor error, ya que con ella pretendemos predecir el comportamiento de la variable dependiente en valores no determinados por la distribución (normalmente futuros). $[e = y_i - y'_i]$, dado que los errores pueden tomar valores positivos o negativos, realmente utilizamos la suma de los errores al cuadrado $\sum (y_i - y'_i)^2$.

Nuestro problema queda reducido a hallar los valores de $a$ y $b$ que hagan nuestra función de la recta más ajustada a nuestra distribución. Para ello hay diferentes métodos de solución:

$$y = a + bx$$

(consideramos en este caso $x$ la variable independiente).

1. $a = \bar{y} - \frac{S_{xy}}{S_x^2} \cdot \bar{x}$ (media de $y$ menos la covarianza dividida por la varianza de la variable independiente multiplicado por la media de X).
2. Mientras que $b = \frac{S_{xy}}{S_x^2}$ (covarianza dividida por la varianza de la variable independiente).
3. Si se sustituye el valor de $b$ en $a$, se obtiene que $a = \bar{y} - b\bar{x}$.

Si sustituimos los valores de $a$ y $b$ en la función de la recta podemos conseguir otra fórmula para obtener la recta de regresión:

$$y = \left(\bar{y} - \frac{S_{xy}}{S_x^2} \cdot \bar{x}\right) + \frac{S_{xy}}{S_x^2} x; \qquad \boxed{y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})}$$





Si se hubiera tomado la $y$ como variable independiente, sería:

$$x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

Se debe tener en cuenta que la pendiente de la recta $y$ sobre $x$ es la covarianza dividida por la varianza de X, mientras que la pendiente de $x$ sobre $y$ es la covarianza dividida por la varianza de Y. Esta última función nos permite decir que la recta hallada pasa por el punto formado por las dos medias, llamado centro de gravedad de la distribución bidimensional $(x, y)$. Otra propiedad de la regresión es que el producto de las pendientes de las rectas de regresión es igual al coeficiente de correlación al cuadrado. $b \times b' = r^2$.

**Veamos un ejemplo[1].**

Un profesor desea establecer la relación entre el tiempo de estudio y las notas que han sacado sus alumnos, para ello cuenta con los siguientes datos.

Para dar respuesta a esta pregunta haremos una regresión tomando como variable independiente el tiempo de estudio: $y = a + bx$. Pasamos a calcular $a$ y $b$, para ello seguiremos los pasos descritos: $a = \bar{y} - b\bar{x}$; $b$ es la covarianza dividida por la varianza de la variable independiente.

| Tiempo de estudio ($x_i$) | Nota ($y_i$) | $x_i \cdot n_i$ | $y_j \cdot n_j$ | $x_i^2 \cdot n_i$ | $x_i \cdot y_j \cdot n_{ij}$ |
|---|---|---|---|---|---|
| 60 | 7 | 60 | 7 | 3.600 | 420 |
| 64 | 8 | 64 | 8 | 4.096 | 512 |
| 67 | 8 | 67 | 8 | 4.489 | 536 |
| 70 | 9 | 70 | 9 | 4.900 | 630 |
| 72 | 10 | 72 | 10 | 5.184 | 720 |
| 90 | 8 | 90 | 8 | 8.100 | 720 |
| | | 423 | 50 | 30.369 | 3.538 |

[1] Es importante que nos demos cuenta de que a la hora de trabajar en situación real es difícil que nos pidan directamente calcular un regresión, es por ello que hay que hacer hincapié en la importancia que tienen los enunciados para familiarizarnos con lo que realmente significa la regresión, que no es otra cosa que una forma de predecir un comportamiento de dos variables relacionadas. Nos permite hacer pronósticos de valores intermedios y mayores.

La media de:

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i \cdot n_i}{N} = 70{,}5$$

La media de Y (previamente hallamos la marca $m$, al tratarse de una variable con intervalos):

$$\overline{y} = \sum_{j=1}^{n} \frac{y_j \cdot n_j}{N} = 8{,}33$$

Al tomar como variable independiente el tiempo de estudio, calculamos la varianza:

$$S_x^2 = \sum \frac{x_i^2 \cdot n_i}{N} = 5.061{,}5$$

Y finalmente calculamos la covarianza para poder determinar el valor de la pendiente $b$:

$$S_{xy} = \frac{\sum\sum x_i y_j n_{ij}}{n} - \overline{x}\,\overline{y} = 589{,}6666 - 587{,}4765 = 2{,}1901$$

$$b = \frac{2{,}1901}{5.061{,}5} = 0{,}0004$$

$$a = 8{,}33 - 0{,}0004(70{,}5) = 8{,}33 - 0{,}0305 = 8{,}2995$$

Quedando la función representativa de la relación entre el tiempo de estudio y la nota obtenida por los alumnos:

$$y = 8{,}2995 + 0{,}0004x$$

Es decir, que si se quisiera predecir la nota que puede tener un alumno con un tiempo de estudio de 50 minutos, obtendríamos el siguiente resultado:

$$y = 8{,}32$$





Otro método para el cálculo de los valores de $a$ y $b$ es mediante un sistema de ecuaciones. En el caso de y sobre $x$:

$$\begin{cases} \sum y_i = aN + b\sum x_i \\ \sum x_i y_i = a\sum x_i + b\sum x_i^2 \end{cases}$$

Mientras que si tomamos $x$ sobre $y$ ($y$ es la variable independiente):

$$\begin{cases} \sum x_i = a'N + b'\sum y_i \\ \sum x_i y_i = a'\sum y_i + b'\sum y_i^2 \end{cases}$$

Vamos a calcular el sistema con los datos del ejercicio anterior (tomando nuevamente la variable $x$ como independiente):

$$50 = 6a + 423b$$
$$3.538 = 423a + 30.369b$$

Despejamos en la primera $a$:

$$a = \frac{50 - 423b}{6}$$

y sustituimos en la segunda ecuación:

$$3.538 = 423\,\frac{50 - 423b}{6} + 30.369b$$

obtenemos: $b = 0{,}0024$, mientras que $a = 8{,}1641$, quedando la función resultante $y = 8{,}1641 + 0{,}0024x$; si pedimos, al igual que en el ejemplo anterior, la previsión de nota con un tiempo de estudio de 50 minutos, obtendremos el siguiente resultado:

$$y = 8{,}1641 + 0{,}0024(50) = 8{,}2841$$

La interpretación de los coeficientes es muy sencilla: *b* equivale a la pendiente de la recta. El signo de *b* nos permite decir si la relación es directa, cuando ambas variables X e Y crecen (+); o inversa cuando la variable X crece, e Y decrece (–), mientras que *a* se puede interpretar como el valor mínimo del que partimos, sea cual sea el valor de *x* (cuando esta actúa como variable independiente).

Los resultados obtenidos (las diferencias pueden deberse a la exactitud y cantidad de decimales utilizados en los cálculos) nos han dado la ecuación de una recta, pero debemos asegurarnos de que realmente nuestra distribución queda definida mediante esta ecuación de la recta. Es decir, cuando calculamos la recta de regresión, debemos asegurarnos de su representatividad, debemos saber si realmente ambas variables siguen una distribución que puede representarse mediante una función de una recta; debemos calcular su *bondad de ajuste*.

## 6.4. BONDAD DE AJUSTE (COEFICIENTE DE DETERMINACIÓN)

Las predicciones que realizamos con el cálculo de la regresión deben ajustarse lo más posible a la nube de puntos, es por ello por lo que debemos utilizar una medida que nos permita aceptar o rechazar dicho cálculo; esa medida es la *bondad de ajuste*.

Se trata de la proporción de variabilidad de la variable dependiente, que es explicada por el modelo de regresión. O lo que es lo mismo, el porcentaje de varianza de y que se puede explicar por *x*.

Este coeficiente sirve para cualquier tipo de regresión realizada, se representa mediante $R^2$ y sus valores oscilan entre 0 y 1 (0 % y 100 %), siendo el valor 0 una representación nula, mientras que 1 es la representación perfecta o ajuste perfecto. Se suelen recomendar valores superiores a 0,85 (85 %) para dar por buena la ecuación obtenida, mientras que valores inferiores a 0,60 tienen escasa fiabilidad, necesitando buscar otro tipo de representatividad.





Queda resuelto mediante la expresión (el cuadrado de la covarianza dividido por las varianzas):

$$R^2 = \frac{S_{xy}^2}{S_x^2 \cdot S_y^2} 100$$

En nuestro ejemplo la *bondad de ajuste* sería (previamente debemos calcular la varianza de y, ya que el resto de valores los hemos obtenido durante el cálculo de la regresión[2]):

$$R^2 = \frac{2{,}1901^2}{5.061{,}5 \cdot 70{,}3333} = 0{,}666, \text{ prácticamente } 0$$

La conclusión final es que la representación y el poder predictivo de la recta de regresión obtenida son nulos. A la hora de interpretar el coeficiente de determinación en una regresión podemos concluir que los resultados obtenidos en la regresión tienen un poder predictivo de $R^2$ (× 100), en nuestro caso sería de un 6,66 %.

## 6.5. COEFICIENTE DE CORRELACIÓN DE PEARSON ($r_{xy}$)

Con el coeficiente de correlación de Pearson podemos decir que se termina el estudio de la relación de dos variables, ya que nos permite establecer la fuerza de la relación de ambas variables y el sentido de la misma.

Para calcular el coeficiente de correlación de Pearson se utiliza la expresión (la covarianza dividida por el producto de las desviaciones típicas):

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

[2] A la hora de hallar la regresión lineal puede resultar más ventajoso, si nos piden la *bondad de ajuste*, las fórmulas que se desprenden de la varianza y covarianza, ya que tendremos muchos valores ya calculados.

Si observamos la fórmula estudiada de la *bondad de ajuste*, vemos que existe una relación matemática entre ambas, ya que la raíz cuadrada del coeficiente de determinación es la correlación de Pearson:

$$r_{xy} = \pm\sqrt{R^2}$$

El signo muestra el tipo de dependencia lineal (directa o inversa) y, como ya hemos apuntado, viene establecido por el valor de la covarianza. Es importante recordar esto, ya que $R^2$ oscila entre los valores 0 y 1, mientras que la correlación puede obtener valores numéricos inferiores a 0, estando comprendida entre –1 y +1. Este valor numérico nos indica la fuerza de asociación entre las variables estudiadas y su interpretación es:

*a*) $r_{xy} > 0$: la correlación lineal es positiva, siendo más fuerte cuanto más se acerque a 1.

*b*) $r_{xy} < 0$: la correlación lineal es negativa, siendo más fuerte cuanto más se acerque a –1.

*c*) $r_{xy} = 0$: no existe relación lineal entre ambas variables, aunque eso no indica que no pueda existir una relación de otro tipo, no lineal, o bien que ambas variables sean independientes.

Existen una serie de valores establecidos para la estimación de la correlación: se considera que no existe relación suficiente cuando $r_{xy}$ es inferior a 0,25; entre 0,25 y 0,50, correlación baja; entre 0,5 y 0,75, moderada; por encima de este valor se considera una correlación buena. Cuando se investiga en Ciencias Sociales se considera que existe una relación de dependencia lineal cuando esta alcanza valores superiores a ±0,80.

En nuestro ejemplo:

$$r_{xy} = \pm\sqrt{R^2} = +0{,}258$$

es decir, una correlación baja, prácticamente inexistente.





Cuando se interpreta este coeficiente debemos tener en cuenta la diferencia que existe entre correlación y causalidad, ya que el hecho de que la relación entre dos variables sea muy fuerte no significa que exista una dependencia entre ambas, no se puede establecer una relación de causalidad, ya que en la realidad estas relaciones pueden estar mediadas por una gran cantidad de variables intervinientes. Un ejemplo: si medimos la relación entre calzar más de un 42 y estar en la cárcel encontramos valores de $r_{xy}$ muy altos y positivos, ya que ambas están dependiendo del sexo de las personas encarceladas y sabemos que existe una relación de 2 a 10 a favor de los hombres, que suelen tener una talla de calzado más alta. No obstante, no podemos decir que la talla de calzado es una causa para ir a la cárcel.

# 7 Series temporales

Las Ciencias Sociales estudian fenómenos en constante cambio, dinámicos, lo que significa que deben estudiarse cada cierto tiempo (periódicamente) de forma cronológica y ordenada para poder ver el alcance de nuestros estudios y predicciones.

Si se quieren estudiar las variables asociadas a este tipo de fenómenos en profundidad, debemos contextualizarlos en períodos de tiempo que nos permitan ver los cambios y nos proporcionen toda la información posible sobre ellas. En este caso, podemos realizar estudios sobre el comportamiento de individuos en un período regular de tiempo o bien estudiar series bivariadas de datos en los cuales la variable independiente es el tiempo. Cuando hablamos de este tipo de estudios nos estamos refiriendo a **series temporales.**

Una definición sencilla sería una secuencia de observaciones recogidas de una variable a lo largo de un tiempo, siendo las observaciones ordenadas y ajustadas a un determinado intervalo temporal.

## 7.1. COMPONENTES Y CLASIFICACIÓN

Podemos clasificar las series temporales en deterministas o aleatorias. Esta diferenciación se hace en función de poder predecir el comportamiento de la variable de estudio en el futuro conociendo sus valores pasados de forma totalmente certera, sin error. Si no es así, decimos que nos encontramos ante series no deterministas o aleatorias, que son de las que se ocupa el análisis de series temporales. Ejemplo:

— **Determinista,** por ejemplo el número de matriculados en España en el grado de medicina de los últimos 20 años.

— **No determinista** (aleatoria), por ejemplo el número de graduados en medicina en esos mismos años.

En Ciencias Sociales nos encontramos con variables que tienen aspectos deterministas y aleatorios: por ejemplo el número de visitantes a un parque temático en el mes de febrero. Estamos





ante un número seguramente muy parecido, pero no será exactamente igual.

Independientemente del tipo de serie ante el que nos encontremos, todas ellas tienen los mismos componentes:

*a)* Tendencia secular o tendencia[1], se representa mediante la letra *T*: se trata de la dirección que mantiene la serie a lo largo de un tiempo amplio. Mediante este componente podemos determinar si sigue algún tipo de patrón establecido: crece, decrece, es alternante o estable.

*b)* Variaciones cíclicas, factor cíclico o ciclo, representada por la letra *V*: se trata de las variaciones más o menos estables, no fijas, que aparecen repetidas en períodos superiores a un año. Son percibidas cuando las observaciones se hacen en períodos medios o largos.

*c)* Variaciones estacionales, estacionalidad o componente estacional, se representa mediante la letra *S*: analiza las variaciones de la serie en períodos de un año o inferiores. Ocurren normalmente en aquellas variables en las que un período concreto de tiempo influye de forma clara en el conjunto de la serie. En el caso del turismo es muy claro, al ser un sector que tiene actividades sujetas a temporalidad por motivos climáticos, culturales, sociales...

*d)* Variaciones accidentales, fluctuaciones irregulares o componente irregular, representadas por *I*: se trata de variaciones asociadas a fenómenos aislados, sin ningún tipo de relación entre ellos y que pueden producir variaciones imprevistas. Acontecimientos concretos, como el hecho de que un país organice los juegos olímpicos pueden hacer que el análisis de ciertas variables contenga cambios bruscos e impredecibles.

Los componentes de las series temporales se comportan de forma diferente en función del tipo de variable; esto hace que el

[1] Los componentes de las series temporales, aunque siempre son los mismos, pueden aparecer con diferente nomenclatura; nosotros para una mejor comprensión hemos recogido las más citadas.

análisis de las series temporales esté sujeto a una serie de hipótesis. Existen dos hipótesis mayormente aceptadas: la hipótesis aditiva y la multiplicativa, aunque la mayoría de los paquetes estadísticos informatizados han optado por una combinación de ellas:

1. Hipótesis aditiva: $Yt = T + V + S + I$.
2. Hipótesis multiplicativa: $Yt = T \cdot V \cdot S \cdot I$.
3. Combinación de hipótesis: $Yt = T \cdot V \cdot S + I$.

Estas hipótesis tratan de decirnos que el valor de la serie temporal en un momento dado es igual a la suma, el producto o la combinación de ambas de los valores de estos cuatro componentes.

Debemos tener en cuenta que para una serie concreta no tienen por qué darse todos los componentes. En algunas concretamente, si no son muy largas, es difícil distinguir entre la tendencia y la variación cíclica; pensemos que tenemos los datos de una variable recogida durante una período de tres años, puede ocurrir que en los años posteriores decrezca (sería cíclica) o que siga creciendo (sería tendencia). Otro caso puede ser que tengamos series con períodos anuales de datos, en este caso carecería del componente variación estacional.

EJEMPLO. En el caso antes expuesto el número de graduados en medicina en los últimos 20 años.

*Tt*: lo que han crecido o decrecido en ese período de tiempo.

*Ct*: períodos asociados a la demanda laboral.

*St*: en este caso no se da al tener datos con periodicidad anual.

*It*: cambios de normativa de *numerus clausus* que pueden variar el número de matriculados, cambios de leyes educativas...

Debido a todos estos factores intervinientes a la hora de analizar las series temporales, debemos centrarnos en aquellos componentes considerados fundamentales: la tendencia y el componente estacional, siendo la tendencia considerada como el componente que más influye a largo plazo.





## 7.2. CÁLCULO DE LA TENDENCIA

Es sin duda el componente más importante, el que más influye a largo plazo, siendo fundamental para realizar predicciones sobre cómo se va a desarrollar la serie en el futuro. Los otros componentes pueden introducir modificaciones parciales o mínimas, como es el caso de la estacionalidad que veremos en este capítulo.

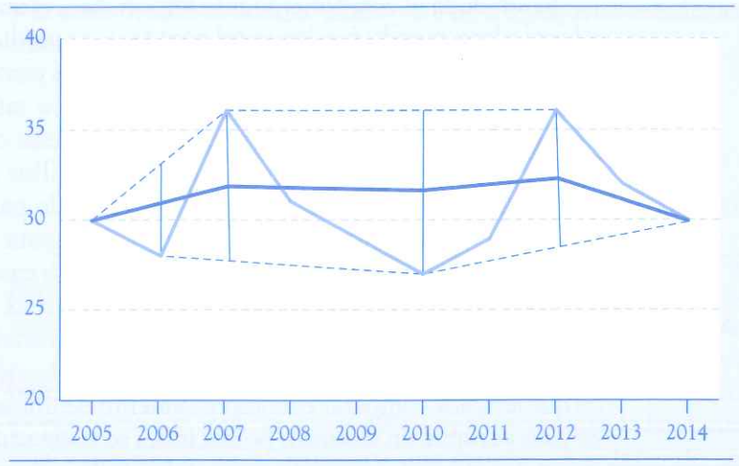
Existen cuatro métodos para el cálculo y análisis de la tendencia. A la hora de elegir el método más adecuado para calcular la tendencia de nuestra serie hay que tener en cuenta si la variable que estamos analizando es determinista, en cuyo caso deberíamos utilizar el método de mínimos cuadrados, mientras que si por el contrario tenemos una serie temporal de una variable evolutiva (no determinista), podemos utilizar el método de medias móviles. Si quisiéramos comparar el comportamiento de una serie temporal de una variable en distintas poblaciones se suele utilizar el método de dos mitades o semipromedio, sobre todo por su simplicidad y facilidad para interpretar los resultados. No obstante, cualquiera de los métodos es susceptible de uso para el análisis de la tendencia.

### 1. Método gráfico

Hemos querido incluir este método de cálculo para que tengamos una visión lo más amplia posible del cálculo de la tendencia. Sin embargo, podemos afirmar que se trata de una aproximación gráfica a la tendencia secular; se recoge en pocos manuales debido a su escaso poder predictivo. Para poder realizar este método partimos de la representación en un sistema de ejes cartesianos de la serie: en el eje horizontal de tiempo y en el vertical la variable Y. Una vez dibujada, se unen los puntos máximos de la serie mediante una poligonal y se hace lo mismo con los valores mínimos. Una vez realizadas, podemos dibujar una línea equidistante que se obtiene uniendo los puntos medios de los segmentos obtenidos por los valores máximos y mínimos. Esta nueva línea poligonal nos muestra la tendencia secular.

EJEMPLO. Se pretende estudiar el número de aprobados de la asignatura de estadística en los últimos 10 años, obteniendo los siguientes resultados:

**FIGURA 7.1**




Las líneas punteadas corresponden a los valores máximos y mínimos, las azules a los segmentos verticales y la línea gruesa azul es la tendencia secular.

### 2. Método de las medias móviles

Este método se encuentra entre los tres clásicos modelos de análisis y cálculo de la tendencia. Una media móvil no es más que un promedio (media) de un número determinado de datos. Se usa para evitar la influencia de un dato en particular; para ello se halla la media aritmética de los valores de la variable con los valores contiguos. Una vez hallados, se construye una nueva serie con los valores obtenidos mostrando la tendencia secular de la serie, ya que esta nueva serie que se obtiene tiene una menor dispersión.

Su cálculo varía en función de los datos promedio que utilicemos para obtener la media móvil: las más utilizadas son las medias de tamaño impar (3, 5, ...), esto es debido a que si se utilizan valores pares, estos no quedan bien centrados y es necesario volver a promediar. El número de datos que se usa para promediar se denomina *tamaño de las medias móviles*.





Hemos de tener en cuenta que cuanto mayor sea el orden, más información perderemos, ya que siempre quedan medias móviles sin determinar al principio y al final de la serie; otro problema es la elección del tamaño de las medias móviles, ya que se trata de una decisión aleatoria y, por tanto, es variable en función del número elegido. Dada una serie temporal, la tendencia secular se calcula:

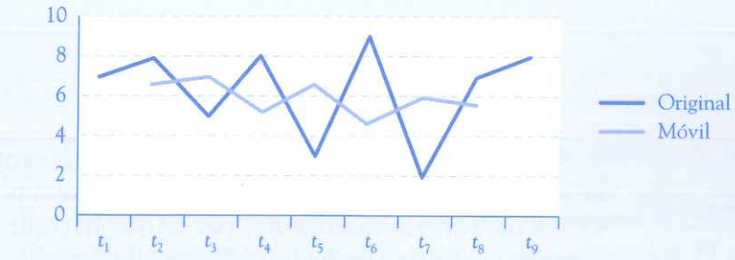
| $t_i$ | $y_i$ | $\bar{y}_i$ | $y_i$ | $\bar{y}_i$ |
|---|---|---|---|---|
| $t_1$ | $y_1$ | — | 7 | — |
| $t_2$ | $y_2$ | $\bar{y}_2$ | 8 | 6,66 |
| $t_3$ | $y_3$ | $\bar{y}_3$ | 5 | 7,00 |
| $t_4$ | $y_4$ | $\bar{y}_4$ | 8 | 5,33 |
| $t_5$ | $y_5$ | $\bar{y}_5$ | 3 | 6,66 |
| $t_6$ | $y_6$ | $\bar{y}_6$ | 9 | 4,66 |
| $t_7$ | $y_7$ | $\bar{y}_7$ | 2 | 6,00 |
| $t_8$ | $y_8$ | $y_8$ | 7 | 5,66 |
| $t_9$ | $y_9$ | — | 8 | — |

Tomaremos como tamaño de las medias móviles tres:

$$y_2 = \frac{y_1 + y_2 + y_3}{3}; \quad y_3 = \frac{y_2 + y_3 + y_4}{3}; \quad y_4 = \frac{y_3 + y_4 + y_5}{3}$$

y así sucesivamente con todos los valores hasta completar la serie.

**FIGURA 7.2**



Como se puede observar, la serie obtenida muestra una menor dispersión. Si en lugar de haber utilizado como tamaño de muestra tres se hubiera elegido un número par, por ejemplo cuatro, la tabla de datos sería la siguiente:

| $t_i$ | $y_i$ | $\bar{y}_i$ | $y_i$ | $\bar{y}_i$ | $\bar{\bar{y}}_i$ |
|---|---|---|---|---|---|
| $t_1$ | $y_1$ | | 7 | | |
| $t_2$ | $y_2$ | | 8 | | |
| | | $\bar{y}_3$ | | 7,00 | |
| $t_3$ | $y_3$ | | 5 | | 6,25 |
| | | $\bar{y}_4$ | | 5,25 | |
| $t_4$ | $y_4$ | | 8 | | 5,75 |
| | | $\bar{y}_5$ | | 6,25 | |
| $t_5$ | $y_5$ | | 3 | | 5,90 |
| | | $\bar{y}_6$ | | 5,50 | |
| $t_6$ | $y_6$ | | 9 | | 5,40 |
| | | $\bar{y}_7$ | | 5,25 | |
| $t_7$ | $y_7$ | | 2 | | 5,90 |
| | | $\bar{y}_8$ | | 6,50 | |
| $t_8$ | $y_8$ | | 7 | | |
| $t_9$ | $y_9$ | | 8 | | |

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4}{4}$$

$$\bar{y}_4 = \frac{y_2 + y_3 + y_4 + y_5}{4}$$

$$\bar{y}_5 = \frac{y_3 + y_4 + y_5 + y_6}{4}$$

Tal como hemos dicho, la serie no quedaría centrada, ya que el valor dos no corresponde a ningún tiempo concreto $t_2$ ni $t_3$, sino a un período intermedio, por lo que hay que ajustar a los tiempos mediante el cálculo de una segunda media, pero esta vez ajustada:





$$\bar{y}_3 = \frac{y_3 + y_4}{2}$$

nuevamente hasta terminar con los valores de la variable.

Como podemos ver, en este caso se pierden más valores, aunque el resultado es una serie con una tendencia más delimitada.

### 3. Método de las dos medias, dos mitades o semipromedios

El objetivo de este método es encontrar dos puntos que determinen la recta que forma la tendencia de la serie: $(\bar{t}_1, \bar{y}_1)$ y $(\bar{t}_2, \bar{y}_2)$.

Para su cálculo el primer paso es distribuir los datos que tenemos en dos grupos **impares** aproximadamente iguales y se calcula la media aritmética de cada grupo (en ocasiones será necesario que un dato aparezca en los dos grupos o bien podremos eliminar uno). Una vez calculadas las medias de los dos grupos, se procede a trasladar los datos a la fórmula de la recta siguiente:

$$y = \bar{y}_1 + \frac{\bar{y}_2 - \bar{y}_1}{\bar{t}_2 - \bar{t}_1}(t - \bar{t}_1)$$

Se puede obtener por semipromedios, y la función será:

$$y = a + bx;$$

siendo:

$$a = y_2 - b\bar{t}_2$$

$$b = \frac{\bar{y}_2 - \bar{y}_1}{\bar{t}_2 - \bar{t}_1}$$

EJEMPLO. El Principado de Asturias quiere calcular la tendencia de cancelaciones realizadas en sus hoteles de cuatro estrellas durante los últimos años cara a instaurar un sistema de calidad. Para ello ha construido una tabla con los datos obtenidos. El método elegido es por mitades, de las dos medias.

| Año | Datos | Período | Semipromedios o medias |
|---|---|---|---|
| 2004 | 5.523 | 1 | |
| 2005 | 5.544 | 2 | |
| 2006 | 6.304 | 3 | 6.066,6 |
| 2007 | 6.480 | 4 | |
| 2008 | 6.482 | 5 | |
| 2009 | 6.572 | 6 | |
| 2010 | 6.605 | 7 | |
| 2011 | 6.545 | 8 | |
| 2012 | 6.175 | 9 | 6.199,4 |
| 2013 | 5.833 | 10 | |
| 2014 | 5.839 | 11 | |

El primer paso es dividir nuestra serie en dos grupos de cinco años, eliminando el año central.

Hemos insertado una columna que nos permite cuantificarlos para poder hacer el cálculo por semipromedios más sencillo y poder expresarlo en función de $x$.

Sustituimos primero en la función por mitades:

$$y = 6.066,6 + \frac{6.199,4 - 6.066,6}{2012 - 2006}(t - 2006)$$

Si queremos estimar el número de cancelaciones para el año 2016, obtendríamos:

$$y_{2016} = 6.066,6 + 22,13(2016 - 2102) = 6.287,9$$

Por semipromedios obtendremos:

$$b = 6.199,4 - \frac{6.066,6}{9 - 3} = 22,13$$

$$a = 6.199,4 - (22,13 \times 9) = 6.199,4 - 199,2 = 6.000,2$$

La recta resultante es, por tanto:

$$y = 6.000,2 + 22,13x.$$





Si queremos calcular el número de cancelaciones para el año 2016, entonces:

$$y = 6.000,2 + (22,13 \times 13) = 6.287,9$$

### 4. Método de mínimos cuadrados (método del ajuste)

Es el método más utilizado; se trata básicamente de una regresión, tal como se ha visto en el capítulo anterior, pero la variable independiente explicativa es el tiempo.

Se trata del método más exacto y tiene la ventaja de contar con el coeficiente de determinación o bondad de ajuste, que nos permite poder determinar la fiabilidad de los resultados y poder establecer si las predicciones pueden ser tomadas o no en consideración.

La recta de tendencia será, por tanto, la siguiente: $y = a + bt$.

La variable tiempo puede simplificarse para realizar los cálculos; para ello se hace un cambio en la variable en función de si los datos son pares o impares:

Cuando el número de datos es impar, la transformación se hace $t' = t - O_t$, siendo $O_t$ el valor central de los tiempos; si es par $t' = 2(t - O_t)$, siendo el valor de $O_t$ la media de los dos valores centrales.

La función de la recta quedará entonces $\mathbf{y = a + b \times t'}$.

Ejemplo.

| Años impar | Período $t'$ | $(t - \bar{t}')$; $\bar{t}' = 5$ |
|---|---|---|
| 2002 | 1 | −4 |
| 2003 | 2 | −3 |
| 2004 | 3 | −2 |
| 2005 | 4 | −1 |
| 2006 | 5 | 0 |
| 2007 | 6 | 1 |
| 2008 | 7 | 2 |
| 2009 | 8 | 3 |
| 2010 | 9 | 4 |

| Años par | Período $t'$ | $2(t - \bar{t}')$; $\bar{t}' = 5,5$ |
|---|---|---|
| 2002 | 1 | −9 |
| 2003 | 2 | −7 |
| 2004 | 3 | −5 |
| 2005 | 4 | −3 |
| 2006 | 5 | −1 |
| 2007 | 6 | 1 |
| 2008 | 7 | 3 |
| 2009 | 8 | 5 |
| 2010 | 9 | 7 |
| 2011 | 10 | 9 |

Se halla la media en ambos casos; impar 2006 = 5; par 2006-2007; 5 + 6/2 = 5,5. La recta queda $y = a + bt'$, haciendo el cambio $y = a + b(t - O_t)$ o si es par $y = a + 2b(t - O_t)$.

Una vez realizado el cambio se procede como en el caso de la recta de regresión: se halla $a$ y $b$, la recta de regresión mediante un sistema de ecuaciones será:

$$\begin{cases} \sum y_i = a \cdot N \\ \sum t'_i \cdot y_i = b \cdot \sum t'^2_i \end{cases}$$

o bien mediante las fórmulas que hemos utilizado en el capítulo correspondiente a la recta de regresión, teniendo en cuenta que la variable tiempo es en este caso la variable independiente:

$$y = a + bt'$$

(consideramos en este caso $t$ la variable independiente).

1. $a = \bar{y} - \frac{s_{xy}}{s_t^2}\bar{x}$ (media de y menos la covarianza dividida por la varianza de la variable independiente multiplicado por la media de X).
2. Mientras que $b = \frac{s_{ty}}{s_t^2}$ (covarianza dividida por la varianza de la variable independiente).
3. Si se sustituye el valor de $b$ en $a$, se obtiene que $a = \bar{y} - b\bar{t}$.

Ejemplo. El número de turistas en fronteras que han llegado a España en los últimos 10 años, según datos de FRONTUR, se recoge en la tabla siguiente; calcular la tendencia secular mediante mínimos cuadrados y la bondad de ajuste de dicho cálculo.





| Año | Número (millones) | $t$ | $t' = 2(\bar{t} - t')$ | $t'_i \times y_i$ | $t'^2_i$ |
|---|---|---|---|---|---|
| 2005 | 55,91 | 1 | –9 | –503,19 | 81 |
| 2006 | 58,00 | 2 | –7 | –406,00 | 49 |
| 2007 | 58,66 | 3 | –5 | –293,30 | 25 |
| 2008 | 57,19 | 4 | –3 | –171,57 | 9 |
| 2009 | 52,17 | 5 | –1 | –52,17 | 1 |
| 2010 | 52,67 | 6 | 1 | 52,67 | 1 |
| 2011 | 56,17 | 7 | 3 | 168,50 | 9 |
| 2012 | 57,46 | 8 | 5 | 287,30 | 25 |
| 2013 | 60,60 | 9 | 7 | 424,20 | 49 |
| 2014 | 64,90 | 10 | 9 | 584,10 | 81 |
| | 573,73 | 55 | 0 | 90,54 | 330 |

En primer lugar procedemos a la transformación de la variable tiempo, al tratarse de una serie par utilizamos $t' = 2(t - O_t)$. Obtenemos las medias $t' = 5,5$ e $y = 57,373$.

Vamos a utilizar el cálculo de la tendencia mediante el sistema de ecuaciones:

$\sum y_i = a \times N$; $573,73 = 10a$, obtenemos $a = 57,373$

$\sum t'_i \times y_i = b \times \sum t'^2_i$; $90,54 = b330$, obtenemos $b = 0,2743$

La recta de regresión quedaría $y = 57,373 + 0,2743t'$.

Vamos a utilizar los mismos valores, pero calcularemos la recta de regresión con la covarianza y la varianza.

La tendencia tendría la siguiente forma:

$$y = a + bt'$$

siendo:

$$a = \bar{y} - b\bar{t}; \quad \text{y} \quad b = \frac{s_{ty}}{s_t^2}$$

| Año | Número (millones) | $t$ | $t'=2(\bar{t}-t')$ | $(y-\bar{y})$ | $(y-\bar{y})(\bar{t}-t')$ | $(y-\bar{y})^2$ | $(\bar{t}-t')^2$ |
|---|---|---|---|---|---|---|---|
| 2005 | 55,91 | 1 | –9 | –1,463 | 13,167 | 2,1403 | 81 |
| 2006 | 58,00 | 2 | –7 | 0,625 | –4,375 | 0,3906 | 49 |
| 2007 | 58,66 | 3 | –5 | 1,287 | –6,435 | 1,6563 | 25 |
| 2008 | 57,19 | 4 | –3 | –0,183 | 0,549 | 0,0334 | 9 |
| 2009 | 52,17 | 5 | –1 | –5,203 | 5,203 | 27,0712 | 1 |
| 2010 | 52,67 | 6 | 1 | –4,703 | –4,703 | 22,1182 | 1 |
| 2011 | 56,17 | 7 | 3 | -1,203 | –3,609 | 1,4472 | 9 |
| 2012 | 57,46 | 8 | 5 | 0,087 | 0,435 | 0,0007 | 25 |
| 2013 | 60,60 | 9 | 7 | 3,227 | 22,589 | 10,4135 | 49 |
| 2014 | 64,90 | 10 | 9 | 7,117 | 64,053 | 50,6516 | 81 |
| | 573,73 | 55 | 0 | | 86,874 | 115,9231 | 330 |

La media de $y$ = 57,373.
La covarianza $S_{yt}$ = 86,874/10 = 8,6874.
La varianza de $t$:

$$S_t^2 = \frac{330}{10} = 33$$

$$b = \frac{S_{yt}}{S_t^2} = 0{,}2632$$

$$a = \bar{y} - b\bar{t} = 57{,}373 - 0{,}2632 \cdot 5{,}5 = 55{,}9254$$

La recta de tendencia quedaría $y = 55{,}9254 + 0{,}2632t'$.

Si con ambas soluciones tratamos de predecir el valor de $y$ en el año 2015, obtendremos:

$$y = 57{,}373 + 0{,}2743t' = 57{,}373 + 0{,}2743 \times 11 = 60{,}4$$

mientras que en el otro cálculo sería:

$$y = 56 + 0{,}2632 \times 11 = 58{,}9$$





es posible que las diferencias, aunque pequeñas, se deban a los ajustes.

En estos casos, como ya hemos dicho, debemos calcular la bondad de ajuste, que nos dirá si estos valores de predicción que salen de esta recta de tendencia pueden ser considerados como válidos:

$$R^2 = \frac{(S_{ty})^2}{S_t^2 \cdot S_y^2}$$

la covarianza al cuadrado dividido por las varianzas:

$$R^2 = \frac{75{,}4709}{33 \cdot 11{,}5923} = 0{,}1973$$

Tal como hemos dicho en el capítulo 6, este resultado se da en porcentaje, por tanto el valor predictivo será de un 19,73 %, por lo que debemos concluir que la tendencia no sigue una distribución lineal y su poder predictivo es muy pequeño. A la hora de interpretar los resultados, se procede de forma similar a la regresión: podemos identificar si crece o decrece en función del signo de la pendiente ($b$), intrapolar o predecir (extrapolar) resultados y mediante la bondad de ajusta determinar el valor de los resultados obtenidos.

## 7.3. CÁLCULO Y ANÁLISIS DE LAS VARIACIONES ESTACIONALES

Cuando estudiamos series temporales en períodos más o menos largos observamos que aparecen datos que forman figuras en la gráfica recurrentes, esto puede deberse a ciertos patrones horarios, mensuales, trimestrales, anuales, etc. Cuando estas variaciones tienen una duración superior a un año, se habla de variaciones cíclicas, mientras que si son inferiores, estaremos ante variaciones estacionales.

La mayoría de las variables que se analizan en sectores de las Ciencias Sociales suelen verse afectadas por este tipo de variaciones estacionales, por ejemplo el turismo, los efectos de las rebajas en las ventas...

Este tipo de efectos pueden afectar a la capacidad de tomar decisiones o generar planes de actuación en las empresas, de ahí la importancia que tiene eliminar el efecto que produce este tipo de variaciones para poder controlar y analizar las variables de estudio sin el efecto que producen las variaciones estacionales; a este procedimiento se denomina desestacionalización. Con ello podemos obtener una nueva serie sin el efecto estacional y poder analizar la variable en un período concreto de una forma más fiable. Existen varios métodos para este cálculo, nosotros vamos a elegir uno de ellos.

Pasos para el cálculo del índice de variación estacional (IVE):

1. Se calcula la media anual para cada uno de los años de la serie: $\bar{y}_t$.
2. Se ajusta la recta de regresión a las medias anuales: $\bar{y}_t = a + bt$.
   - I. $b$ es el incremento del valor de la media anual.
   - II. Número de divisiones tomadas en la estacionalización anual; si es trimestral $m = 4$; semestral $m = 2$, etc.
   - III. $b/m$ es el incremento por estación.
3. En este paso se resta a la media de cada estación la proporción del incremento anual:

$$\bar{y}'_k = \bar{y}_k - \frac{b(k-1)}{m}$$

en donde toma los valores $k = 1, 2, 3, ..., m$, realizando tantas como estaciones hemos tomado en el estudio ($m$).

4. El último caso es calcular el IVE para cada una de las estaciones determinadas:

$$IVE = \frac{\bar{y}'_k}{\bar{y}'} 100$$





Veamos un ejemplo.

Disponemos de la afluencia por trimestres ($x10^3$) de un tren de alta montaña en los últimos cuatro años. Queremos calcular la variación estacional de los datos.

| Trimestres ($k$) | 2011 | 2012 | 2013 | 2014 | Total | $\bar{y}_k$ | $y'_k$ | IVE |
|---|---|---|---|---|---|---|---|---|
| Primero | 6,0 | 8,00 | 9,00 | 11,00 | 34 | 8,50 | 8,500 | 49,18 |
| Segundo | 16,0 | 17,00 | 18,00 | 21,00 | 72 | 18,00 | 17,520 | 101,37 |
| Tercero | 29,0 | 31,00 | 34,00 | 39,00 | 133 | 33,25 | 32,300 | 186,86 |
| Cuarto | 11,0 | 13,00 | 12,00 | 14,00 | 50 | 12,50 | 10,820 | 62,59 |
| Total (anual) | 62,0 | 69,00 | 73,00 | 85,00 | | | 69,130 | |
| $\bar{y}_i$ | 15,5 | 17,25 | 18,25 | 21,25 | | $\bar{y}'$ | 17,283 | |

El siguiente paso es calcular $b$, incremento del valor de la media anual:

| Año | $\bar{y}_i$ |
|---|---|
| 2011 | 15,50 |
| 2012 | 17,25 |
| 2013 | 18,25 |
| 2014 | 21,25 |

$$b = \frac{(17{,}25 - 15{,}50) + (18{,}25 - 17{,}25) + (21{,}25 - 18{,}25)}{3} = 1{,}91$$

Una vez que hemos calculado $b$, pasamos a calcular las medias corregidas de las estaciones mediante la fórmula:

$$\bar{y}'_k = \bar{y}_k - \frac{b(k-1)}{m}$$

una por cada estación delimitada en los datos.

$$y_{k1} = 8{,}5 - \frac{1{,}91(1-1)}{4} = 8{,}5$$

$$y_{k2} = 18 - \frac{1{,}91(2-1)}{4} = 17{,}52$$

$$y_{k3} = 33{,}25 - \frac{1{,}91(3-1)}{4} = 32{,}295$$

$$y_{k4} = 12{,}25 - \frac{1{,}91(4-1)}{4} = 10{,}8175$$

El último paso es calcular el índice VE de cada estación estudiada:

$$IVE_1 = \frac{8{,}5}{17{,}283} \times 100 = 49{,}18$$

$$IVE_2 = \frac{17{,}52}{17{,}283} \times 100 = 101{,}37$$

$$IVE_3 = \frac{32{,}295}{17{,}283} \times 100 = 186{,}86$$

$$IVE_4 = \frac{10{,}8175}{17{,}283} \times 100 = 62{,}59$$

Con este cálculo hemos conseguido eliminar el efecto de la estacionalidad y podemos interpretar de forma fácil los datos que tenemos. En nuestro caso vemos que el peso de la componente estacional se da sobre todo en el tercer trimestre y algo menos en el segundo; podemos afirmar que el uso de este transporte de montaña desarrolla mayoritariamente su actividad en los meses centrales del año y sobre todo en los meses de julio a agosto.

Si queremos estudiar una serie temporal en la que pretendemos calcular la tendencia de la misma y vemos que puede tratarse de una serie con una influencia estacional, debemos introducir un paso más **la desestacionalización.**





## 7.4. DESESTACIONALIZACIÓN DE UNA SERIE

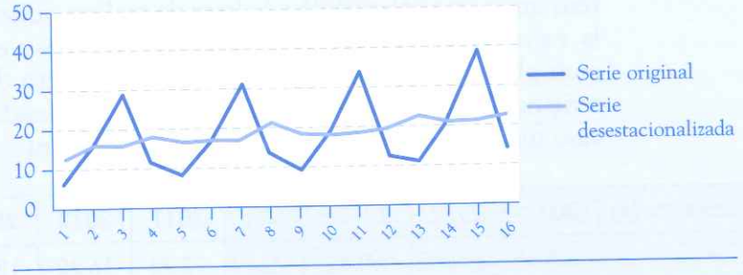
Es en este momento cuando debemos incluir el tipo de hipótesis que hemos planteado a la hora de realizar nuestro estudio de la variable, ya que dependiendo de si la hipótesis es sumativa o multiplicativa restaremos o dividiremos el efecto del IVE de la serie original para obtener la serie desestacionalizada. En nuestro caso tomamos la hipótesis multiplicativa, por tanto dividimos:

| Trimestres (*k*) | 2011 | 2012 | 2013 | 2014 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|
| Primero | 6/0,49 | 8/0,49 | 9/0,49 | 11/0,49 | 12,24 | 16,32 | 18,36 | 22,45 |
| Segundo | 16/1,01 | 17/1,01 | 18/1,01 | 21/1,01 | 15,84 | 16,83 | 17,82 | 20,79 |
| Tercero | 29/1,86 | 31/1,86 | 34/1,86 | 39/1,86 | 15,59 | 16,66 | 18,28 | 20,97 |
| Cuarto | 11/0,62 | 13/0,62 | 12/0,62 | 14/0,62 | 17,74 | 20,97 | 19,36 | 22,58 |

El paso siguiente es calcular la recta de tendencia para la serie original y la serie desestacionalizada. Para ello los datos deben colocarse en el orden temporal:

| Datos | Serie original | Serie desestacionalizada |
|---|---|---|
| 1 | 6 | 12,24 |
| 2 | 16 | 15,84 |
| 3 | 29 | 15,59 |
| 4 | 11 | 17,74 |
| 5 | 8 | 16,32 |
| 6 | 17 | 16,83 |
| 7 | 31 | 16,66 |
| 8 | 13 | 20,97 |
| 9 | 9 | 18,36 |
| 10 | 18 | 17,82 |
| 11 | 34 | 18,28 |
| 12 | 12 | 19,36 |
| 13 | 11 | 22,45 |
| 14 | 21 | 20,79 |
| 15 | 39 | 20,97 |
| 16 | 14 | 22,58 |

**FIGURA 7.3**




Como se observa en el gráfico, la serie desestacionalizada muestra una tendencia más clara que la serie original al eliminar el efecto de la estacionalización de los datos.

Si calculamos las rectas de tendencia, veremos que la capacidad de predicción es mucho más potente:

| Datos | $t' = 2(t - t')$ media = 8,5 | Serie original | Serie desest. | $(y - \bar{y})$ original | $(y - \bar{y})(t - t')$ | $(t - t')^2$ | $(y - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | –15 | 6 | 12,24 | –12,0625 | 180,9375 | 225 | 145,5039 |
| 2 | –13 | 16 | 15,84 | –2,0625 | 26,8125 | 169 | 4,2539 |
| 3 | –11 | 29 | 15,59 | 10,9375 | –120,3125 | 121 | 119,6289 |
| 4 | –9 | 11 | 17,74 | –7,0625 | 63,5625 | 81 | 49,8789 |
| 5 | –7 | 8 | 16,32 | –10,0625 | 70,4375 | 49 | 101,2539 |
| 6 | –5 | 17 | 16,83 | –1,0625 | 5,3125 | 25 | 1,2890 |
| 7 | –3 | 31 | 16,66 | 12,9375 | –38,8125 | 9 | 167,3789 |
| 8 | –1 | 13 | 20,97 | –5,0625 | 5,0625 | 1 | 25,2890 |
| 9 | 1 | 9 | 18,36 | –9,0625 | –9,0625 | 1 | 82,1289 |
| 10 | 3 | 18 | 17,82 | –0,0625 | –0,1875 | 9 | 0,0039 |
| 11 | 5 | 34 | 18,28 | 15,9375 | 79,6875 | 25 | 254,0039 |
| 12 | 7 | 12 | 19,36 | –6,0625 | –42,4375 | 49 | 36,7539 |
| 13 | 9 | 11 | 22,45 | –7,0625 | –63,5625 | 81 | 49,8789 |
| 14 | 11 | 21 | 20,79 | 2,9375 | 32,3125 | 121 | 8,6289 |
| 15 | 13 | 39 | 20,97 | 20,9375 | 272,1875 | 169 | 438,3789 |
| 16 | 15 | 14 | 22,58 | –4,0625 | –60,9375 | 225 | 16,0390 |
| | | 289 | | 0 | 401 | 1.360 | 1.500,2927 |





Cálculo para la serie original:

La media de y:

$$\bar{y} = \frac{289}{16} = 18{,}0625$$

La covarianza:

$$S_{yt} = \frac{401}{16} = 25{,}0625$$

La varianza de $t$:

$$S_t^2 = \frac{1.360}{16} = 85$$

$$b = \frac{S_{yt}}{S_t^2} = 0{,}2948$$

$$a = \bar{y} - b\bar{t} = 18{,}0625 - 0{,}2948 \cdot 8{,}5 = 15{,}5567$$

La recta de tendencia quedaría:

$$y = 15{,}5567 + 0{,}2948t'.$$

Para hallar la bondad de ajuste ($R^2$), necesitamos calcular la varianza de y, $S_y^2$:

$$S_y^2 = \frac{1.500{,}2927}{16} = 93{,}7682$$

$$R^2 = \frac{(25{,}0625)^2}{85} \cdot 93{,}7682 = 0{,}0788 \times 100 = 7{,}88\,\%$$

La capacidad de predicción es casi nula. Veamos la nueva recta de tendencia desestacionalizada:

| Datos | $t' = 2(t - t')$ media = 8,5 | Serie original | Serie desest. | $(y - \bar{y})$ desest. | $(y - \bar{y})(t - t')$ | $(t - t')^2$ | $(y - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | –15 | 6 | 12,24 | –6,06 | 90,90 | 225 | 36,7236 |
| 2 | –13 | 16 | 15,84 | –2,46 | 31,98 | 169 | 6,0516 |
| 3 | –11 | 29 | 15,59 | –2,71 | 29,81 | 121 | 7,3441 |
| 4 | –9 | 11 | 17,74 | –0,56 | 5,04 | 81 | 0,3136 |
| 5 | –7 | 8 | 16,32 | –1,98 | 13,86 | 49 | 3,9204 |
| 6 | –5 | 17 | 16,83 | –1,47 | 7,35 | 25 | 2,1609 |
| 7 | –3 | 31 | 16,66 | –1,64 | 4,92 | 9 | 2,6896 |
| 8 | –1 | 13 | 20,97 | 2,67 | –2,67 | 1 | 7,1289 |
| 9 | 1 | 9 | 18,36 | 0,06 | 0,06 | 1 | 0,0036 |
| 10 | 3 | 18 | 17,82 | –0,48 | –1,44 | 9 | 0,2304 |
| 11 | 5 | 34 | 18,28 | –0,02 | –0,10 | 25 | 0,0004 |
| 12 | 7 | 12 | 19,36 | 1,06 | 7,42 | 49 | 1,1236 |
| 13 | 9 | 11 | 22,45 | 4,15 | 37,35 | 81 | 17,2225 |
| 14 | 11 | 21 | 20,79 | 2,49 | 27,39 | 121 | 6,2001 |
| 15 | 13 | 39 | 20,97 | 2,67 | 34,71 | 169 | 7,1289 |
| 16 | 15 | 14 | 22,58 | 4,28 | 64,20 | 225 | 18,3184 |
| | | | 292,80 | 0 | 350,78 | 1.360 | 116,5606 |

La media de y:

$$\bar{y} = \frac{292,8}{16} = 18,3$$

La covarianza:

$$S_{yt} = \frac{350,78}{16} = 21,9237$$

La varianza de $t$:

$$S_t^2 = \frac{1.360}{16} = 85$$





$$b = \frac{S_{yt}}{S_t^2} = 0,0161$$

$$a = \bar{y} - b\bar{t} = 18,3 - 0,0161 \times 8,5 = 18,1632$$

La recta de tendencia quedaría:

$$y = 18,1632 + 0,0161t'.$$

Para hallar la bondad de ajuste ($R^2$) necesitamos calcular la varianza de y, $S_y^2$:

$$S_y^2 = \frac{116,5606}{16} = 7,2850$$

$$R^2 = \frac{(21,9237)^2}{85} \times 7,2850 = 0,7762 \times 100 = 77,62\ \%$$

La capacidad de predicción ha mejorado considerablemente, quedando demostrada la importancia que tiene este proceso en las series temporales.

## 7.5. AUTOCORRELACIÓN

En muchos de los libros que analizan las series temporales incluyen el concepto de autocorrelación; con este cálculo se trata de ver la dependencia que tiene una serie consigo misma en un período anterior, por ejemplo la relación entre los valores de un mes concreto en los diferentes años recogidos en una serie temporal mensual. Este cálculo se hace mediante el coeficiente de correlación de la serie original y la serie que se obtiene de aplicar un retardo $h$. La fórmula sería la siguiente:

$$r = \frac{S_{t,t-h}}{S_t S_{t-h}}$$

Este cálculo permite establecer relaciones en series grandes en períodos concretos y poder establecer conclusiones del comportamiento de la variable, para poder hacer inferencias sobre este comportamiento.

# 8 Los números índice

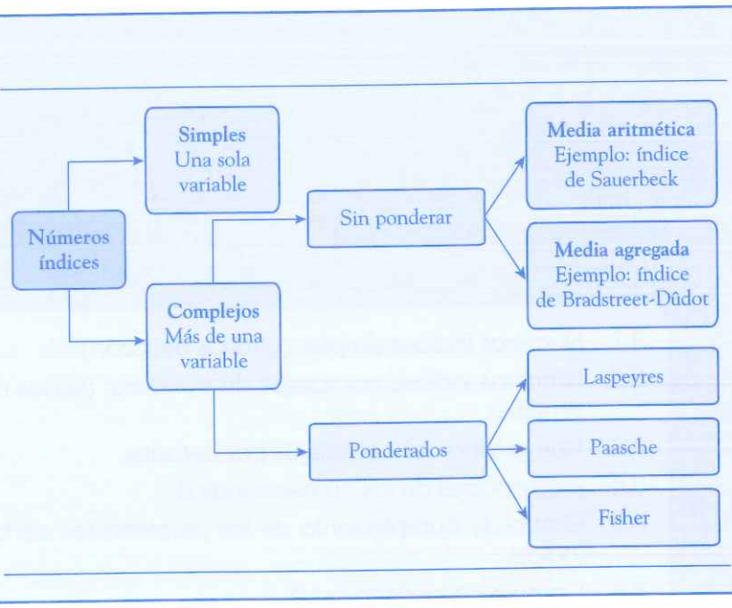
Se trata de una medida que pone de manifiesto los cambios de una variable o variables relacionadas con respecto a una característica, que suele ser el tiempo (o el espacio), con respecto a un momento dado o punto de referencia que se toma como base.

Uno de sus usos es comparar diferentes series de forma sencilla, para ello se expresa en tantos por ciento, lo que permite poder comparar las variables, aunque originalmente tengan diferentes medidas. El punto de partida para el análisis es una referencia arbitraria que se denomina período base (este período base debe ser un período normal).

Estos cálculos tienen muchas aplicaciones, las más usada son los índices de precios, cotizaciones, crecimiento o decrecimiento de empresas, gastos...

En el siguiente esquema podemos ver una clasificación de los números índice[1].

**FIGURA 8.1**




[1] Esquema adaptado tomado de Ronquillo, A. (1997). *Estadística aplicada al sector turístico. Técnicas cuantitativas y cualitativas de análisis turístico* (p. 120). Madrid: Centro de Estudios Ramón Areces.





## 8.1. NÚMEROS ÍNDICES SIMPLES (O RATIOS DE RAZÓN)

Tal como parece en el esquema, se trata de una sola variable. Tenemos una serie temporal $(t_i, y_i)$, en primer lugar se fija un valor $t$ de la serie como período base (tomamos un año normal arbitrario): el número índice simple se halla dividiendo cada valor de la serie por el valor correspondiente al período tomado como base y multiplicando el resultado por cien.

Al valor del período base lo denominamos $t_0$:

$$I_i = \frac{y_i}{y_0} 100$$

EJEMPLO. Tenemos los datos de ocupación de un establecimiento hotelero de una ciudad a lo largo de un período de diez años y queremos estudiar su evolución:

| $t$ | $y_i$ | Índice simple | Índice |
|---|---|---|---|
| 2006 | 2.950 | $(y_0/y_0) \times 100$ | 100,00 |
| 2007 | 2.500 | $(y_1/y_0) \times 100$ | 84,74 |
| 2008 | 2.650 | $(y_2/y_0) \times 100$ | 89,83 |
| 2009 | 2.900 | | 98,30 |
| 2010 | 3.000 | | 101,70 |
| 2011 | 3.100 | | 105,10 |
| 2012 | 3.200 | | 108,50 |
| 2013 | 3.150 | | 106,80 |
| 2014 | 3.300 | | 111,90 |
| 2015 | 3.500 | $(y_n/y_0) \times 100$ | 118,60 |

La interpretación de estos resultados es la siguiente:

En los primeros años, hasta 2010, experimenta una bajada que en 2007 es del 15,26 %, que se va reduciendo en 2008 a un 10,17 %, hasta una reducción del 1,7 % en 2009. A partir de ahí sufre unos incrementos que van desde el 1,7 % de 2010 hasta el 18,6 % del año 2015.

Como se puede comprobar, se trata de una cálculo sencillo y fácil de interpretar.

## 8.2. NÚMEROS ÍNDICES COMPLEJOS SIN PONDERAR (VARIAS MAGNITUDES SIMPLES)

Se trata de establecer la evolución a través del tiempo de un grupo de variables de forma conjunta, estableciendo un único índice para todo el grupo de variables. Este tipo de cálculo puede ser muy útil en diferentes campos de la industria y comercio, ya que nos permite establecer una unidad de control de varios productos simultáneamente. Por ejemplo, en una planta de envasado podemos obtener un índice de la evolución de los diferentes productos que se trabajan en conjunto, en un alojamiento con varios tipos de precios en sus instalaciones, etc.

Siguiendo el esquema inicial, vemos que aparecen dos tipos de números índices complejos sin ponderar: los más utilizados son el índice de Sauerbeck para la media aritmética simple y el índice de Bradstreet-Dûtot para la media agregada simple. El cálculo de ambos métodos es muy sencillo.

### 8.2.1. Índice de Sauerbeck

Se trata de las medias de los diferentes índices simples de las variables de estudio en una serie temporal:

$$S_t = \frac{\sum_{i=1}^{n} \frac{Y_{it}}{Y_{i0}}}{N} 100$$

EJEMPLO. En la empresa Socitia, S. L., desean saber la evolución de precios experimentada en conjunto durante los últimos siete años; esta empresa cuenta con tres diferentes productos cuyo valor por pieza se recoge en la tabla siguiente:





| Año | Producto 1 | Producto 2 | Producto 3 | $Ip_1$ | $Ip_2$ | $Ip_3$ | $S_t$ |
|---|---|---|---|---|---|---|---|
| 2009 | 3,5 | 5,2 | 6,0 | 100,00 | 100,00 | 100,00 | 100,00 |
| 2010 | 3,7 | 5,4 | 6,3 | 105,71 | 103,84 | 105,00 | 104,85 |
| 2011 | 3,9 | 5,9 | 6,8 | 111,42 | 113,46 | 113,33 | 112,73 |
| 2012 | 4,3 | 6,4 | 7,0 | 122,85 | 123,07 | 116,66 | 120,86 |
| 2013 | 4,5 | 6,8 | 7,8 | 128,57 | 130,76 | 130,00 | 129,77 |
| 2014 | 5,1 | 7,1 | 8,0 | 145,71 | 136,53 | 133,33 | 138,52 |
| 2015 | 5,9 | 7,8 | 8,4 | 168,57 | 150,00 | 140,00 | 152,85 |

El primer paso es elegir un año base, que en este caso, al querer establecer una evolución desde el inicio de la actividad, tomaremos el año 2009. En segundo lugar procedemos a calcular los índices simples de los tres productos. El tercer paso es hallar la media aritmética de los índices simples de los artículos por año, que en nuestro caso consiste en dividir entre tres.

Una de las ventajas de este tipo de cálculos es que nos permite ver no solo la evolución de los productos en conjunto, sino también cuál de los productos tiene un mayor peso en la evolución de la empresa y los cambios en todos los productos.

Hemos dicho que el índice de Sauerbeck es el más utilizado, pero podemos hallar otros índices complejos sin ponderar; para ello podemos modificar el paso dos e introducir, en función del tipo de variable que estemos estudiando, la media geométrica o la media armónica[2].

### 8.2.2. Índice de Bradstreet-Dûtot

La diferencia con la media aritmética simple es que en este caso se suman primero los valores de las variables de estudio por año y se divide por el valor obtenido en el año tomado como base:

$$B - Dt = \frac{\sum_{i=1}^{n} Y_{it}}{\sum_{i=1}^{n} Y_{i0}} 100$$

[2] Se procede a su cálculo tal como hemos visto en el capítulo de las medidas de concentración.

Para nuestro ejemplo hemos tomado las notas medias académicas de los alumnos aprobados en sociología en los últimos cinco años de tres facultades de la Universidad de Oviedo:

| Año | F1 | F2 | F3 | F1 + F2 + F3 | $B - D_t$ |
|---|---|---|---|---|---|
| 2011 | 6,9 | 6,7 | 6,8 | 20,4 | 100,00 |
| 2012 | 7,3 | 6,9 | 7,0 | 21,2 | 103,90 |
| 2013 | 7,5 | 7,3 | 7,8 | 22,6 | 110,78 |
| 2014 | 7,7 | 7,5 | 8,0 | 23,2 | 113,72 |
| 2015 | 7,9 | 7,8 | 8,1 | 23,8 | 116,66 |

Se puede observar la evolución en el tiempo de las notas medias de aprobados, en este caso el mayor incremento se ha dado entre los años 2012 y 2013, que sube 6,88 % la nota media. Las ventajas de este cálculo son su rapidez; sin embargo, perdemos la información de las diferentes facultades, que sí tendríamos con el índice anterior.

## 8.3. NÚMERO ÍNDICES COMPLEJOS PONDERADOS

Existen varias razones para el uso de números índices ponderados, la más relevante es la posibilidad de poder otorgar un valor diferente a cada una de las variables que componen el índice proporcional a su aportación al conjunto de las variables. Los números índices anteriores otorgan la misma importancia a todas las variables; sin embargo, en la realidad existen variables que tienen una mayor aportación al valor global que estamos calculando y mediante esta ponderación podemos compensar estas diferencias. En economía, turismo y otras disciplinas profesionales encontramos variables que se ven afectadas por variaciones relativas a características que no se pueden separar de la propia variable.

Por ejemplo, si queremos saber el índice de precios de un hotel o el crecimiento de los ingresos de empresas que trabajan con diferentes artículos o productos, debemos tener en cuenta que cada uno de los artículos o productos puede verse condicionado por la





cantidad o el precio. Podemos encontrar productos cuyo precio es alto, pero su volumen es pequeño, o viceversa, con lo que a la hora de tomar decisiones basándonos en los valores de una determinada empresa puede ser importante tener en cuenta estas variaciones.

Habitualmente se trabaja con variables que incluyen precio y cantidad, es por ello por lo que cuando se quiere establecer un índice de precios se usan para su ponderación las cantidades y, por el contrario, si consideramos las cantidades, se usará el precio para su ponderación.

Existen varios tipos de números índices ponderados, como se puede ver en el esquema inicial, y aunque en este capítulo nos centraremos en los tres más utilizados: Laspeyres, Paasche y Fisher, comenzaremos por un cálculo más sencillo basado en el uso de la media.

### 8.3.1. Índices de medias ponderadas de índices simples[3]

Queremos elaborar un índice ponderado para un grupo de productores vinícolas teniendo en cuenta sus producciones. Contamos con su índice de evolución de los últimos cinco años para el cálculo de la evolución del grupo. El primer paso es elaborar la tabla de ponderación $w$:

| Productor | Producción (l) | Ponderador (%) |
|---|---|---|
| P1 | 15.000 | 15,62 |
| P2 | 17.000 | 17,70 |
| P3 | 22.000 | 22,92 |
| P4 | 19.000 | 19,89 |
| P5 | 23.000 | 23,96 |
| **Total** | **96.000** | **100,00** |

[3] Este tipo de cálculo trabaja sobre índices, no directamente con datos de variables.

Si consideramos 96.000 = 100, tenemos que la tabla de ponderación será un cálculo sencillo en función de la aportación de cada uno de los miembros de grupo. En un segundo momento aplicamos a la tabla de índices* de los productores este ponderador y aplicamos la fórmula:

| Año* | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| 2011 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| 2012 | 102,8 | 100,9 | 103,3 | 102,7 | 104,8 |
| 2013 | 103,9 | 97,3 | 105,4 | 97,5 | 101,5 |
| 2014 | 105,6 | 103,6 | 102,3 | 99,3 | 102,6 |
| 2015 | 110,7 | 107,8 | 108,7 | 100,5 | 107,2 |

$$I = \frac{I_1 w_1 + I_2 w_2 + I_3 w_3 + \cdots + I_N w_N}{w_1 + w_2 + w_3 + \cdots + w_N} 100 = \frac{\sum_{i=1}^{n} I_i w_i}{\sum_{i=1}^{n} w_i} 100 =$$

$$I_{2012} = \frac{102{,}8 \cdot 0{,}1562 + 100{,}9 \cdot 0{,}177 + 103{,}3 \cdot 0{,}2292 + 102{,}7 \cdot 0{,}1989 + 104{,}8 \cdot 0{,}2396}{0{,}1562 + 0{,}177 + 0{,}2292 + 0{,}1989 + 0{,}2396}$$

Hacemos el mismo cálculo para cada uno de los años y obtenemos los índices ponderados en función de la producción de este grupo de viticultores. Si fuera necesario, podemos utilizar, en función de las variables que estemos analizando, la media geométrica o la media armónica, tal como se ha visto en el capítulo de las medidas de concentración.

### Índice de Laspeyres

Este el índice que utiliza el INE (Instituto Nacional de Estadística) para elaborar el IPC (índice de precios al consumo). La razón de esta elección se basa en las características específicas de su cálculo, ya que para ello es necesario saber únicamente los valores de ponderación del año base y, una vez determinados, solo es necesario investigar la evolución de los precios en los períodos sucesivos.





Su uso no obstante tiene un inconveniente, que se infiere de su ventaja: y es que a medida que nos alejemos del año base los datos de ponderación están más desfasados y, por tanto, se alejan de la realidad.

Para su cálculo podemos utilizar sus dos formulaciones, una en la que la ponderación es la cantidad y otra en la que la ponderación es el precio, de ahí que se hable de índice de Laspeyres de precios y de cantidades. Si calculamos el índice de variación de precios ($L_{pt}$) se usa como ponderador la cantidad, y al contrario, si queremos el índice de variación de cantidades ($L_{qt}$), usaremos como ponderador el precio. Su formulación es la siguiente:

$$L_{pt} = \frac{\sum_{i=1}^{n} P_{it} \cdot Q_{i0}}{\sum_{i=1}^{n} P_{i0} \cdot Q_{i0}} \cdot 100; \qquad L_{qt} = \frac{\sum_{i=1}^{n} Q_{it} \cdot P_{i0}}{\sum_{i=1}^{n} Q_{i0} \cdot P_{i0}} \cdot 100$$

Ejemplo. Disponemos de los precios y pernoctaciones de un alojamiento en los últimos seis años en las diferentes categorías de habitaciones que ofrece (individual, doble y *suite*), y queremos establecer la evolución de precios de nuestro alojamiento en este período de tiempo.

Para ello vamos a utilizar el índice de precios de Laspeyres:

| Año | Individual *P* | *Q* | Doble *P* | *Q* | *Suite P* | *Q* | $L_{pt}$ |
|---|---|---|---|---|---|---|---|
| 2010 | 45 | 500 | 41 | 834 | 125 | 248 | 100,00 |
| 2011 | 48 | **523** | 43 | **852** | 128 | **239** | 104,40 |
| 2012 | 50 | **511** | 48 | **811** | 130 | **245** | 110,92 |
| 2013 | 53 | **498** | 50 | **793** | 135 | **240** | 115,94 |
| 2014 | 55 | **503** | 52 | **832** | 140 | **253** | 120,40 |
| 2015 | 60 | **524** | 55 | **861** | 145 | **250** | 127,52 |

Hemos marcado en negrita las cantidades de los años 2011 a 2015 para recordar que en este tipo de cálculo no es necesario contar con todos los datos de la variable que usaremos para ponderar los datos, basta con tener las cantidades de un año (que se usará como año base). En nuestro caso hemos tomado como año base 2010. Procedemos a colocar nuestros datos en la fórmula:

$$Lp_{2010} = \frac{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 100$$

$$Lp_{2011} = \frac{48 \cdot 500 + 43 \cdot 834 + 128 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 104{,}4$$

$$Lp_{2012} = \frac{50 \cdot 500 + 48 \cdot 834 + 130 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 110{,}92$$

$$Lp_{2013} = \frac{53 \cdot 500 + 50 \cdot 834 + 135 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 115{,}94$$

$$Lp_{2014} = \frac{55 \cdot 500 + 52 \cdot 834 + 140 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 120{,}40$$

$$Lp_{2015} = \frac{60 \cdot 500 + 55 \cdot 834 + 145 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 127{,}52$$

Como ya hemos dicho en varias ocasiones a lo largo de los diferentes capítulos, lo importante de la estadística es la interpretación de los datos, no solamente el cálculo.

En el caso de los números índice la interpretación es muy intuitiva. En nuestro ejemplo el índice de precios ha ido subiendo en estos años, siendo las mayores subidas las recogidas en los años 2012, con un 7,3 %, y en el año 2015, con un incremento del 7,12 %. Además, se puede calcular el crecimiento medio de todos estos años y poder marcar objetivos empresariales si fuera necesario.

Hemos de tener en cuenta a la hora de interpretar los resultados que el índice de Laspeyres tiene la ventaja de poder comparar todos los años del período estudiado con referencia del año base,





lo que nos permite establecer una línea evolutiva continua, es decir, podemos concluir que desde el año 2010 al año 2013 hay un incremento del 15,94 %, y así podemos ir comparando cualquier dato de la serie.

### Índice de Paasche

Se trata nuevamente de un índice con dos formulaciones, al igual que Laspeyres: un índice de precios ($P_p$) y un índice de cantidades ($P_q$); el funcionamiento de la ponderación sigue el mismo razonamiento que en el caso anterior. Si estamos calculando el índice de precios, ponderamos con cantidades, y si estamos calculando el índice de cantidades, ponderamos con el precio.

Las ponderaciones en el caso de Paasche son el producto del precio del año base por la cantidad del año en que estemos calculando el índice de precios, y en el caso de cantidades es el producto de la cantidad del año base por el precio del año en que se esté calculando el índice. Podemos ver que es muy parecido al índice de Laspeyres, pero tiene como ventaja que las ponderaciones son actuales; para ello es necesario contar con todos los datos del período de estudio y esto hace que sea más costoso.

Esta ventaja de actualización de la ponderación a la hora de interpretar los datos se convierte en un inconveniente, ya que al variar cada año solo nos permite comparar los índices del período de estudio con el año base y no entre ellos. Esta razón y su alto coste hacen que sea menos utilizada.

Su formulación es la siguiente:

$$P_{pt} = \frac{\sum_{i=1}^{n} P_{it} \cdot Q_{it}}{\sum_{i=1}^{n} P_{i0} \cdot Q_{it}} \cdot 100; \qquad P_{qt} = \frac{\sum_{i=1}^{n} Q_{it} \cdot P_{it}}{\sum_{i=1}^{n} Q_{i0} \cdot P_{it}} \cdot 100$$

EJEMPLO. Vamos a tomar los datos del ejercicio anterior, pero esta vez calcularemos el índice de Paasche de precios:

| Año | Individual $P$ | $Q$ | Doble $P$ | $Q$ | Suite $P$ | $Q$ | $P_{pt}$ |
|---|---|---|---|---|---|---|---|
| 2010 | 45 | 500 | 41 | 834 | 125 | 248 | 100,00 |
| 2011 | 48 | 523 | 43 | 852 | 128 | 239 | 104,51 |
| 2012 | 50 | 511 | 48 | 811 | 130 | 245 | 110,88 |
| 2013 | 53 | 498 | 50 | 793 | 135 | 240 | 115,92 |
| 2014 | 55 | 503 | 52 | 832 | 140 | 253 | 120,34 |
| 2015 | 60 | 524 | 55 | 861 | 145 | 250 | 127,64 |

$$Pp_{2010} = \frac{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248}{45 \cdot 500 + 41 \cdot 834 + 125 \cdot 248} 100 = 100$$

$$Pp_{2011} = \frac{48 \cdot 523 + 43 \cdot 852 + 128 \cdot 239}{45 \cdot 523 + 41 \cdot 852 + 125 \cdot 239} 100 = 104{,}51$$

$$Pp_{2012} = \frac{50 \cdot 511 + 48 \cdot 811 + 130 \cdot 245}{45 \cdot 511 + 41 \cdot 811 + 125 \cdot 245} 100 = 110{,}88$$

$$Pp_{2013} = \frac{53 \cdot 498 + 50 \cdot 793 + 135 \cdot 240}{45 \cdot 498 + 41 \cdot 793 + 125 \cdot 240} 100 = 115{,}92$$

$$Pp_{2014} = \frac{55 \cdot 503 + 52 \cdot 832 + 140 \cdot 253}{45 \cdot 503 + 41 \cdot 832 + 125 \cdot 253} 100 = 120{,}34$$

$$Pp_{2015} = \frac{60 \cdot 524 + 55 \cdot 861 + 145 \cdot 250}{45 \cdot 524 + 41 \cdot 861 + 125 \cdot 250} 100 = 127{,}64$$

La interpretación de los resultados nos dice que del año 2010 a 2011 experimenta un incremento del 4,51 %; a 2012 un incremento del 10,88 %; a 2013 un incremento del 15,92 %; en el año 2014 un incremento del 20,34 % y con el año 2015 un incremento del 27,64 %. Debido a la variación de la ponderación de año en año realizar una interpretación entre cada uno de los años del período de estudio sería un error.





### Índice de Fisher

Este índice es la media geométrica de los dos índices anteriores, se trata pues de un cálculo a partir de números índices y no de valores directos de las variables de estudio.

Su formulación es:

$$Fp_t = \sqrt{Lp \cdot Pp}; \quad Fq_t = \sqrt{Lq \cdot Pq}$$

Ejemplo. Tomaremos los datos obtenidos en los ejercicios de Laspeyres y Paasche de precios:

| Año | $L_p$ | $P_p$ | $F_p$ |
|---|---|---|---|
| 2010 | 100,00 | 100,00 | 100,00 |
| 2011 | 104,40 | 104,51 | 104,40 |
| 2012 | 110,92 | 110,88 | 110,89 |
| 2013 | 115,94 | 115,92 | 115,92 |
| 2014 | 120,40 | 120,34 | 120,36 |
| 2015 | 127,52 | 127,64 | 127,57 |

$$Fp_{2011} = \sqrt{104{,}40 \cdot 104{,}40} = 104{,}40$$

$$Fp_{2012} = \sqrt{110{,}92 \cdot 110{,}88} = 110{,}89$$

$$Fp_{2013} = \sqrt{115{,}94 \cdot 115{,}92} = 115{,}92$$

$$Fp_{2014} = \sqrt{120{,}40 \cdot 120{,}34} = 120{,}36$$

$$Fp_{2015} = \sqrt{127{,}52 \cdot 127{,}64} = 127{,}57$$

La interpretación de los resultados se hace igual que en los casos anteriores, comparando los índices anuales y determinando su crecimiento o decrecimiento.

Existen otra serie de números índices menos utilizados, entre ellos:

1. El índice de Drovisch-Bowley, que es la media aritmética de los índices de Laspeyres y Paasche:

$$I_{DB} = \frac{I_p + P_p}{2}$$

Al igual que en los anteriores, existe también para cantidades.

2. Índice de Edgeworth-Marshall: se trata de una media agregada ponderada de las cantidades o precios del año base y el año en estudio:

$$I^n_{EM_0} = \frac{\sum_{i=1}^{N} P_n(Q_0 + Q_1)}{\sum_{i=1}^{N} P_0(Q_0 + Q_1)} 100$$

3. Índice de Walch:

$$IPW^n_0 = \frac{\sum_{i=1}^{N} P_n(Q_0 \cdot Q_1)}{\sum_{i=1}^{N} P_0(Q_0 \cdot Q_1)} 100$$

## 8.4. PROPIEDADES DE LOS NÚMEROS ÍNDICE[4]

Se puede decir que tienen cinco propiedades básicas:

1. *Existencia*. Todo número índice debe existir, tiene un valor finito distinto de 0.

[4] Escuder, R. (1987). *Métodos estadísticos aplicados a la economía*. Barcelona: Ariel. Martín-Guzmán, P. y Martín Pliego, F. J.(2012). *Curso básico de estadística*. Madrid: AC. Martín Pliego, F. J. (2012). *Introducción a la estadística económica y empresarial*. Madrid: AC.





2. *Identidad*. Si de hace coincidir el período base y el actual, el valor debe ser igual a la unidad o al 100 %: $I^n_n = 100$.
3. *Inversión*. El valor del índice debe ser invertible al cambiar los períodos entre sí. Esto se traduce en que si calculamos el índice para el año base = 0 con la base del año $t$, tiene que ser igual al inverso del índice del año $t$ calculado basándose en el año 0.

$$I^0_t = \frac{1}{I^t_o}$$

4. *Proporcionalidad*. Si las magnitudes de un período manifiestan una variación proporcional, el número índice tiene que manifestar esa variación.
5. *Homogeneidad*. Un número índice no puede verse afectado por cambios en las unidades de media de las variables.

## 8.5. GRADO DE CUMPLIMIENTO DE LAS PROPIEDADES DE LOS NÚMEROS ÍNDICE

La existencia y la identidad la cumplen todos los índices estudiados, la inversión solo es cumplida por los índices de Bradstreet-Dûtot, Edgeworth y Fisher. La proporcionalidad, aunque se cumple, produce ciertos problemas con los índices ponderados, excepto con Laspeyres. Finalmente, la homogeneidad no se cumple en ninguno de los números índice analizado. Teniendo en cuenta este breve análisis de las propiedades, vemos que el mejor de los índices es Lapeyres, ya que es el único índice ponderado que cumple la proporcionalidad sin presentar grandes errores.

## 8.6. CAMBIO DE PERÍODO BASE

En ocasiones es importante actualizar los índices que hemos establecido; imaginemos que hemos utilizado el índice de Las-

peyres o cualquier otro y tenemos un año base alejado de la fecha actual, es muy probable que hayamos perdido poder de medida, en estos casos es necesario cambiar el año base para recuperar rigurosidad en nuestras medidas. El nuevo año elegido como año base pasará a tomar el valor 100 y el resto de valores se obtiene dividiendo cada uno de los valores de la serie entre el valor del índice del nuevo período base. Solamente estamos actualizando los datos, en estos casos no hay cambios en las variables de medida ni en otro dato utilizado.

Tomamos para nuestro ejemplo uno de los índices del ejercicio anterior. En este caso el año que teníamos como base era el año 2010 y queremos actualizar los datos tomando como nuevo año base el 2013:

| Año | $L_p$ | Índice base 2013 |
|---|---|---|
| 2010 | 100,00 | (100,00/115,94) × 100 = 86,25 |
| 2011 | 104,40 | (104,40/115,94) × 100 = 90,04 |
| 2012 | 110,92 | (110,92/115,94) × 100 = 95,67 |
| 2013 | 115,94 | 100,00 |
| 2014 | 120,40 | (120,40/115,94) × 100 = 103,85 |
| 2015 | 127,52 | (127,52/115,94) × 100 = 109,88 |

La interpretación de los resultados con el nuevo año base sería la siguiente:

El resultado del año 2012 es un 4,33 % menor que en el año 2013, el resultado del año 2011 un 9,96 % menor, mientras que el año 2014 pone de manifiesto un subida del 3,85 %, y así procederíamos con el resto de los valores obtenidos en la nueva serie.

## 8.7. RENOVACIÓN Y ENLACE

Tal como hemos dicho anteriormente, los números índice necesitan renovarse para que los resultados sean un buen reflejo de la realidad. En ocasiones ocurre que los productos para el cálculo del índice pueden cambiar por la propia medida de la variable,





como el caso del IPC, en el cual hay productos que dejan de ser relevantes y otros ocupan su lugar, puede que unos valores desaparezcan y aparezcan nuevos elementos a formar parte de la variable.

Cuando esto ocurre, no se pueden comparar entre sí, ya que estaríamos cambiando el sentido y contenido de los índices. Es entonces cuando debemos utilizar lo que se denomina enlace. Lo que tenemos delante es una serie con diferentes medidas y que no nos permitiría establecer una trayectoria temporal del producto o productos estudiados.

Para poder hacer el enlace o renovación debemos disponer de los números índices calculados de las diferentes maneras que aparecen a lo largo del período que deseamos calcular.

Ejemplo. Disponemos de los índices de precios de un establecimiento hotelero, que en los últimos seis años ha cambiado de categoría y de tipo de servicios ofrecidos. Los resultados de las dos series son los siguientes:

| Año | Índice 1, base 2005 | Índice 2, base 2012 | Serie enlazada |
|---|---|---|---|
| 2005 | 100 | — | (100/110) × 100 = 90,90 |
| 2006 | 103 | — | (103/110) × 100 = 93,63 |
| 2007 | 107 | — | (107/110) × 100 = 97,27 |
| 2008 | 104 | — | (104/110) × 100 = 94,54 |
| 2009 | 102 | — | (102/110) × 100 = 92,72 |
| 2010 | 110 | 100 | 100,00 |
| 2011 | | 104 | 104,00 |
| 2012 | | 110 | 110,00 |
| 2013 | | 113 | 113,00 |
| 2014 | | 107 | 107,00 |
| 2015 | | 115 | 115,00 |

La forma de calcular la nueva serie es realizar un cambio de base en los índices anteriores al año del cambio de productos y/o servicios ofrecidos. Se interpreta igual que en cualquier otro cálculo de número índice.

## 8.8. DEFLACTACIÓN DE SERIES TEMPORALES

La deflactación es un concepto muy ligado al ámbito económico-financiero, aunque puede resultar de mucha utilidad en cualquier tipo de actividad económica que se desarrolle por períodos más o menos largos de tiempo. Se trata de determinar las variaciones monetarias o fluctuaciones del poder adquisitivo de las personas o entidades; otra definición puede ser la eliminación en el estudio de una serie de valores del efecto de la subida de precios.

Para ello debemos diferenciar dos conceptos:

— El **valor nominal** (moneda corriente), que es el valor del momento, independientemente de las fluctuaciones.

— El **valor real** (moneda constante), que es la cantidad de poder adquisitivo que tenemos una vez eliminada la influencia de la depreciación de la moneda.

Para poder deflactar una serie debemos calcular sus índices utilizando una referencia adecuada (índice de precios adecuado), denominado deflector. Su formulación es muy sencilla: se trata de un cociente entre la serie original o valor nominal dividido entre el deflactor, que habitualmente es el IPC (índice de precios al consumo):

$$VR = \frac{VN}{IPC}100$$

Si efectuamos nuestros cálculos en números índice, la expresión es:

$$IVR = \frac{IVN}{IPC}100$$

Es evidente que tanto el IPC como el IVN deben estar calculados en el mismo año base para poder hallar el IVR (índice de variación real).

Tal como hemos dicho en la primera definición, podemos utilizar para la deflactación cualquier tipo de índice, en cuyo caso





podemos encontrar bibliografía en la que la expresión de la fórmula tenga la siguiente nomenclatura:

$$\frac{Vt}{Lp}$$

En este caso hemos utilizado el índice de Laspeyres, pero podríamos utilizar el de Paasche o cualquier otro.

EJEMPLO. Queremos saber el valor en moneda constante de los ingresos obtenidos por una persona a lo largo de una serie de años; se trata de determinar cómo ha fluctuado su poder adquisitivo. Para ello contamos con los ingresos medios mensuales de los diferentes años de estudio y el valor del IPC en los mismos años*:

| Año* | VN* | IPC* | IPC (2000) | VR | IVN | IVR |
|---|---|---|---|---|---|---|
| 2009 | 1.265 | 135 | 100,0 | 1.265,00 | **100,00** | **100,00** |
| 2010 | 1.290 | 140 | 103,7 | 1.243,97 | **101,97** | **98,33** |
| 2011 | 1.312 | 148 | 109,6 | 1.197,08 | **103,71** | **94,62** |
| 2012 | 1.398 | 152 | 112,6 | 1.241,56 | **110,51** | **98,14** |
| 2013 | 1.450 | 156 | 115,6 | 1.254,32 | **114,62** | **99,15** |
| 2014 | 1.490 | 159 | 117,8 | 1.264,85 | **118,25** | **100,38** |
| 2015 | 1.620 | 163 | 120,7 | 1.342,17 | **128,06** | **106,09** |

### Paso 1

El primer paso es cambiar de base el IPC al año 2009. Para ello hacemos la transformación siguiente:

$$\frac{IPC}{IPC_{base}}100$$

Así elaboramos una serie nueva del IPC. Llevamos a cabo este paso debido a que la serie formada por el IPC tiene una referencia

anterior al período de estudio; si comenzara con el 100 no haría falta.

**Paso 2**

Calcular el valor real (VR) en función del nuevo IPC para el año 2000, con ello se cumple la premisa de que tanto el IPC —deflactor— y el VR están en la misma base. Operamos:

$$\frac{VN}{IPC_{\text{serie de la nueva base}}}100 = \frac{1.290}{103{,}7}100 = 1.243{,}97$$

Y hacemos así sucesivamente, hasta completar la serie.

**Paso 3**

El índice del valor nominal (*IVN*) se obtiene como un índice simple:

$$\frac{VN}{VN_{base}}100 = \frac{1.290}{1.265}100$$

de:

$$\frac{VN}{VN_{base}}100 = \frac{1.290}{1.265}100 = 101{,}97;$$

$$\frac{1.312}{1.265}100 = 103{,}71 \quad \ldots \quad \frac{1.620}{1.265}100 = 128{,}06$$

Y hacemos así sucesivamente, hasta completar la serie.

**Paso 4**

La columna IVR se obtiene como índice simple de:

$$\frac{IVN}{IPC_{2009}}100$$

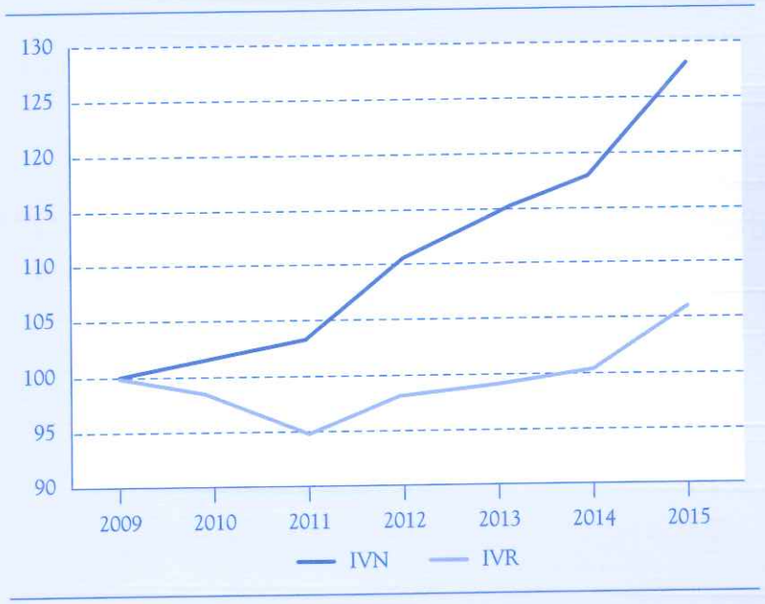




Para interpretar los resultados debemos fijarnos en las columnas de los índices de valor nominal y real (trama). Podemos ver cómo el IVN (salario) crece anualmente, pero al hacerlo por debajo del crecimiento de los precios el IVR (salario real de la persona) ha ido disminuyendo poco a poco hasta el año 2013, en que experimenta una recuperación, aunque por debajo del 100 %, siendo ya en los dos últimos años donde su salario ha crecido por encima de los precios. Se puede decir que este cálculo muestra cómo ganando más cada año tenemos menos poder adquisitivo.

En la gráfica podemos ver la comparativa de ambos números índices.

**FIGURA 8.2**



# Referencias bibliográficas

Álvarez, S. J. (2011). *Estadística aplicada. Teoría y problemas*. Madrid: CLAG.
Álvaro, J. L. y Garrido, A. (1995). *Análisis de datos con SPSS/PC+*. Cuadernos Metodológicos, 14. Madrid: CIS (Centro de Investigaciones Sociológicas).
Amón, J. (1980). *Estadística para psicólogos: 1 Estadística descriptiva*. Madrid: Pirámide.
Arnau, J. (1996). Técnicas de análisis avanzadas y diseño de investigación: tendencias actuales y líneas futuras de desarrollo. En J. Arnau (ed.): *Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento*. Barcelona: EUB.
Cao, R. et al. (2001). *Introducción a la estadística y sus aplicaciones*. Madrid: Pirámide.
Carrasco, S. (2005). *Aproximación a la estadística desde las ciencias sociales*. Valencia: Universitat de Valencia.
De Burgos, J. (2011). *Estadística: definiciones, teoremas y resultados*. Madrid: García-Maroto Editores.
Escuder, R. (1987). *Métodos estadísticos aplicados a la economía*. Barcelona: Ariel.
Ferrán, M. (1999). *SPSS para Windows, programación y análisis estadístico*. Madrid: McGraw-Hill.
Glass, G. y Stanley, J. (1980). *Métodos estadísticos aplicados a las ciencias sociales*. Madrid: Prentice Hall.
Kendall, M. G. (1968). *«Francis Ysidro Edgeworth, 1845-1926», Studies in the History Probability*. Londres: Charles Griffin.
Martín-Guzmán, P. y Martín Pliego, F. J. (2012). *Curso básico de estadística*. Madrid: AC.
Martín Pliego, F. J. (2012). *Introducción a la estadística económica y empresarial*. Madrid: AC.
Mood, A. y Gaybill, F. A. (1978). *Introducción a la teoría de la estadística*. Madrid: Aguilar.
Osvaldo, P. y Fernández de la R., P. (1988). La estadística, una ciencia en la controversia. *Revista Universitaria, 25*.