# Introducción a la Estadística Económica

**Book** · April 2012

CITATIONS
7

READS
29,975

**4 authors:**

Rigoberto Perez Suarez
University of Oviedo
**108** PUBLICATIONS **820** CITATIONS
SEE PROFILE

Covadonga Caso
University of Oviedo
**22** PUBLICATIONS **124** CITATIONS
SEE PROFILE

Rio Maria Jesus
University of Oviedo
**8** PUBLICATIONS **70** CITATIONS
SEE PROFILE

Ana Jesus Lopez-Menendez
University of Oviedo
**161** PUBLICATIONS **2,185** CITATIONS
SEE PROFILE

# Introducción a la Estadística Económica


Rigoberto Pérez, Covadonga Caso, María Jesús Río y Ana Jesús López
*rigo@uniovi.es,ccaso@uniovi.es,mjrio@uniovi.es,anaj@uniovi.es,*
Dpto. de Economía Aplicada, Campus del Cristo. Universidad de Oviedo
https://sites.google.com/a/uniovi.es/libros/iee

Abril 2012

*A nuestras familias*

ISBN13 978-84-693-9868-5
Depósito Legal: AS-6241-2010
Edición 2010
Revisión V.1.0.2

## Breve reseña de autores

Los autores de libro son profesores del Departamento de Economía Aplicada de la Universidad de Oviedo (Unidad de Estadística y Econometría).

**Rigoberto Pérez Suárez** es Catedrático de Universidad y su amplia experiencia docente incluye asignaturas de Estadística Econometría y Series temporales tanto en primer y segundo ciclo como en doctorados y másteres. Es autor de varios libros de texto (Nociones Básicas de Estadística, Análisis de datos económicos I: Métodos descriptivos, Análisis de datos económicos II: Métodos inferenciales) y del software docente ADE+, así como de numerosas publicaciones relativas a la innovación educativa y el e-learning.
También ha sido Director de Area de Innovación de la Universidad de Oviedo (2000-2006) y Director del Campus Virtual Compartido del grupo G9 (2004-2006).
En el ámbito investigador es autor de diversas publicaciones en revistas de impacto y ha dirigido numerosas tesis doctorales y proyectos de investigación, generalmente referidos a la predicción económica y al análisis de la desigualdad.

**Covadonga Caso Pardo** es Profesora Titular de Universidad y su docencia está centrada en asignaturas de Estadística de licenciaturas y grados, y en cursos de postgrado de Análisis Multivariante. Es una de las autoras del manual Análisis de datos económicos I: Métodos descriptivos.

**María Jesús Río Fernández** es Profesora Titular de Escuela Universitaria y su experiencia docente incluye diversas asignaturas de Estadística en primer y segundo ciclo. Es autora del manual Análisis de datos económicos I: Métodos descriptivos.

**Ana Jesús López Menéndez** es Profesora Titular de Universidad y su docencia abarca asignaturas de Estadística, Econometría y Series temporales. Es autora de los manuales Análisis de datos económicos I: Métodos descriptivos y Análisis de datos económicos II: Métodos inferenciales, así como de numerosas publicaciones relativas a la innovación educativa y el e-learning.
En el ámbito investigador es autora de diversos artículos publicados en revistas de impacto, ha dirigido seis tesis doctorales y ha participado en numerosos proyectos de investigación.

# Índice general

## Presentación

La elaboración de un manual universitario es una experiencia de gran interés e intensidad que conlleva un proceso previo de reflexión sobre el papel de la asignatura, sus contenidos y la metodología docente.

En nuestro caso esta experiencia fue abordada hace ya casi veinte años por un conjunto de profesores ante la puesta en marcha en la Universidad de Oviedo de las licenciaturas en Economía y Administración y Dirección de Empresa. En aquella ocasión, afirmábamos que nuestro manual "Análisis de datos económicos" pretendía aproximar nuestros programas a las necesidades reales de nuestros alumnos, tratando de compatibilizar la exposición amena e intuitiva de problemas con un tratamiento serio de los contenidos. Con este objetivo, el texto incluía varias novedades como la utilización de tres niveles diferenciados de lectura, la incorporación de numerosas ilustraciones y un disquette con ejemplos resueltos con hoja de cálculo.

Transcurrido el tiempo y agotadas varias ediciones de aquel manual, en la actualidad nos situamos en un nuevo contexto, caracterizado por la puesta en marcha de los nuevos grados universitarios adaptados al Espacio Europeo de Educación Superior, en los que se contempla un papel más activo del estudiante, tal y como refleja la definición del crédito europeo ECTS, que "computa el número de horas de trabajo requeridas para la adquisición por los estudiantes de los conocimientos, capacidades y destrezas correspondientes", por lo que en su asignación "deberán estar comprendidas las horas correspondientes a las clases lectivas, teóricas o prácticas, las horas de estudio, las dedicadas a la realización de seminarios, trabajos, prácticas o proyectos, y las exigidas para la preparación y realización de los exámenes y pruebas de evaluación".

Por otra parte, la constante evolución de las Tecnologías de la Información y la Comunicación (TIC) abre nuevas posibilidades para la generación y transmisión del Conocimiento. De ahí, que en este libro, por una parte, hayamos cambiado el formato impreso por el digital y por otra, nos hayamos centrado en los contenidos docentes, que serán complementados con materiales online, tanto de acceso libre (en la web del libro) como restringidos a los estudiantes de nuestras asignaturas (accesibles en el campus virtual de la Universidad de Oviedo, http://www.campusvirtual.uniovi.es)

Con este planteamiento, presentamos "Introducción a la Estadística Económica", texto que se adapta a la asignatura del mismo nombre incluida en el primer curso de los grados de Economía, Administración y Dirección de Empresas, Contabilidad y Finanzas y Relaciones Laborales y Recursos Humanos de la Universidad de Oviedo. A lo largo de doce temas presentamos de forma sencilla pero con rigor los conceptos y resultados relativos a los principales métodos estadísticos descriptivos.

En los temas iniciales se analiza el origen, organización y resumen de la información, presentando las principales fuentes de información económica y su representación mediante tablas y gráficos (tema 1), así como las principales medidas de posición (tema 2), dispersión y forma (tema 3) y desigualdad y pobreza (tema 4).

A continuación se aborda el análisis conjunto de variables, estudiando las principales medidas de correlación y asociación (tema 5) y las técnicas de regresión lineal tanto simple (tema 6) como múltiple (tema 7).

Los contenidos más específicos de estadística económica incluyen los números índices y tasas (tema 8), los principales índices económicos de precios, cantidades y valor (tema 9) y el Indice de Precios de Consumo (IPC) con sus principales aplicaciones (tema 10).

Por último, los temas finales estudian la evolución temporal de las magnitudes económicas. El tema 11 describe las series temporales y los métodos de aproximación de su tendencia, mientras el tema 12 analiza la estacionalidad y la elaboración de predicciones a partir de modelos temporales.

Como ya hemos anticipado, este libro se publica en formato PDF y está disponible en la Red para que cualquier persona pueda descargarlo de forma libre y gratuita. La última versión de este libro y material complementario se encuentra en:

https://sites.google.com/a/uniovi.es/libros/iee

Confiamos en que este material pueda resultar de utilidad y agradecemos de antemano vuestros comentarios y sugerencias.

# 1 Organización y presentación de los datos. Fuentes estadísticas

El análisis estadístico de cualquier problema económico requiere en una primera etapa determinar los objetivos y el colectivo informante. A modo de ilustración, supongamos que se desea hacer un estudio estadístico sobre el sector sanitario en España. ¿Quién suministrará la información que necesitamos? Los hospitales o centros de salud, el personal sanitario o la población española en general. Las conclusiones que se obtengan, ¿a qué colectivo afectarán? La información necesaria para alcanzar los objetivos del estudio posiblemente se transmitirá por un cuestionario o acudiremos a bases de datos ya elaboradas por algún organismo, pero ¿la recogeremos de forma cualitativa o cuantitativa? Una vez recabada la información y como fase previa a la aplicación de las técnicas estadísticas pertinentes se procederá a la organización de los datos y se presentarán una serie de tablas y gráficos con un resumen de la información disponible.

A lo largo de este tema se introducirán los conceptos básicos vinculados a esta fase inicial de un estudio estadístico. Asimismo se ofrecerá una panorámica de los principales organismos y fuentes que proporcionan información estadística.

## 1.1. Origen de la información: censos y muestras

**Definición 1.1.** Se denomina *población* o *universo* al conjunto de personas o cosas a las que va referida una investigación estadística.

Desde el punto de vista estadístico, el término población puede aludir tanto a personas como a hogares, hospitales o empresas. Cada una de las personas o cosas que integran la población recibe el nombre de *elemento* y el número total de elementos que la integran se denomina *tamaño* poblacional.

La recogida de información se realiza, generalmente, por medio de cuestionarios, siendo el entrevistado o informador una especie de "socio anónimo" de todo el proceso estadístico. En su sentido más amplio, entendemos por *encuesta* el procedimiento global que se sigue para la recogida de información. Su extensión, es decir, el conjunto de elementos de la población a los que se solicita información unidades informantes da lugar a dos tipos de encuestas: censales y muestrales.

**Definición 1.2.** Una *encuesta censal* o *censo* es aquélla que se realiza a todos los componentes de la población.

Los distintos países llevan a cabo periódicamente recuentos exhaustivos de sus habitantes, viviendas, explotaciones agrarias, ..., conocidos como Censos de Población, de

Viviendas, Agrario, ... . Históricamente, los recuentos de población son el primer tipo de estadística del que se tiene noticia. Los gobiernos de las civilizaciones antiguas ya realizaban este tipo de recuentos con el fin de recaudar tributos y de reclutar hombres aptos para la guerra.

El análisis exhaustivo de poblaciones no es la forma más habitual de desarrollar encuestas. A pesar de que los avances informáticos permiten procesar volúmenes de información que hace unos años resultaban impensables, hay dos razones fundamentales para ello:

1. La necesidad de limitar recursos -motivada por los elevados costes de los censos-
2. La rapidez en la obtención de resultados.

Estos argumentos conducen a plantear estudios parciales, llevando a cabo posteriormente una generalización de los resultados obtenidos. En este contexto surgen los conceptos de subpoblación y muestra.

> El hecho de trabajar con encuestas censales no garantiza la ausencia de errores en los resultados, pues siempre pueden aparecer errores vinculados al proceso de observación: preguntas confusas, errores de memoria por parte del entrevistado, negativas a responder, etc.

**Definición 1.3.** Una *subpoblación* es una parte de la población integrada por un conjunto de elementos que presentan alguna característica común.

Los centros sanitarios de titularidad pública o los hospitales ubicados en la Comunidad de Madrid son ejemplos de subpoblaciones en un estudio sobre el sector sanitario en España. ¿Pueden generalizarse a toda la población los resultados obtenidos a partir de la información proporcionada por los elementos de una subpoblación? En principio la respuesta es negativa pues sólo hay garantías de que representen a la subpoblación en cuestión y no a todo el colectivo. La alternativa será considerar estudios basados en muestras.

**Definición 1.4.** Una *muestra* es una parte de la población cuyos elementos se eligen de modo que sean representativos de todo el colectivo. Las encuestas basadas en muestras se denominan *encuestas muestrales*.

El concepto de muestra abre algunos interrogantes importantes: ¿qué significa representativa?, ¿cómo garantizar que una muestra sea representativa? Una muestra será representativa cuando constituya una réplica a escala de la población. ¿Cómo podríamos definir nosotros la réplica? Una muestra de hospitales debería tener el mismo porcentaje que la población de centros públicos y privados, el mismo porcentaje por provincias, por número de empleados, por gastos, ... . En realidad, serán muchas las características a tener en cuenta para que la muestra pueda ser calificada como una réplica de la población. La estadística proporciona métodos para la selección de muestras, en su mayor parte basados en la elección de sus elementos al azar, lo cual garantizará la imparcialidad en el proceso de selección.

Las encuestas muestrales presentan ciertas ventajas frente a las censales. Por una parte, hay que notar el ahorro considerable tanto monetario como de tiempo que puede suponer el tener que entrevistar sólo a unos pocos individuos de una población numerosa. Pero además, el hecho mismo de poder trabajar con muestras de tamaño relativamente pequeño permite, a su vez, afinar en la calidad de los datos y, en consecuencia, controlar la fiabilidad de los resultados. Obviamente, no todo son ventajas, pues el hecho de trabajar con información parcial, proporcionada por una pequeña parte de la población, puede generar errores en los resultados, cuya magnitud estará estrechamente relacionada con la representatividad de la muestra.

Ambos tipos de encuesta -censal y muestral- deben convivir en una especie de simbiosis, complementándose mutuamente. En algunos casos es conveniente la utilización de muestras, en otros resulta imprescindible, y en cualquiera de ellos el censo correspondiente proporciona el marco de referencia.

## 1.2. Tipos de información estadística

Uno de los aspectos importantes en el diseño de una encuesta es la elaboración de un *cuestionario*, mediante el cual se recogerá la información necesaria sobre los rasgos o caracteres de interés para el estudio, que pueden ser tanto cuantitativos como cualitativos.

**Definición 1.5.** Los caracteres cuantitativos, expresados mediante números, reciben el nombre de *variables* y se representan habitualmente mediante mayúsculas $X$, $Y$, ... Los resultados de la observación de una variable se denominan *valores* y se designan por las correspondientes letras minúsculas $x_1, x_2, \ldots$; $y_1, y_2, \ldots$ Dependiendo de los valores que puedan presentar se distinguen a su vez dos tipos de variables:

- *Discretas*: Variables que sólo pueden tomar cierto número de valores aislados o, de forma equivalente, si el número de valores diferentes que pueden asumir es finito o infinito numerable.
- *Continuas*: Variables que pueden tomar cualquiera de los infinitos valores de uno o varios intervalos de la recta real.

El número de asignaturas matriculadas en un grado o el número de empleados de una empresa son ejemplos de variables discretas que pueden tomar valores 1,2,3, ... La altura de los estudiantes, el tiempo diario de estudio o el coste de las materias primas en una industria son ejemplos de variables continuas.

Cuando observamos en concreto el valor de una variable continua anotaremos una serie de valores aislados; por ejemplo, la altura será 155, 165 o 180 cm, es decir, su cuantificación tendrá una precisión limitada, determinada por la unidad de medida que pueda captar el observador según el instrumento utilizado. Con ello queremos expresar que, desde un punto de vista empírico, las variables presentan un comportamiento discreto, para señalar también inmediatamente que la distinción entre variables continuas y discretas es muy importante desde la perspectiva teórica, esto es, el concepto de continuidad garantiza el paso al límite y como consecuencia permite aplicar una potente

metodología matemática -el cálculo diferencial e integral -. Por tanto, es conveniente distinguir el carácter continuo o discreto de una variable porque condicionará el modelo teórico a aplicar para su estudio y para ello atenderemos a su naturaleza y no a los resultados de su observación empírica.

**Definición 1.6.** Los caracteres cualitativos, expresados mediante palabras, reciben el nombre de *atributos*. Los resultados de la observación de un atributo se denominan *modalidades o categorías*.

Ejemplos de atributos son los estudios de grado realizados (con categorías economía, derecho, medicina, ...), el sexo, el estado civil, la nacionalidad, ...; el sector de actividad, el municipio de ubicación de una empresa, ...

**Definición 1.7.** En general, denominamos serie estadística, o sencillamente *estadística*, a la información o colección de datos disponible. Estas series pueden ser clasificadas en diferentes categorías que pasamos a examinar a continuación.

- Según el número de caracteres estudiados, se distingue entre:
  - *Estadísticas univariantes*: son aquéllas que se obtienen cuando se estudia un único carácter.
  - *Estadísticas multivariantes*: analizan de forma conjunta varios caracteres, opción que resulta adecuada cuando puede existir alguna relación en su comportamiento.
- Según la óptica del estudio se distingue entre:
  - *Estadísticas temporales* o de corte longitudinal, cuando se toma el tiempo como referencia y se analiza la evolución temporal de una o varias variables.
  - *Estadísticas de corte transversal*, que aparecen cuando se abandona la óptica temporal y el estudio se efectúa sobre distintos individuos o unidades espaciales en un momento del tiempo concreto.
  - *Datos de panel*, que se corresponden con situaciones en las que se dispone de datos que combinan ambas perspectivas, longitudinal y transversal.

## 1.3. Presentación de datos

### 1.3.1. Tabulación

Una vez recogida la información, debemos preocuparnos de su presentación, procurando que ésta sea útil y manejable a efectos de su análisis estadístico. El proceso de ordenación y agrupación de los datos se denomina *tabulación*, y su resultado será una tabla estadística.

En este tema se presentarán las tablas correspondientes a estadísticas univariantes, posponiendo a los temas específicos la presentación de las tablas para estadísticas multivariantes.

## Tablas estadísticas univariantes (datos no agrupados)

Sea $X$ una variable que puede tomar $k$ valores diferentes designados por $x_1, x_2, \ldots, x_k$, que se asumen ordenados en sentido creciente y para la que se dispone de un total de $N$ observaciones.

**Definición 1.8.** Se definen los siguientes tipos de frecuencias asociados a cada valor $x_i$ $(i = 1, \ldots, k)$ de una variable $X$:

1. *Frecuencia absoluta* $n_i$: número de observaciones en las que se presenta el valor $x_i$.
2. *Frecuencia relativa* $f_i$: proporción de observaciones en las que se presenta el valor $x_i$; se obtiene como cociente entre su frecuencia absoluta y el número total de datos $f_i = \frac{n_i}{N}$. Se expresa habitualmente en términos porcentuales $(f_i \times 100)$.
3. *Frecuencia absoluta acumulada* $N_i$: número de observaciones menores o iguales que $x_i$; se obtiene como $N_i = n_1 + \cdots + n_i = \sum_{j=1}^{i} n_j$.
4. *Frecuencia relativa acumulada* $F_i$: proporción de observaciones menores o iguales que $x_i$; se obtiene como cociente entre su frecuencia absoluta acumulada y el número total de datos $F_i = \frac{N_i}{N}$. Se expresa habitualmente en términos porcentuales $(F_i \times 100)$.

El conjunto de los diferentes valores asumidos por una variable junto con cualquiera de las frecuencias correspondientes se denomina *distribución de frecuencias* y, genéricamente, se representa por $(x_i, n_i)$ o $(x_i, f_i)$. Suele representarse mediante tablas del tipo siguiente:

| $x_i$ | $n_i$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $n_k$ |

| $x_i$ | $f_i$ |
|---|---|
| $x_1$ | $f_1$ |
| $x_2$ | $f_2$ |
| $\vdots$ | $\vdots$ |
| $x_k$ | $f_k$ |

Las definiciones anteriores, a excepción de las frecuencias acumuladas, son aplicables también para el caso de caracteres cualitativos.

**Propiedad 1.1.** *Propiedades de las frecuencias*

*a)* $0 \leq n_i \leq N$; $\sum_{i=1}^{k} n_i = N$

*b)* $0 \leq f_i \leq 1$; $\sum_{i=1}^{k} f_i = 1$

*c)* $0 \leq N_i \leq N$; $N_k = N$.

*Fórmula de recurrencia:* $N_1 = n_1$, $N_i = N_{i-1} + n_i$, $i = 2, \ldots, k$

*d)* $0 \leq F_i \leq 1$; $F_k = 1$; $F_i = \sum_{j=1}^{i} f_j$

*Demostración.* La mayor parte de las propiedades son consecuencia inmediata de la propia definición de las frecuencias. Nos centraremos en la demostración de las propiedades b) y d).

b) Dado que por la propiedad a) se tiene que $0 \leq n_i \leq N$ y teniendo en cuenta que $f_i = \frac{n_i}{N}$, se deduce que $0 \leq f_i \leq 1$. Por otra parte, para demostrar que la suma de las frecuencias relativas es siempre la unidad (o 100 si se expresan en términos porcentuales), basta considerar la definición de frecuencia relativa y sacar factor común el denominador $N$ en el operador suma:

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

d) La primera parte es consecuencia inmediata a partir de la propiedad c). Por otra parte, las frecuencias relativas acumuladas pueden obtenerse mediante sumas acumuladas de frecuencias relativas, ya que:

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_j$$

□

## Tablas estadísticas univariantes (datos agrupados en intervalos)

En los estudios empíricos se dispone habitualmente de un número de observaciones elevado, para las que las variables estudiadas pueden presentar muchos valores diferentes. En otras ocasiones, a lo anterior debe añadirse que la variable puede ser clasificada como continua. Estas dos razones, conjuntamente o por separado, dan lugar a que las tablas estadísticas que manejamos puedan ser de gran tamaño y, por consiguiente, poco manejables. En estos casos es habitual clasificar los datos en intervalos o clases.

Supongamos que los valores de la variable $X$ están agrupados en $k$ intervalos que denotamos por $L_{i-1} - L_i$, donde $L_{i-1}$ es el extremo inferior de cada intervalo y $L_i$ el extremo superior ($i = 1, \ldots, k$). La frecuencia absoluta $n_i$, asociada al intervalo i-ésimo ($i = 1, \ldots, k$), se obtendrá como suma de las frecuencias correspondientes a los valores pertenecientes a dicho intervalo. Se obtienen así *tablas de datos agrupados* en intervalos del tipo siguiente:

| $L_{i-1} - L_i$ | $n_i$ |
|---|---|
| $L_0 - L_1$ | $n_1$ |
| $L_1 - L_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $L_{k-1} - L_k$ | $n_k$ |

La *amplitud* de un intervalo se denota por $a_i$ y viene dada por la diferencia entre los valores extremos de dicho intervalo: $a_i = L_i - L_{i-1}$.

La *marca de clase* $x_i$ es un valor que representa al intervalo. Puesto que en las tablas agrupadas se desconocen los valores que se presentan en cada intervalo, suele asumirse que los valores se reparten de modo uniforme dentro del mismo y, por lo tanto, quedarán bien representados por el valor situado en el centro, lo que conduce a tomar como marca de clase el punto medio del intervalo: $x_i = \frac{L_{i-1}+L_i}{2}$.

El agrupamiento de datos presenta algunos puntos de discusión acerca de los que no hay criterios unánimes; entre ellos destacaremos los referentes a:

- Número de intervalos.- La determinación del número de intervalos suele efectuarse intentando buscar un equilibrio entre la pérdida de información que se deriva de la agrupación y la operatividad. Así, la consideración de muchos intervalos presenta la ventaja de respetar la información inicial, pero en cambio no simplifica el estudio. Por el contrario, si se opta por agrupar los datos en pocos intervalos la ventaja sería la síntesis y operatividad conseguida pero llevaría asociado el inconveniente de una pérdida excesiva de información.

- Amplitud de los intervalos.- La amplitud puede ser constante para todos los intervalos, lo cual simplifica el tratamiento de los datos, o bien variable según el recorrido, opción que permite una mejor adecuación a las características de la variable en estudio.

- Extremos que se incluyen en cada intervalo.- Habitualmente se consideran intervalos contiguos y pueden presentarse observaciones coincidentes con los extremos de los intervalos, por lo que es necesario establecer si los intervalos incluyen el extremo inferior o el superior, es decir, si son semiabiertos del tipo $[L_{i-1}, L_i)$ o $(L_{i-1}, L_i]$. Por otra parte, los intervalos extremos pueden ser no acotados del tipo "Menos de 150 cm" o "Más de 2 metros".

## Tablas temporales

Sea $Y$ una variable que se observa a lo largo de distintos periodos de tiempo $t$ (años, meses, etc.), siendo $Y_t$ el valor observado en el periodo $t$. La descripción numérica de una variable de este tipo puede realizarse a través de una tabla con dos columnas, una para el tiempo ($t$) y otra para las observaciones ($Y_t$). A continuación se muestra una tabla temporal con datos (en tantos por mil) de la tasa de natalidad en España en el periodo 2000-2009:

| $t$ | $Y_t$ |
|---|---|
| 2000 | 9,85 |
| 2001 | 9,95 |
| 2002 | 10,11 |
| 2003 | 10,49 |
| 2004 | 10,61 |
| 2005 | 10,71 |
| 2006 | 10,92 |
| 2007 | 10,94 |
| 2008 | 11,37 |
| 2009 | 10,73 |

### 1.3.2. Representaciones gráficas

En un sentido amplio, entendemos por *representación gráfica* de una serie estadística cualquier tipo de dibujo que nos permita detectar a primera vista algunas de sus características más notables, esto es, que nos ofrezca una visión general del fenómeno en estudio. La representación gráfica es un instrumento que ayuda a resumir o desglosar la información que se encuentra contenida en la tabla estadística y al mismo tiempo puede descubrir una parte de esa información que esté oculta en la representación numérica.

Aquí estudiaremos algunos de los gráficos usuales para estadísticas univariantes, que serán complementados en temas posteriores con las representaciones gráficas asociadas a estadísticas multivariantes.

#### Gráficos para información cualitativa

- *Diagrama de sectores*. El esquema básico de esta representación consiste en dividir un círculo en tantos sectores como modalidades tenga el atributo, de manera que el área de cada sector sea proporcional a la frecuencia de la modalidad que representa. El diagrama de sectores de la figura 1.1 refleja la distribución por sectores de actividad de la población ocupada en una región.

- *Diagrama de rectángulos*. Sobre un par de ejes cartesianos se trazan tantos rectángulos como modalidades tenga el atributo, todos con idéntica base, situada en el eje de abscisas, y con altura proporcional a la frecuencia de la modalidad correspondiente. [Figura 1.2]

#### Gráficos para información cuantitativa

- *Diagrama de barras*. Es la representación gráfica de la distribución de frecuencias (absolutas o relativas) de una tabla de datos no agrupados. En un plano de coordenadas, se representan en el eje de abscisas los distintos valores de la

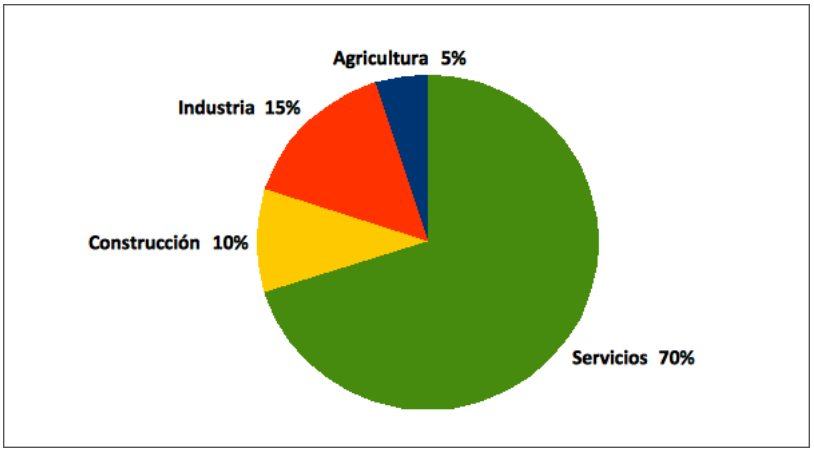



Figura 1.1: Diagrama de sectores. Población ocupada

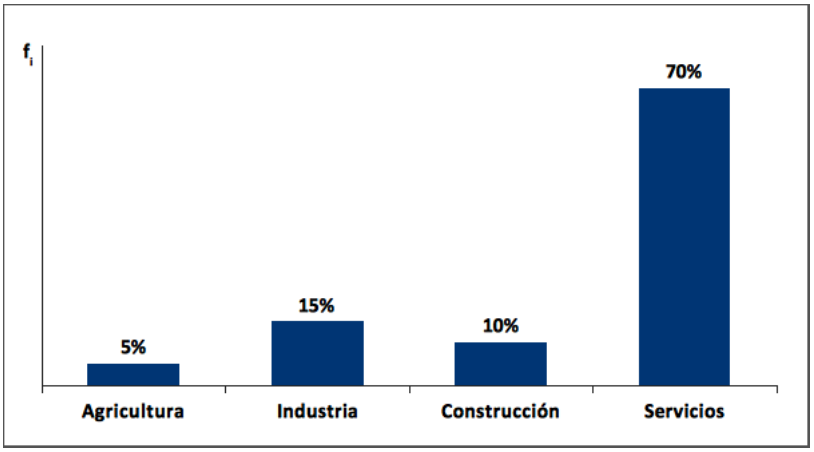



Figura 1.2: Diagrama de rectángulos. Población ocupada

variable y en el eje de ordenadas las frecuencias correspondientes, obteniéndose los puntos $(x_i, n_i)$ o $(x_i, f_i)$; para realzar la representación se traza el segmento vertical que une cada punto con su abscisa. De esta manera el dibujo consiste en una serie de barras verticales cuya altura refleja la importancia del valor al que están asociadas. [Figura 1.3]

- *Diagrama en escalera*. Es la representación gráfica de la distribución de frecuencias acumuladas (absolutas o relativas) de una tabla de datos no agrupados. En un plano de coordenadas se asigna a cada observación $x_i$ una altura igual a su frecuencia acumulada $N_i$, punto que se une mediante un trazo horizontal a la ordenada del valor siguiente. El gráfico se completa asignando el valor 0 hasta llegar al primer valor de la variable $(x_1)$ y el valor $N$ (o 1 en el caso de

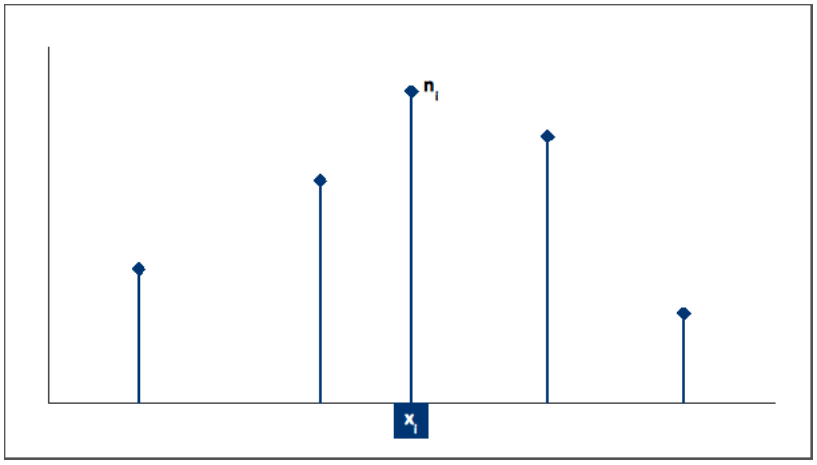



Figura 1.3: Gráfico de barras

frecuencias relativas acumuladas) a partir del último valor ($x_k$). Se obtiene así la representación gráfica de una función que asigna a cada número real su frecuencia acumulada, con discontinuidades en cada uno de los $k$ valores diferentes observados de la variable $x_i$, siendo la altura de cada salto coincidente con su frecuencia absoluta $n_i$ (o relativa $f_i$). [Figura 1.4]

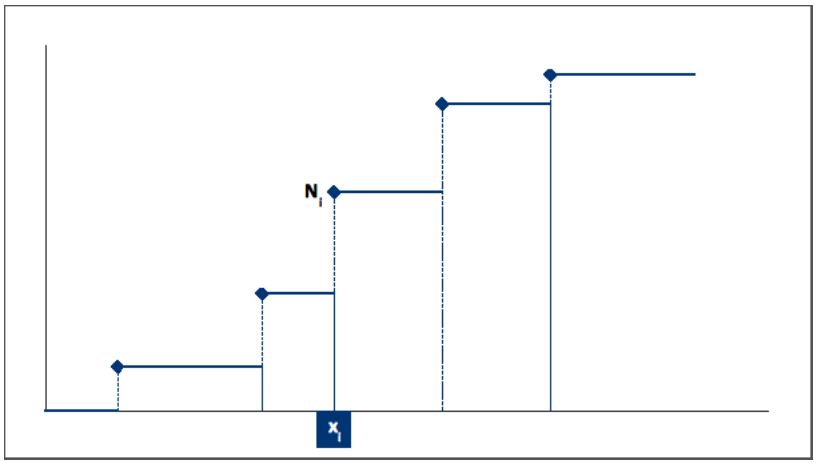



Figura 1.4: Diagrama en escalera

- *Histograma*. Es la representación gráfica de la distribución de frecuencias absolutas (o relativas) para tablas de datos agrupados en intervalos. Se obtiene construyendo sobre cada intervalo $L_{i-1} - L_i$, representado en el eje de abscisas, un rectángulo cuya base es igual a la amplitud del intervalo $a_i$ y cuya altura $h_i$

se determina de forma que el área del rectángulo sea proporcional a su frecuencia $n_i$, para lo cual bastará calcular la altura mediante la expresión: $h_i = \frac{n_i}{a_i}$ (o $h_i = \frac{f_i}{a_i}$ en el caso de frecuencias relativas). [Figura 1.5]

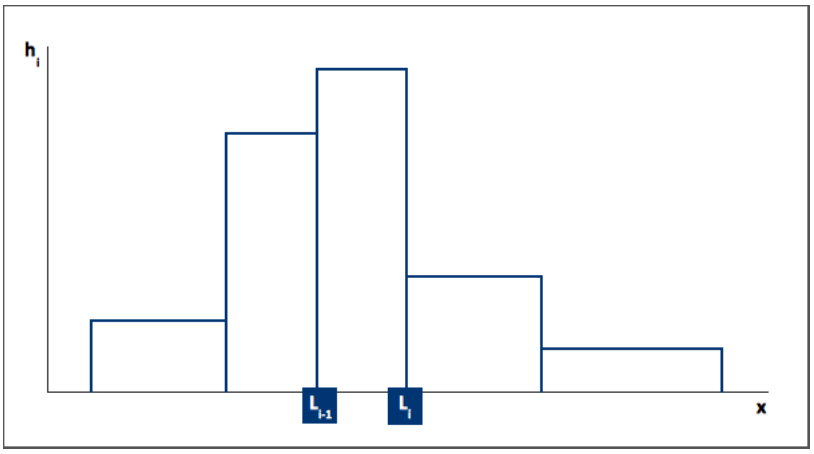



Figura 1.5: Histograma

Dado que el área de cada rectángulo coincide con la frecuencia de un intervalo, el área total del gráfico se identificará con el número total de datos $N$ (o será 1 si se representan las frecuencias relativas). Así, la forma del histograma nos indicará cómo se distribuyen las observaciones a lo largo de todo el recorrido de la variable:

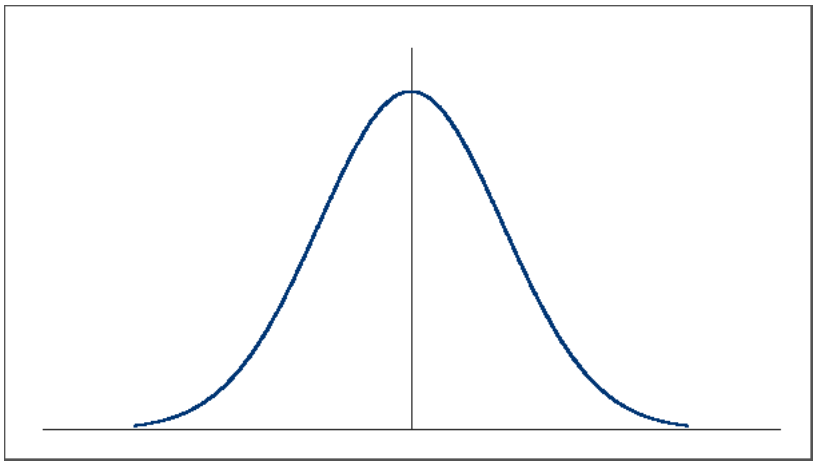


Figura 1.6: Curva normal

- Un histograma con forma de campana reflejará una situación en la que la mayor parte de los datos se concentran en la parte central, con un peso

relativo muy pequeño de los datos en los extremos y repartidos además de forma simétrica a ambos lados. Un gráfico de este tipo se identifica con la conocida *distribución normal o de Gauss*, que juega un papel central en los desarrollos de la Inferencia Estadística. [Figura 1.6]

- Un histograma con forma de U se identificará con situaciones en las que la parte central tiene poca importancia, mientras que la mayor parte de las observaciones se concentran en ambos extremos del recorrido.

- *Polígono de frecuencias acumuladas*. Es la representación gráfica de la distribución de frecuencias acumuladas en tablas de datos agrupados en intervalos. Este gráfico muestra cómo se van acumulando paulatinamente las observaciones, para lo cual se asocia al extremo superior de cada intervalo su frecuencia acumulada (absoluta o relativa) y se unen todos estos puntos mediante una línea poligonal, teniendo en cuenta además que la frecuencia acumulada correspondiente a cualquier valor anterior al extremo inferior del primer intervalo ($L_0$) es nula y que la correspondiente a valores superiores al extremo superior del último intervalo ($L_k$) es $N$ (o 1 si se trata de frecuencias relativas). [Figura 1.7]

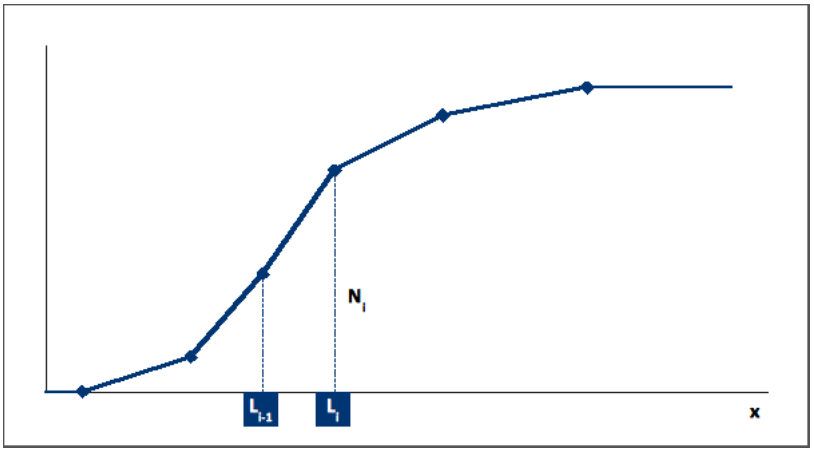



Figura 1.7: Polígono de frecuencias acumuladas

- *Gráfico temporal*. Para representar gráficamente una serie temporal utilizaremos un plano de coordenadas en el que a cada unidad temporal $t$ en el eje de abscisas se asigna una ordenada que se identifica con el valor de la variable observado en el periodo $t$, $Y_t$. Normalmente, al objeto de hacer más visible la evolución temporal de la variable, se unen los puntos $(t, Y_t)$. [Figura 1.8]

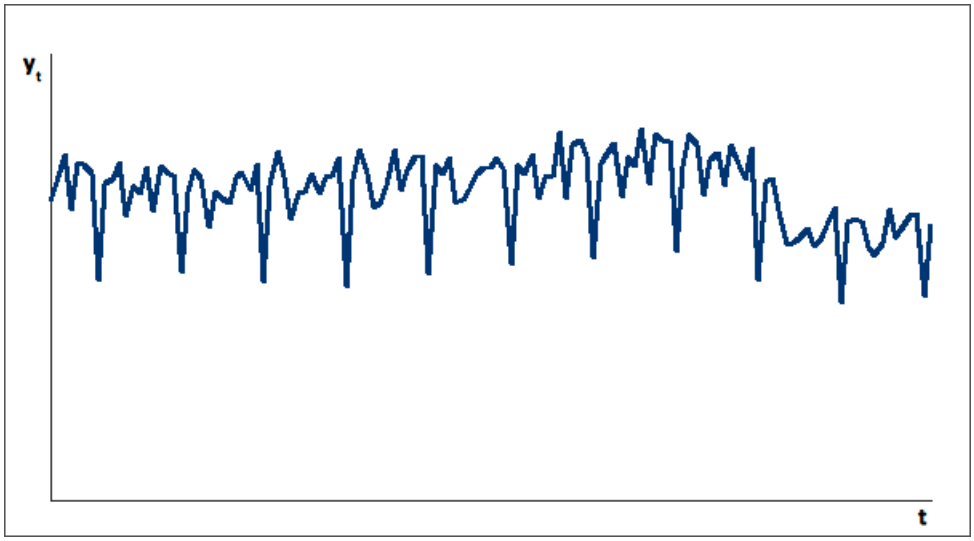



Figura 1.8: Serie temporal

## 1.4. Organismos y fuentes estadísticas de información económica

El Estado, como administrador de los intereses de los ciudadanos, precisa de información estadística para desarrollar sus funciones y planificar sus políticas en materia económica y social en general (etimológicamente el término "estadística" deriva de la palabra "estado"). El gobierno precisa de cifras estadísticas de subida de los precios para revisar las pensiones, negociar los salarios en convenios colectivos, ..., necesita disponer de datos estadísticos sobre natalidad para prever la dotación de plazas escolares, sobre incidencia de ciertas enfermedades para planificar las infraestructuras sanitarias, etc. En este apartado se ofrece una panorámica de los principales organismos oficiales dedicados a la elaboración y publicación de estadísticas, tanto en el ámbito nacional como internacional. Asimismo, se presenta un resumen de las principales características de algunas de las estadísticas de uso más generalizado en el ámbito económico.

### 1.4.1. Organización estadística oficial

#### Sistema estadístico nacional: el INE

Dentro de la organización de la Administración General del Estado y con el fin de cubrir sus propias necesidades de información para la toma de decisiones existen servicios dedicados a la producción de estadísticas, que en su conjunto constituyen el *Sistema Estadístico Nacional*. La actividad del Sistema Estadístico Nacional está regulada por una serie de normas legales cuyo punto de partida es la Constitución Española de 1978, que en el artículo 149.1.31 establece que "*la Estadística para fines*

*estatales es competencia exclusiva del Estado*". Sin embargo, esta competencia debe ser considerada desde una perspectiva más amplia a distintos niveles administrativos (Comunidades Autónomas y Ayuntamientos). Por otra parte, la Unión Europea también legisla en materia estadística a través de reglamentos, decisiones y directivas de cumplimiento obligatorio para los países miembros.

En la actualidad, el marco legal vigente es la *Ley de la Función Estadística Pública* (LFEP) de 9 de mayo de 1989. Entre los aspectos regulados por la LFEP se encuentran la recogida de datos, el secreto estadístico, la difusión y conservación de la información estadística y los Servicios Estadísticos del Estado.

> La garantía del secreto estadístico, regulada por la LFEP, resulta especialmente importante en la sociedad actual, en la que existe una preocupación permanente por salvaguardar los derechos fundamentales de los individuos y, en particular, el de intimidad en lo que concierne a información privada de las unidades informantes (personas, hogares, empresas, etc.).

La organización de la producción de estadísticas para fines estatales tiene como pilar básico el *Instituto Nacional de Estadística* (INE), creado en 1945 con el fin de ser la oficina central de estadística, y que en la actualidad es un organismo autónomo adscrito al Ministerio de Economía y Competitividad. En la LFEP de 1989 se describe todo el conjunto de funciones encomendadas al INE, y que pueden ser resumidas en las siguientes grandes líneas de actuación:

- Ser el principal productor de estadísticas para fines estatales
- Ocuparse de la coordinación y planificación del Sistema Estadístico Nacional

Además, el INE debe proponer normas metodológicas (sobre conceptos, unidades estadísticas, clasificaciones, etc.), que serán de uso común en todos los servicios estadísticos con el fin de garantizar la homogeneidad y comparabilidad de los resultados. Asimismo, son competencia del INE las relaciones en materia estadística con los organismos internacionales especializados y, en particular, con la Oficina de Estadística de la Unión Europea (EUROSTAT).

En el desarrollo del Sistema Estadístico Nacional están también implicados los servicios estadísticos de las distintas Administraciones del Estado, entre los que cabe destacar los correspondientes a los distintos ministerios, encargados de elaborar estadísticas relativas a las actividades de su competencia, y el Banco de España, institución que cuenta con servicios estadísticos responsables de la elaboración de las estadísticas monetarias y financieras.

En cuanto a los servicios estadísticos de las administraciones regionales, la mayor parte de las comunidades autónomas cuentan con una regulación y un instituto de estadística propios, si bien en algunos casos dichos servicios están vinculados directamente a la administración regional.

### Sistema estadístico europeo: EUROSTAT

El proceso de integración de los distintos países en la Unión Europea ha venido acompañado de una demanda creciente de información estadística. La elaboración de políticas comunitarias ha de estar apoyada en estadísticas que sean comparables entre los distintos países miembros. Así, las estadísticas regionales europeas sirven para orientar a la Comisión Europea en la asignación de fondos regionales, las estadísticas de precios armonizados juegan un papel clave en la política monetaria de los Bancos Centrales, etc.

> Las estadísticas europeas pueden contribuir a describir al "ciudadano medio europeo"[1]:
>
> - En el caso de la mujer, tendría 42 años de edad, y podría llegar a vivir otros 41 años. Tendría unos 28 años de edad cuando tuvo su primer hijo y tendría menos de dos niños. Trabaja en los servicios públicos o sociales alrededor de 33 horas a la semana y ha completado al menos la educación secundaria superior.
> - El hombre tendría 39 años de edad y una esperanza de vida de otros 39 años. Trabaja en el sector de servicios de mercado alrededor de 40 horas a la semana y ha completado al menos la educación secundaria superior.

El pilar básico del sistema estadístico europeo es la *Oficina de Estadística de la Unión Europea*, también conocida como EUROSTAT. Se trata de un organismo dependiente de la Comisión Europea, cuya misión fundamental es proporcionar a las instituciones europeas estadísticas fiables y comparables entre países y regiones miembros de la Unión Europea, países candidatos y países de la Asociación Europea de Libre Comercio (AELC). EUROSTAT trabaja en estrecha colaboración con los institutos nacionales de estadística con el fin de desarrollar un sistema estadístico europeo integrado, estableciendo un lenguaje común, en cuanto a conceptos y metodología en general, entre los sistemas estadísticos nacionales de los países miembros y garantizar así la comparabilidad de los resultados.

En el marco de sus competencias, Eurostat publicó en el año 2005 el *Código de buenas prácticas de las estadísticas europeas*, documento que constituye un instrumento fundamental de la armonización estadística europea y en el que se recogen una serie de quince principios que se comprometen a respetar las autoridades estadísticas nacionales y comunitarias. Algunos de estos principios se refieren a aspectos institucionales y organizativos (por ejemplo, independencia profesional, confidencialidad estadística, imparcialidad y objetividad), un segundo bloque se refiere a aspectos metodológicos de la elaboración de estadísticas y, en tercer lugar, se señalan una serie de principios relativos a la producción de estadísticas con el fin de garantizar que las estadísticas elaboradas satisfagan las necesidades de los usuarios (por ejemplo, oportunidad y puntualidad, coherencia y comparabilidad, accesibilidad y claridad).

### Otros organismos internacionales

En el ámbito internacional, cabe destacar el papel relevante de la Organización de Naciones Unidas (ONU). Su División de Estadística constituye en la actualidad la

[1] Eurostat News Release 154/2010

mayor autoridad en el mundo en materia de estadística con una importante labor en materia de coordinación estadística internacional. Entre sus funciones se encuentra el desarrollo de directrices y normativas comunes de actuación en materia estadística, una labor de apoyo a los países para fortalecer sus sistemas estadísticos nacionales y la recopilación y difusión de información estadística global. En relación con este último punto, la División de Estadística de Naciones Unidas coordina *UNdata*, un servicio de base de datos on-line a nivel mundial, que permite conocer cifras oficiales por países sobre un amplio rango de temas: agricultura, población, educación, empleo, energía, medio ambiente, salud, industria, tecnología, desarrollo humano, ...

Además, numerosos organismos internacionales desarrollan trabajos en materia estadística y en sus páginas web ofrecen información estadística de diferentes ámbitos a nivel mundial: la Organización Internacional del Trabajo (OIT), la Organización para la Cooperación y el Desarrollo Económico (OCDE) o la Organización Mundial de la Salud (OMS), entre otros.

### 1.4.2. Algunas estadísticas económicas

Por lo general, los usuarios del ámbito económico no realizan encuestas para obtener la información estadística que precisan, sino que utilizan estadísticas elaboradas por distintos organismos: el INE, las oficinas regionales de estadística, EUROSTAT, etc. Basta consultar INEbase, la base de datos temática del INE, para comprobar la amplia disponibilidad de estadísticas sobre los temas más diversos: cifras de población, precios, costes laborales, ocupación hotelera, hipotecas, ... Dentro del amplio abanico de estadísticas disponibles se presentan a continuación las características fundamentales de dos estadísticas demográficas de tipo censal, el Padrón Municipal y los Censos Demográficos y de dos estadísticas muestrales dirigidas a hogares, la Encuesta de Población Activa (EPA), que es la principal referencia para conocer la dinámica del mercado laboral a nivel nacional, y la Encuesta de Presupuestos Familiares (EPF), enfocada al estudio de los gastos de los hogares españoles.[2]

#### El Padrón Municipal

El *Padrón Municipal* es un registro administrativo donde constan los vecinos de un municipio, constituyendo prueba de residencia en el municipio y del domicilio habitual en el mismo. Toda persona que viva en España está obligada a inscribirse en el padrón del municipio en el que resida habitualmente (quien viva en varios municipios debe inscribirse únicamente en el que habite durante más tiempo al año). Se trata por tanto de un registro permanentemente actualizado de los residentes en un municipio.

La información recogida en los padrones es muy reducida, la estrictamente necesaria para la gestión municipal, y contiene como obligatorios sólo los siguientes datos de cada vecino: nombre y apellidos, sexo, domicilio habitual, nacionalidad, lugar y

[2] La descripción de estas estadísticas es un resumen de las metodologías detalladas que están disponibles en la web del INE www.ine.es.

fecha de nacimiento y número de Documento Nacional de Identidad o, tratándose de extranjeros, del documento que lo sustituya.

> Todos los aspectos relativos a la elaboración del Padrón y sus usos administrativos y estadísticos están regulados por la Ley 4/1996, por la que se modifica la Ley 7/1985, Reguladora de las Bases del Régimen Local.

La elaboración de los padrones es responsabilidad de los ayuntamientos, con la coordinación y supervisión del INE. A partir de la revisión de los padrones a 1 de enero de cada año, el INE publica las cifras de población declaradas oficiales por el Gobierno y que sirven de base para aspectos tales como la toma de decisiones que afectan a la financiación y competencia de los municipios o la determinación del número de diputados por circunscripción en los procesos electorales. Asimismo, los padrones municipales constituyen el documento base para la elaboración del *Censo Electoral*.

### Los Censos Demográficos

Los *Censos Demográficos* constituyen el proyecto estadístico de mayor envergadura que deben acometer periódicamente los oficinas de estadística de cualquier país. Bajo esta denominación se engloban realmente tres censos diferentes: el Censo de Población, que es el de mayor repercusión y tradición, el Censo de Viviendas y el Censo de Edificios.

Los Censos Demográficos se definen como el conjunto de operaciones estadísticas que permiten determinar el número de habitantes, viviendas y edificios del Estado y sus distintas áreas geográficas (comunidades autónomas, provincias y municipios).

En particular, el *Censo de Población* permite conocer características demográficas y sociales de la población, tales como su estructura por sexo y edad, el estado civil, los movimientos migratorios, los estudios, la situación laboral, etc.

> El primer censo moderno de población en España fue realizado en 1768 por el Conde de Aranda, bajo el reinado de Carlos III. Tras varios censos realizados en los siglos XVIII y XIX, desde el año 1900 vienen realizándose censos oficiales de población de forma ininterrumpida con periodicidad decenal. El cuadro adjunto permite comprobar, a través de las cifras de los censos, el importante incremento de la población española desde el Censo de Aranda.

| Año | Población |
|---|---|
| 1768 | 9.309.804 |
| 1900 | 18.830.649 |
| 2001 | 40.847.371 |

La información de los censos es de gran valor para la toma de decisiones en temas tan importantes para la vida cotidiana como dónde construir nuevos colegios, hospitales o residencias, cómo diseñar incentivos a la natalidad, cómo mejorar el transporte público..., además de la asignación de recursos económicos del Estado o la Unión Europea a Comunidades y Ayuntamientos para desarrollo rural, fomento del empleo, etc.

Los Censos en España son realizados por el INE cada 10 años, siendo las últimas cifras publicadas las correspondientes al Censo de 2001. El próximo Censo de Población tiene como fecha de referencia el 1 de noviembre de 2011 y en él se incluirán todas las personas con residencia habitual en el territorio nacional.

Hasta el año 2001 los censos de población eran operaciones exhaustivas en las que agentes censales del INE visitaban todas las viviendas del país para distribuir y recoger los cuestionarios censales. Gracias a los avances metodológicos y tecnológicos, el Censo de 2011 se basará en la combinación de registros y encuestas por muestreo:

- En primer lugar, se elaborará un "fichero precensal" realizado a partir de un aprovechamiento máximo de los registros administrativos disponibles, tomando como base el Padrón.
- En segundo lugar, se realizará un trabajo de campo con dos grandes operaciones:
  - Un Censo de Edificios exhaustivo que permita la georreferenciación de todos los edificios.
  - Una encuesta por muestreo para conocer las características de las personas y las viviendas. El tamaño muestral será de aproximadamente dos millones y medio de viviendas y la selección muestral se basará en mecanismos aleatorios. Los hogares seleccionados podrán responder por Internet o por correo y los agentes censales únicamente acudirán a los domicilios que no respondan por alguna de las vías mencionadas.

La nueva metodología para la elaboración de los Censos Demográficos de 2011 presenta numerosas ventajas. El aprovechamiento de la información ya existente en múltiples registros administrativos y el porcentaje de respuestas que se obtendrán por canales diferentes al de la entrevista tradicional conllevarán una menor carga de trabajo. Esto permitirá al INE trabajar con una organización más reducida y por tanto mejorar su formación y control contribuyendo así a incrementar la calidad y puntualidad de los resultados, con unos costes más reducidos (se estima que con la nueva metodología se precisará un 90 % menos de personal que en el año 2001).

**¿Puede sustituir el Padrón Municipal al Censo de Población?**

Tanto el Padrón Municipal como el Censo de Población son recuentos de habitantes y el Padrón será el punto de partida para la elaboración del Censo en 2011, pero no son fuentes de información sustitutivas ya que difieren entre sí en cuanto a su finalidad y contenido.

El Censo de Población es un documento estadístico que se realiza cada diez años y que no permite la difusión de los datos personales de los ciudadanos (nombre, apellidos, DNI), con el fin de preservar el secreto estadístico. Todo lo contrario que el Padrón, que es un documento administrativo que se actualiza permanentemente y en el que los datos nominales de los residentes en el muncipio son imprescindibles. En resumen, el Censo de Población es una foto fija de la población que incluye muchos datos pero totalmente anónimos; en cambio, el Padrón es un registro vivo que contiene menos información pero perfectamente identificada.

Una representación gráfica asociada habitualmente a los datos demográficos de censos y padrones es la *pirámide de población*. Se trata de una representación de tipo mixto mediante la que se analizan conjuntamente la variable "edad" (agrupada en intervalos) y el atributo "sexo", cuya construcción se basa en la consideración de los histogramas de edad separadamente para las poblaciones masculina y femenina.

La pirámide de población se utiliza en demografía para tener una visión global de la población de un país o región por sexos y edades, analizando las tendencias de crecimiento o estancamiento de la población. La forma de la pirámide refleja tendencias poblacionales y así, bases amplias junto con vértices apuntados son síntomas de poblaciones expansivas mientras que si la base es pequeña en términos relativos y la cúspide achatada, la población se encuentra en fase de envejecimiento. Ejemplos de ambas situaciones quedan reflejados en las pirámides de la figura 1.9, correspondientes a la población española según la información de los Censos de los años 1900 y 2001.

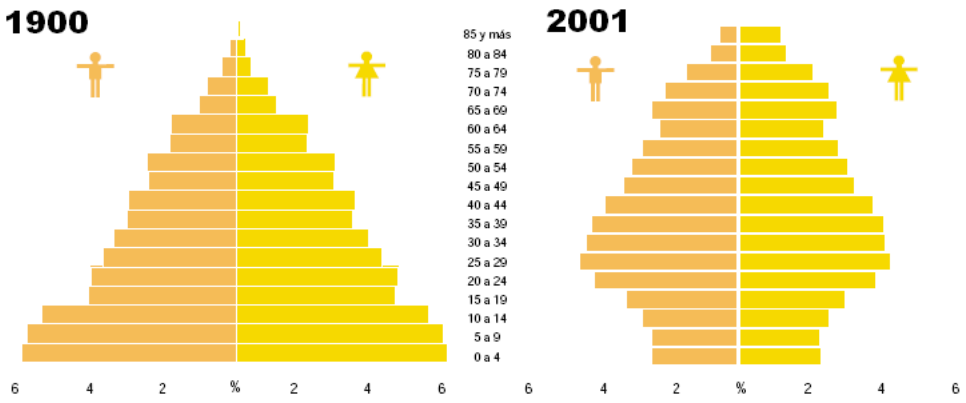



Figura 1.9: Pirámides de la población española (INE)

### La Encuesta de Población Activa

La *Encuesta de Población Activa* (EPA) es una investigación que viene realizando el INE desde 1964, cuya finalidad principal es conocer la actividad económica en lo relativo a su componente humano, proporcionando datos sobre las principales categorías poblacionales en relación con el mercado de trabajo (ocupados, parados, activos e inactivos).

La EPA es una investigación por muestreo de periodicidad trimestral, dirigida a la población que reside en viviendas familiares. Para garantizar que la situación laboral de las personas que integran la muestra represente adecuadamente a la de toda la población española de 16 y más años, el proceso de selección es aleatorio y se realiza en dos etapas: en la primera se eligen al azar zonas geográficas de los municipios (denominadas secciones censales [3]) y, a continuación, en la segunda etapa, se eligen

[3] Las secciones censales se corresponden con las secciones electorales, se trata de áreas geográficas de un municipio con un tamaño entre 500 y 2.000 electores

viviendas de las zonas seleccionadas previamente. Cada trimestre se entrevistan por vía telefónica 65.000 hogares, lo que supone aproximadamente 200.000 personas. Cada trimestre se renueva una sexta parte de la muestra, de modo que los hogares seleccionados colaboran durante seis trimestres consecutivos.

En la Encuesta de Población Activa se clasifica en varias categorías a la población de 16 y más años, que es la que está capacitada legalmente para poder trabajar:

- *Población económicamente activa*, constituida por las personas que durante la semana de referencia suministran mano de obra para la producción de bienes y servicios económicos o que están disponibles y hacen gestiones para incorporarse a dicha producción. Comprende a las personas de al menos 16 años que satisfacen las condiciones necesarias para ser consideradas ocupadas o paradas:
  - Se define como *ocupada* toda persona, de al menos 16 años, que tiene un trabajo por cuenta ajena o ejerce actividad por cuenta propia.
  - Se clasifica como *parada* toda persona de al menos 16 años que cumple simultáneamente los requisitos de estar "sin empleo", "disponible para trabajar" y "busca activamente empleo". Siguiendo la normativa de la Unión Europea, en la actualidad se consideran métodos activos de búsqueda de empleo, entre otros, estar en contacto con una oficina -pública o privada- de empleo con el fin de encontrar trabajo, anunciarse o responder a anuncios de periódicos, participar en una prueba o entrevista en el marco de un procedimiento de contratación ... Por tanto, la mera inscripción como demandante de empleo en las oficinas de empleo públicas, no supone la clasificación de una persona como parado.
- *Población económicamente inactiva*, integrada por el resto de personas, excluidas del mercado laboral. Así, por ejemplo, pertenecen a esta categoría personas que se ocupan exclusivamente de su hogar, estudiantes, jubilados e incapacitados para trabajar.

Los principales resultados de la encuesta son estimaciones trimestrales, tanto nacionales como desagregadas por comunidades autónomas, del número total de activos, ocupados y parados, e inactivos, que son clasificados, a su vez, atendiendo a características demográficas (sexo y edad), de índole cultural (nivel de estudios, formación profesional, etc.) y económica (profesión, rama de actividad, etc.). Se calculan además dos indicadores adicionales, de gran trascendencia para el análisis de la coyuntura económica: la tasa de actividad, definida como cociente del número total de activos entre la población de 16 años y más, y la tasa de paro, que se define como el cociente del número de parados entre el de activos.

### La Encuesta de Presupuestos Familiares

La *Encuesta de Presupuestos Familiares* (EPF) es una investigación realizada por el INE con el objetivo de proporcionar información sobre la naturaleza y destino de

los gastos de consumo de los hogares para el conjunto nacional y para las comunidades autónomas. Por tanto, la variable central de la encuesta es el gasto de consumo, entendiendo como tal tanto el flujo monetario que destina el hogar al pago de determinados bienes y servicios de consumo final, como el valor de determinados consumos no monetarios efectuados por los hogares (entre los que se incluyen el salario en especie o el alquiler estimado de la vivienda en propiedad en la que reside el hogar).

La EPF es una encuesta muestral de periodicidad anual. El procedimiento de selección muestral es similar al de la EPA, considerándose en este caso muestras de 24.000 hogares, que colaboran durante un periodo de dos años. Cada hogar seleccionado presta su colaboración durante dos semanas consecutivas al año en las que debe informar sobre todos los bienes y servicios consumidos.

> La encuesta anual viene realizándose desde el año 2006. Con anterioridad el INE realizaba con periodidad trimestral la Encuesta Continua de Presupuestos Familiares, que sirvió de referencia para la elaboración del Índice de Precios de Consumo base 2006.

La información sobre el gasto que aporta la EPF constituye el elemento básico para la estimación del Consumo Privado en el Sistema de Cuentas Nacionales y para establecer la cesta de la compra y la estructura de ponderaciones del Índice de Precios de Consumo. La EPF publica resultados relativos al gasto medio por hogar y por persona según grupos de gasto, características de los hogares (tamaño y tipo de hogar o principal fuente de ingresos, por ejemplo) y del sustentador principal (sexo, edad, situación laboral, nivel de formación, etc.). Asimismo se proporcionan datos sobre el consumo en cantidades físicas de determinados bienes alimenticios, bebidas, tabaco y combustibles.

# 2 Medidas de posición

La información contenida en una tabla estadística puede ser resumida mediante algunos valores que proporcionen una visión global del comportamiento de la variable. Estos valores sintéticos son representantes de la distribución y se denominan medidas de posición central o promedios.

Sin duda, el promedio más habitual es la media aritmética. No obstante, podemos plantear situaciones muy diversas en las que esta medida no es la idónea para resumir la información. Por ejemplo, si observamos la edad del conjunto de estudiantes matriculados en un centro universitario, seguramente no será aconsejable aplicar la media aritmética sino determinar la edad más frecuente. Sin embargo, si disponemos de información sobre el gasto semanal en ocio de los estudiantes, para resumir este conjunto de datos sería preferible elegir aquel valor central que se sitúa justo en el medio: una mitad de los estudiantes gasta menos de esa cantidad y la otra mitad gasta más.

> Actualmente los valores medios relacionados con la conducta humana son habituales; sin embargo, en sus inicios no parecía que el cálculo de promedios fuese un instrumento adecuado en este tipo de análisis. El primero en realizar este tipo de estudios fue Jacques Quetelet (1796-1874), quien introdujo el concepto de "hombre medio", partiendo de que todo hombre era el resultado de la actuación de causas constantes.
>
> Quetelet desarrolló numerosos estudios sobre estatura, peso, capacidad torácica, etc., comprobando que, para grupos cuantiosos de personas, sus valores se hallaban distribuidos de forma simétrica respecto a la media aritmética. Estos estudios, que hoy podríamos considerar como habituales, fueron duramente criticados en sus comienzos. Se creía que el estudio estadístico de la conducta humana no tenía sentido, porque ésta se ve influenciada por alguna actuación divina.
>
> Quetelet es recordado además de por lo señalado en los párrafos anteriores, por el enorme impulso que proporcionó a las estadísticas oficiales en Europa

Las tres opciones presentadas no son las únicas a la hora de buscar un representante de la distribución. A lo largo del tema se desarrollarán otras medidas que también son necesarias, bien porque ninguna de las anteriores se adapta al planteamiento del problema, o bien porque la información disponible exige alguna consideración que cualquiera de las anteriores no tiene en cuenta.

Los promedios proporcionan valores que ocupan un lugar central en la distribución. No obstante, resulta también de interés determinar otros valores que ocupan una posición señalada aunque no sea central, por ejemplo la renta máxima por debajo de la cual se encuentra el 10 % de hogares más pobres. En general, estos valores se denominan medidas de posición no central o cuantiles y serán introducidos al final del tema.

## 2.1. Medidas de posición central: promedios

### 2.1.1. La media aritmética

**Definición 2.1.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, $\left(\sum_{i=1}^{k} n_i = N,\right)$ llamamos *media aritmética* de $X$, que denotamos por $\bar{x}$, al valor de la siguiente expresión:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} \tag{2.1.1}$$

En otros términos, la media aritmética es el resultado de dividir la suma de todos los valores entre el número total de datos.

Dado que $f_i = \frac{n_i}{N}$, la media aritmética puede expresarse también como:

$$\bar{x} = \sum_{i=1}^{k} x_i f_i$$

Para calcular la media aritmética de una distribución con datos agrupados los valores de $x_i$ representan las marcas de clase de los intervalos.

**Propiedad 2.1.** *La suma de las desviaciones de los valores de una variable respecto a su media es cero:*

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

*Demostración.* Teniendo en cuenta la definición de la media aritmética y operando con el primer miembro de la ecuación se obtiene que:

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

□

Esta propiedad permite interpretar la media aritmética como "centro de gravedad" de la distribución en el sentido de que, al resumir toda la información en este valor, se compensan los errores que se puedan cometer por exceso y por defecto.

**Propiedad 2.2.** *Si todos los valores de una variable se incrementan en una misma cantidad c (cambio de origen), la media también se incrementa en esa constante, esto es:*

$$x'_i = x_i + c; \quad \forall i = 1, 2, \ldots, k \quad \Rightarrow \quad \bar{x'} = \bar{x} + c$$

*Si todos los valores de una variable experimentan un cambio proporcional, es decir, se multiplican por una misma cantidad c (cambio de escala), la media también se multiplica por esa constante, esto es:*

$$x'_i = c\,x_i;\ \forall i = 1, 2, \ldots, k \quad \Rightarrow \quad \bar{x'} = c\,\bar{x}$$

*Demostración.* Representemos por $(x_i, f_i)$ la distribución inicial y por $(x'_i, f_i)$ la distribución resultante de un cambio de origen. Dado que $x'_i = x_i + c$ se verifica que:

$$\bar{x'} = \sum_{i=1}^{k} x'_i f_i = \sum_{i=1}^{k} (x_i + c) f_i = \sum_{i=1}^{k} (x_i f_i + c f_i) = \bar{x} + c \sum_{i=1}^{k} f_i = \bar{x} + c$$

Análogamente, si ahora representamos por $(x'_i, f_i)$ la distribución resultante de un cambio de escala, se cumple que $x'_i = cx_i$, de donde se deduce que:

$$\bar{x'} = \sum_{i=1}^{k} x'_i f_i = \sum_{i=1}^{k} (cx_i) f_i = c \sum_{i=1}^{k} x_i f_i = c\bar{x}$$

□

**Propiedad 2.3.** *(Propiedad de descomponibilidad) Si se divide una población de tamaño N en p subpoblaciones de tamaños* $N_1, N_2, \ldots, N_p$, $\left(\sum_{j=1}^{p} N_j = N\right)$ *y medias* $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$, *la media poblacional se relaciona con las medias de las subpoblaciones mediante la expresión:*

$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \cdots + \bar{x}_p N_p}{N}$$

*Demostración.* Efectuaremos la comprobación para el caso de dos subpoblaciones. Para ello representemos por $(x_i, n_i)$ la distribución poblacional y designemos por $n_{i1}$ y $n_{i2}$ la frecuencia absoluta de $x_i$ en cada subpoblación; estas frecuencias están relacionadas mediante la expresión $n_{i1} + n_{i2} = n_i$.

El tamaño de las subpoblaciones será $N_1 = \sum_{i=1}^{k} n_{i1}$ y $N_2 = \sum_{i=1}^{k} n_{i2}$ y las respectivas medias vendrán dadas por las expresiones:

$$\bar{x}_1 = \frac{\sum_{i=1}^{k} x_i n_{i1}}{N_1}; \quad \bar{x}_2 = \frac{\sum_{i=1}^{k} x_i n_{i2}}{N_2}$$

En consecuencia:

$$\frac{\bar{x}_1 N_1 + \bar{x}_2 N_2}{N} = \frac{\sum_{i=1}^{k} x_i n_{i1} + \sum_{i=1}^{k} x_i n_{i2}}{N} = \frac{\sum_{i=1}^{k} (x_i n_{i1} + x_i n_{i2})}{N} = \frac{\sum_{i=1}^{k} x_i (n_{i1} + n_{i2})}{N} = \bar{x}$$

□

**Ejemplo 2.1.** Supongamos que los estudiantes matriculados en cierta asignatura están divididos en dos grupos. En el primer grupo se presentaron al examen final 40 estudiantes siendo la nota media del grupo 6, mientras que en el otro grupo se presentaron 60 alumnos y la nota media es 7,5. A partir de esta información, aplicando la ecuación 2.3, podemos calcular la nota media de todos los estudiantes como sigue:

$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2}{N} = \frac{6 \times 40 + 7,5 \times 60}{100} = 6,9$$

### 2.1.2. La media ponderada

En algunas situaciones la importancia que tiene un valor dentro del conjunto viene reflejada mediante información complementaria, que se cuantifica a través de ponderaciones o pesos.

Por ejemplo, si las materias primas de una empresa son importadas en un 20 % y nacionales en el 80 % restante, para calcular el coste medio debemos utilizar una media ponderada, donde las ponderaciones o pesos reflejan la importancia relativa de cada tipo de procedencia geográfica. De modo análogo, si conocemos la estructura del presupuesto de las familias (25 % dedicado a alimentación, 10 % a vestido y calzado, 15 % a transporte...) estas ponderaciones reflejan la importancia relativa de cada tipo de gasto y deberán ser tenidas en cuenta, por ejemplo, para calcular la subida media de precios (de hecho, como veremos más adelante, esto es lo que se hace en el Indice de Precios de Consumo, IPC) .

**Definición 2.2.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, cuya importancia es conocida y viene dada por los pesos o ponderaciones $w_1, w_2, \ldots, w_k$, llamamos *media ponderada* de $X$, que denotamos por $\bar{x}_w$, al valor de la siguiente expresión:

$$\bar{x}_w = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i} \tag{2.1.2}$$

En la práctica, el mayor problema a la hora de aplicar esta medida surge por las dificultades de conocer, en muchos casos, las ponderaciones. Estos pesos suelen obtenerse a partir de encuestas o de informaciones complementarias sobre la variable.

**Ejemplo 2.2.** En el último semestre un estudiante se ha examinado de varias asignaturas cuyo número de créditos es diferente. En esta situación, dado que las asignaturas no tienen la misma importancia en el expediente académico, para calcular la nota media se deberá tener en cuenta el número de créditos de cada asignatura. A partir de los datos recogidos en la tabla siguiente:

| Asignatura | Nota ($x_i$) | Nº créditos ($w_i$) |
|---|---|---|
| Matemáticas | 5 | 6 |
| Microeconomía | 9 | 10 |
| Introducción al Derecho | 7 | 4,5 |
| Estadística | 6 | 6 |
| Historia Económica | 9 | 9 |
| Sociología | 7 | 4,5 |

se obtiene que la nota media del estudiante es 7,5. Este resultado es una media ponderada de las notas, donde el peso de cada asignatura viene determinado por el número de créditos.

Algunas propiedades de la media aritmética podrían ser estudiadas como casos ponderados. Así, por ejemplo, al obtener la media de una población a partir de las medias de varias subpoblaciones, la expresión de cálculo no es más que una media ponderada de éstas donde las ponderaciones resultan ser los tamaños de las distintas subpoblaciones.

La media aritmética puede ser considerada como el centro de gravedad de la distribución; los valores bajos llevan a la media a tomar un valor bajo y los altos la llevan a valores altos, de manera que cuando el conjunto de valores es bastante uniforme se compensarán las dos fuerzas y la media resultará representativa. En consecuencia, para aquellas distribuciones que presenten valores anormalmente extremos, es muy probable que la media aritmética supere o quede muy por debajo de la mayoría de las observaciones. En estos casos sería conveniente buscar un representante de la distribución con mayor capacidad descriptiva que la media aritmética.

Los promedios que estudiamos a continuación -mediana y moda- serán complementarios de la media aritmética a la hora de sintetizar la información contenida en una distribución.

### 2.1.3. La mediana

**Definición 2.3.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, ordenados de forma creciente ($x_1 < x_2 < \ldots < x_k$), con frecuencias absolutas $n_1, n_2, \ldots, n_k$, llamamos *mediana*, que denotamos por $Me$, a un valor que divide a la distribución en dos partes iguales, esto es, que deja tantas observaciones a su izquierda como a su derecha.

Supongamos que las notas obtenidas por los 15 estudiantes que han aprobado un examen son las siguientes:

| Notas | Nº estudiantes |
|---|---|
| 5 | 1 |
| 6 | 5 |
| 7 | 3 |
| 8 | 3 |
| 9 | 2 |
| 10 | 1 |

Si se ordenan los datos en sentido creciente: 5, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 10, se puede identificar que el valor 7 ocupa la posición central dejando tanto a su izquierda como a su derecha el mismo número de datos. Dicho valor es la mediana de la distribución.

En este caso resulta inmediato localizar la mediana, pues se dispone de un número impar de datos y, una vez ordenados éstos en forma creciente, se determina fácilmente el valor que ocupa el lugar central.

En general, dada una distribución de datos no agrupados, si el número de datos $N$ es impar, existe un único valor central. Sin embargo, cuando $N$ es par se tienen dos valores centrales; si éstos no coinciden puede decirse que hay infinitos valores medianos, todos los comprendidos entre los dos valores centrales, aunque suele tomarse como mediana la media aritmética de éstos.

Por otro lado, cuando se tiene un gran número de observaciones no resultará operativo hacer una ordenación como la del ejemplo, siendo pues necesario utilizar otro sistema para determinar la mediana. En consecuencia, el método general de cálculo de la mediana de una distribución con datos no agrupados será como sigue:

- Si no existe ningún valor de la distribución cuya frecuencia acumulada coincida con $\frac{N}{2}$, la mediana será el menor valor de la variable que presenta una frecuencia acumulada mayor que $\frac{N}{2}$. En particular, esta situación se dará siempre que $N$ sea impar puesto que en ese caso el valor de $\frac{N}{2}$ no es entero.

- Si $\frac{N}{2}$ coincide con la frecuencia acumulada de un valor $x_i$, la mediana está indeterminada entre los valores $x_i$ y $x_{i+1}$. En tal caso se tomará como mediana la media aritmética de ambos, esto es, $Me = \frac{(x_i + x_{i+1})}{2}$. Esta situación solamente puede aparecer si $N$ es par.

En general, para distribuciones con datos agrupados en intervalos, el método anterior conduce a identificar el intervalo mediano: será aquél que presenta la primera frecuencia acumulada mayor o igual que $\frac{N}{2}$. En el caso de que la frecuencia acumulada del i-ésimo intervalo coincida con $\frac{N}{2}$, la mediana será el extremo superior de dicho intervalo, $L_i$. En otro caso, una vez localizado este intervalo, una primera alternativa sería tomar su marca de clase como mediana, sin embargo se puede obtener una mejor aproximación aplicando el razonamiento que describimos a continuación.

Para determinar cuál es el valor dentro del intervalo mediano que corresponde a la mediana se puede suponer que las observaciones están uniformemente distribuidas

a lo largo del mismo. Entonces utilizando la representación del polígono de frecuencias acumuladas en el tramo que corresponde al intervalo mediano y la semejanza de triángulos (véase figura 2.1) se puede aproximar la mediana como:

$$Me = L_{i-1} + d$$

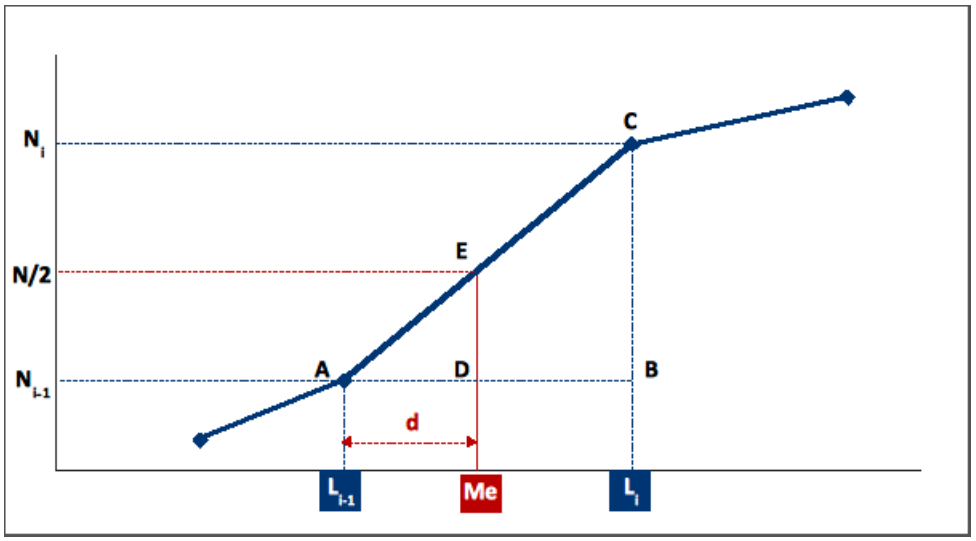



Figura 2.1: Mediana

Para determinar $d$ debemos tener en cuenta la semejanza de los triángulos ABC y ADE, de la que se deriva que:

$$\frac{DE}{AD} = \frac{BC}{AB}$$

de donde, a su vez, teniendo en cuenta la longitud de los lados, se obtiene que:

$$\frac{\frac{N}{2} - N_{i-1}}{d} = \frac{N_i - N_{i-1}}{L_i - L_{i-1}}$$

Finalmente, despejando $d$ en la igualdad anterior y sustituyendo, se llega a la siguiente expresión:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i \tag{2.1.3}$$

**Propiedad 2.4.** *Si la variable $X$ experimenta un cambio de origen, la mediana de la variable transformada $(X' = X + c)$ será $Me' = Me + c$*

*Si la variable $X$ se ve afectada por un cambio de escala, la mediana de la variable transformada $(X' = cX)$ será $Me' = cMe$.*

*Demostración.* Dada la distribución inicial $(x_i, n_i)$, un cambio de origen supone únicamente una traslación de los valores sin afectar a las respectivas frecuencias, esto es, la distribución transformada será $(x_i^{'} = x_i + c, n_i)$ donde $c$ es una constante arbitraria. En consecuencia, dicha transformación no altera la posición que ocupan los valores dentro de la distribución y por lo tanto si $Me$ es la mediana de la distribución inicial, $Me^{'} = Me + c$ será la mediana de la distribución transformada.

El mismo tipo de razonamiento es aplicable a los cambios proporcionales. Ahora la distribución transformada será $(x_i^{'} = cx_i, n_i)$ y su mediana vendrá dada por $Me^{'} = cMe$. □

### 2.1.4. La moda

Al representar gráficamente un conjunto de datos, mediante un diagrama de barras (si no están agrupados) o mediante un histograma (si están agrupados en intervalos), la característica que más resalta a primera vista posiblemente sea su máximo. En este sentido el valor de la variable que determina dicho máximo en la representación gráfica resume la información inicial.

**Definición 2.4.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, se define la *moda*, que denotamos por $Mo$, como aquel valor de la variable que más veces se repite, esto es, que presenta mayor frecuencia.

Su cálculo es inmediato cuando los datos están sin agrupar, salvo que haya más de un valor con esta frecuencia máxima, en cuyo caso se podría hablar de distribuciones bimodales, trimodales, ..., plurimodales en general.

Para distribuciones con datos agrupados, antes de determinar el valor de la moda habrá que localizar el intervalo que la contiene, que llamamos intervalo modal, que será aquél que presenta mayor frecuencia por unidad de amplitud, es decir, el que tiene mayor altura $\left(h_i = \frac{n_i}{a_i}\right)$ en el histograma. En lugar de tomar como valor aproximado de la moda la marca de clase del intervalo modal, asumiremos que la moda se aproxima más al intervalo contiguo de mayor altura (véase figura 2.2).

Este planteamiento supone que las distancias de la moda a los intervalos contiguos son inversamente proporcionales a sus alturas. Entonces, si denotamos por $a$ y $b$ las distancias a los intervalos anterior y posterior respectivamente, se cumplirá que:

$$ah_{i-1} = bh_{i+1}$$

Aplicando una propiedad de las proporciones se tiene:

$$\frac{a}{h_{i+1}} = \frac{b}{h_{i-1}} = \frac{a+b}{h_{i-1}+h_{i+1}}$$

de donde:

$$a = \frac{h_{i+1}(a+b)}{h_{i-1}+h_{i+1}} = \left(\frac{h_{i+1}}{h_{i-1}+h_{i+1}}\right) a_i$$

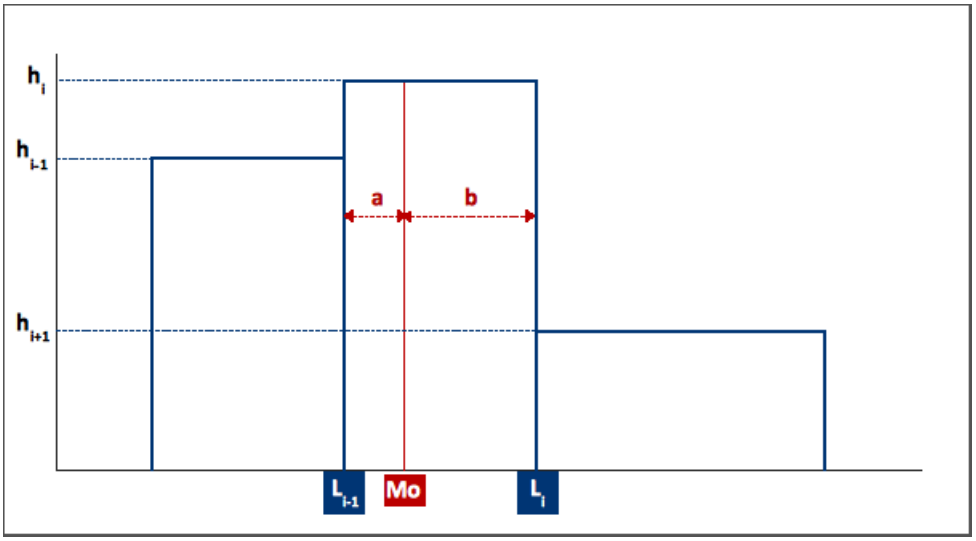



Figura 2.2: Moda

Finalmente, dado que $Mo = L_{i-1} + a$, se tiene:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i \tag{2.1.4}$$

Si los datos están agrupados en intervalos de la misma amplitud, el intervalo o clase modal coincide con el que presenta mayor frecuencia y el valor de la moda puede ser calculado a partir de las frecuencias absolutas por medio de la siguiente expresión:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i$$

**Propiedad 2.5.** *Si la variable $X$ experimenta un cambio de origen, la moda de la variable transformada $(X' = X + c)$ será $Mo' = Mo + c$*

*Si la variable $X$ se ve afectada por un cambio de escala, la moda de la variable transformada $(X' = cX)$ será $Mo' = c\,Mo$*

*Demostración.* La demostración resulta evidente sin más que tener en cuenta que en ambos casos las frecuencias de los valores no se modifican. Así pues, si $x_i$ es el valor modal de la distribución inicial $(x_i, n_i)$, $x_i + c$ lo será de la distribución resultante tras un cambio de origen, cumpliéndose que $Mo' = Mo + c$. Análogamente, $cx_i$ será el valor modal de la distribución transformada tras un cambio de escala. □

**Ejemplo 2.3.** Se ha observado la producción de leche de vaca obtenida el último mes en 100 explotaciones ganaderas, obteniendo la siguiente distribución:

| Producción de leche (miles de litros) | Nº de explotaciones |
|---|---|
| 5-10 | 5 |
| 10-15 | 10 |
| 15-20 | 15 |
| 20-30 | 15 |
| 30-50 | 35 |
| 50-70 | 15 |
| 70-80 | 5 |

Dado que se trata de una distribución con datos agrupados, para calcular la mediana debemos localizar en primer lugar el intervalo mediano, es decir, la clase cuya frecuencia acumulada supera por primera vez a $\frac{N}{2}$.

En este caso $\frac{N}{2} = 50$; si observamos las frecuencias acumuladas recogidas en la tercera columna de la tabla siguiente podemos concluir que el intervalo mediano es el quinto (30-50):

| Producción de leche (miles de litros) | Nº de explotaciones | $N_i$ | $h_i$ |
|---|---|---|---|
| 5-10 | 5 | 5 | 1 |
| 10-15 | 10 | 15 | 2 |
| 15-20 | 15 | 30 | 3 |
| 20-30 | 15 | 45 | 1,5 |
| 30-50 | 35 | 80 | 1,75 |
| 50-70 | 15 | 95 | 0,75 |
| 70-80 | 5 | 100 | 0,5 |

Tabla 2.1: Producción de leche

Finalmente, aplicando la fórmula propuesta anteriormente para el cálculo de la mediana con datos agrupados obtenemos que $Me = 32,857$. Este resultado nos indica que la mitad de las explotaciones han producido el mes pasado 32.857 litros o menos.

Calculemos ahora la moda de la distribución, para lo cual debemos identificar previamente el intervalo modal, esto es, la clase que presenta mayor altura. Observando las alturas recogidas en la última columna de la tabla anterior podemos comprobar que el intervalo modal es el tercero (15-20). A su vez, aplicando la fórmula propuesta anteriormente para el cálculo de la moda con datos agrupados, se llega al siguiente resultado: $Mo = 17,143$, que indica que la producción láctea mas frecuente es de 17.143 litros.

### 2.1.5. Otros promedios: media geométrica y media armónica

Aunque los promedios definidos hasta aquí (media aritmética, mediana y moda) son las medidas aplicadas habitualmente para resumir un conjunto de datos, existen

situaciones en que la propia naturaleza de la información conduce a otro tipo de medidas como son la media geométrica y la media armónica.

**Definición 2.5.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, se define la *media geométrica*, que denotamos por $G$, como el valor dado por la siguiente expresión:

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}} \tag{2.1.5}$$

En otros términos, la media geométrica es la raíz N-ésima del producto de todas las observaciones.

Para situaciones en que la variable de interés presente variaciones acumulativas, la media geométrica será el promedio adecuado para resumir su comportamiento. Así, en el tema 8, aplicaremos esta medida para calcular índices o tasas medias de variación de una magnitud económica (precios, salarios, ...) en un periodo temporal dado.

Si la variable presenta valores positivos y negativos, no tiene sentido calcular la media geométrica; tampoco si alguna observación es nula.

**Definición 2.6.** Dada una variable estadística $X$, que toma un conjunto de valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, se define la *media armónica*, que denotamos por $H$, como el valor dado por la siguiente expresión:

$$H = \frac{N}{\frac{n_1}{x_1} + \cdots + \frac{n_k}{x_k}} \tag{2.1.6}$$

La media armónica no se puede calcular si la variable presenta algún valor nulo.

**Ejemplo 2.4.** Un automovilista hizo el recorrido entre dos ciudades en 4 etapas; en la tabla siguiente se indica la distancia recorrida y la velocidad empleada en cada etapa:

| Etapa | $Velocidad$ $(km/h)$ | $Distancia$ $(km)$ |
|---|---|---|
| 1 | 60 | 45 |
| 2 | 80 | 70 |
| 3 | 100 | 200 |
| 4 | 70 | 85 |

A partir de estos datos podemos plantearnos calcular la velocidad media v del recorrido total. Teniendo en cuenta el concepto de velocidad (espacio/tiempo), el tiempo empleado en cada etapa lo obtendremos a partir de los datos anteriores como cociente de la distancia reccorrida y la velocidad correspondientes y, en consecuencia, la velocidad media vendrá dada por la siguiente expresión:

$$v = \frac{espacio}{tiempo} = \frac{400}{\frac{45}{60} + \frac{70}{80} + \frac{200}{100} + \frac{85}{70}} = 82,66$$

Así pues para calcular la velocidad media hemos aplicado el concepto de media armónica.

- La media armónica de una variable $X$ coincide con la inversa de la media arimética de la variable inversa de $X$, $\frac{1}{X}$.
- Las medias armónica, geométrica y aritmética están relacionadas por las siguientes desigualdades: $H \leq G \leq \bar{x}$. La demostración puede consultarse en [2] pp. 76-77.

### 2.1.6. Ventajas e inconvenientes de los promedios

Con el fin de resumir un conjunto de datos hemos definido distintas medidas de tendencia central o promedios, lo que nos indica que no existe una que sea idónea en todas las situaciones. Cada promedio presenta ventajas e inconvenientes que harán aconsejable o no su aplicación como representante según el tipo de problema a resolver. Señalamos a continuación las características básicas de estas medidas.

- Media aritmética: Su cálculo es sencillo y en el mismo intervienen todas las observaciones; sin embargo, su resultado es sensible ante la presencia de valores extremos (anormalmente bajos o altos). En consecuencia, su aplicación será aconsejable cuando los datos son bastante homogéneos.

- Mediana: Es una medida más robusta que la media aritmética, es decir, menos sensible ante la presencia de valores extremos; ahora bien, presenta el inconveniente de que en su cálculo no intervienen todas las observaciones sino únicamente las observaciones centrales. Por consiguiente será recomendable su utilización cuando los datos son irregulares, es decir, aparecen observaciones anormalmente bajas o altas.

- Moda: Presenta las mismas ventajas e inconvenientes que la mediana. Su aplicación es apropiada cuando algún valor absorbe la mayor parte de las frecuencias, esto es, la mayoría de las observaciones son iguales entre sí. Sin embargo, a los inconvenientes añadiremos que en una distribución pueden existir varios valores modales.

- Media geométrica: es una medida apropiada cuando la variable tiene carácter acumulativo. Por otra parte, como hemos señalado anteriormente, carece de sentido si hay algún valor nulo o si se presentan simultáneamente valores positivos y negativos.

- Media armónica: Su interpretación no es tan clara como la de las anteriores. Aunque utiliza todos los datos, presenta el inconveniente de que es muy sensible ante la presencia de valores bajos. Además, no está definida cuando alguna observación es nula.

Dada la importancia del problema, en el tema siguiente definiremos medidas específicas para analizar la representatividad de los promedios.

## 2.2. Medidas de posición no central: cuantiles

Los *cuantiles* constituyen un grupo de medidas de significado análogo al de la mediana, con la diferencia de que en vez de apuntar al centro de la distribución, ahora el objetivo es determinar valores que la dividan en unas "cuantas" partes iguales.

Entre los cuantiles podemos citar, por ser de uso más frecuente, los cuartiles, los deciles y los centiles o percentiles. Los cuartiles son tres puntos $(Q_1, Q_2, Q_3)$ que dividen a la distribución en cuatro partes iguales, es decir, en cuatro intervalos en cada uno de los cuales está incluido, respectivamente, un 25 % de los valores de la distribución. Los deciles son nueve valores $(D_r,\ r = 1, \ldots, 9)$ que dividen a la distribución en 10 partes tales que dentro de cada intervalo está incluido un 10 % de las observaciones. Siguiendo el mismo esquema de definición diremos que los centiles son 99 números $(C_r,\ r = 1, \ldots, 99)$ que se obtienen al dividir la distribución en 100 partes iguales.

**Definición 2.7.** Dada una variable estadística $X$, cuyos valores suponemos ordenados en sentido creciente, se definen distintos tipos de cuantiles como sigue:

- El *cuartil* de orden $r$, que designamos por $Q_r$, es un valor que divide a la distribución en dos partes, dejando a su izquierda $r\frac{N}{4}$ observaciones $(r = 1, 2, 3)$
- El *decil* de orden $r$, que designamos por $D_r$, es un valor que divide a la distribución en dos partes, dejando a su izquierda $r\frac{N}{10}$ observaciones $(r = 1, 2, \ldots, 9)$
- El *centil* de orden r, que designamos por $C_r$, es un valor que divide a la distribución en dos partes, dejando a su izquierda $r\frac{N}{100}$ observaciones $(j = 1, 2, \ldots, 99)$

Cabe destacar que, por la propia definición de estas medidas, existen coincidencias entre diversos cuantiles. Así, el segundo cuartil coincide con la mediana de la distribución; asimismo, el primer decil es igual al centil de orden 10, el quinto decil coincide con el segundo cuartil y con la mediana, etc.

En general, el procedimiento de cálculo de un cuantil es análogo al descrito para la mediana anteriormente. En particular, detallaremos a continuación las expresiones correspondientes a los cuartiles.

En una distribución de datos sin agrupar, para calcular el cuartil $Q_r$ habrá que buscar el valor de la variable que corresponde a la primera frecuencia acumulada mayor o igual que $r\frac{N}{4}$. Como en el caso de la mediana, si $r\frac{N}{4}$ coincide con la frecuencia acumulada de algún valor $x_i$ de la variable, se tomará como valor del cuartil la media aritmética de este valor y el que le sigue en la ordenación creciente, esto es, $Q_r = \dfrac{(x_i + x_{i+1})}{2}$.

Para distribuciones con datos agrupados en intervalos, el primer paso será identificar el intervalo que contiene al cuartil y el valor de éste vendrá dado por la siguiente expresión:

$$Q_r = L_{i-1} + \frac{r\frac{N}{4} - N_{i-1}}{n_i}\, a_i, \quad (r = 1, 2, 3) \tag{2.2.1}$$

Análogamente, las expresiones de cálculo de deciles y centiles para distribuciones con datos agrupados son las siguientes:

$$D_r = L_{i-1} + \frac{r\frac{N}{10} - N_{i-1}}{n_i} a_i, \quad (r = 1, \ldots, 9)$$

$$C_r = L_{i-1} + \frac{r\frac{N}{100} - N_{i-1}}{n_i} a_i, \quad (r = 1, \ldots, 99)$$

**Ejemplo 2.5.** Retomemos el ejemplo 2.3 sobre producción de leche para determinar el tercer cuartil.

Dado que $3\frac{N}{4} = 75$, observando las frecuencias acumuladas calculadas en la tercera columna de la tabla 2.1, podemos concluir que el tercer cuartil se encuentra en el quinto intervalo (30-50) y, aplicando la fórmula 2.2.1, se obtiene su valor $Q_3 = 47,143$. Este resultado indica que el 75 % de las explotaciones han producido 47.143 litros de leche como máximo.

### El diagrama de cajas

El *diagrama de cajas* es una representación gráfica basada en los cuartiles y en los valores extremos (mínimo y máximo) de la distribución. La denominación del gráfico se debe a que está compuesto por una caja cuya altura viene determinada por el primer cuartil $Q_1$ y el tercero $Q_3$ y de la que salen dos líneas verticales, la inferior limitada por el valor mínimo y la superior por el máximo.

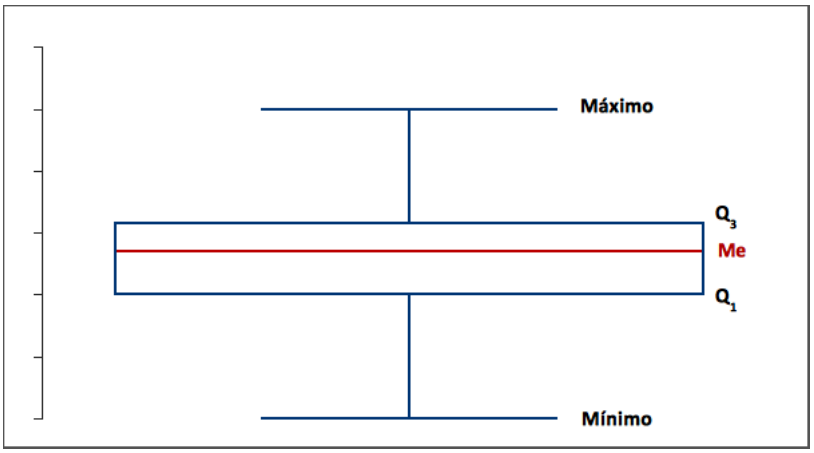



Figura 2.3: Diagrama de cajas

# 3 Medidas de dispersión y forma

Los promedios o medidas de posición central, estudiados en el tema anterior, son números que tratan de representar a un conjunto de datos. De ahí que sea necesario también analizar su grado de representatividad, problema del que nos ocuparemos en este tema y que resolveremos mediante las medidas de dispersión.

## 3.1. Medidas de dispersión absolutas

Las medidas de dispersión tratan de sintetizar en un único número la separación entre los distintos valores de una variable, es decir, su objetivo es cuantificar la variabilidad de un conjunto de datos. Por consiguiente, una primera aproximación a la dispersión vendrá dada por la diferencia entre las observaciones extremas.

**Definición 3.1.** Dada una variable estadística $X$, con distribución de frecuencias $(x_i, f_i)$, se define el *recorrido* o *rango*, que denotamos por $R$, como la diferencia entre el mayor y el menor valor de la variable:

$$R = \max_i(x_i) - \min_i(x_i) \tag{3.1.1}$$

Esta medida presenta la ventaja de que su cálculo es sencillo; sin embargo, al basarse exclusivamente en los valores extremos puede inducir a una sobrevaloración de la variabilidad en la medida en que esos valores sean anómalos, es decir, estén alejados del resto de las observaciones. Asimismo, otro inconveniente a destacar es que en su definición no interviene ningún promedio.

Las distribuciones representadas en los diagramas de cajas de la Figura 3.1 presentan el mismo recorrido, sin embargo se puede apreciar que la variabilidad del 50 % de los valores centrales es significativamente menor en la distribución A, puesto que la altura de la caja correspondiente es más pequeña.

**Definición 3.2.** Dada una variable estadística $X$, se define el *recorrido intercuartílico*, que denotamos por $R_I$, como la diferencia entre el tercero y el primero de los cuartiles de la distribución:

$$R_I = Q_3 - Q_1 \tag{3.1.2}$$

Este resultado nos indica la amplitud del intervalo en el que están comprendidos el 50 % de los valores centrales de la distribución, evitando así el problema de las observaciones anómalas. En el diagrama de cajas, el recorrido intercuartílico se identifica con la altura de la caja (distancia entre el primer cuartil y el tercero).

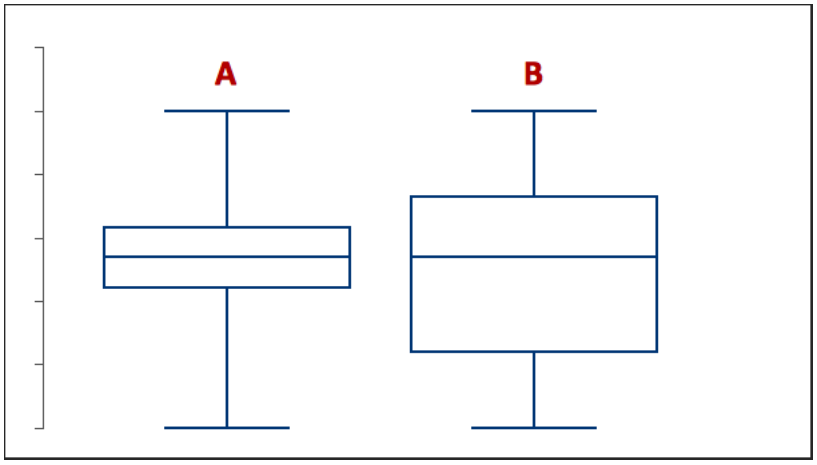



Figura 3.1: Recorrido

Ahora bien, como hemos señalado anteriormente, el principal objetivo que perseguimos al estudiar la dispersión es analizar la representatividad de los promedios, problema que no podemos resolver aplicando ninguna de estas dos medidas. La representatividad de un promedio P dependerá del grado de dispersión que presenten los datos con respecto a dicho valor, y por tanto es necesario definir medidas de dispersión que tengan en cuenta al promedio.

Para determinar el grado de dispersión de los valores respecto a un promedio $(P)$ habrá que medir las desviaciones de los valores respecto a ese valor de referencia, $(x_i - P)$, y utilizar alguna medida de síntesis de estas desviaciones. Para ello una primera medida que podemos considerar es la suma $\sum_{i=1}^{k}(x_i - P)n_i$. Ahora bien, si tenemos en cuenta que el signo de las desviaciones puede ser positivo o negativo, el resultado de la suma puede ser un valor muy pequeño y, no obstante, existir grandes desviaciones en la distribución. En particular, si el promedio de referencia fuese la media aritmética, la suma de todas las desviaciones resultaría siempre nula con independencia de la distribución de partida. El problema en la medida propuesta es que aparecen desviaciones con distinto signo. Sin embargo, para medir la dispersión debemos tener en cuenta la cuantía de esta desviación y no su signo ya que éste refleja sólo un carácter cualitativo respecto a la aproximación que el promedio $P$ hace del valor $x_i$: si es por exceso, $(x_i < P)$, el signo es negativo y en caso contrario será positivo. Para solucionar este inconveniente (evitar la influencia del signo), consideraremos dos posibilidades: elevar al cuadrado las desviaciones o tomar el valor absoluto de éstas.

**Definición 3.3.** Llamamos *desviación absoluta media* respecto a un promedio $P$, que denotamos por $D_P$, al valor de la siguiente expresión:

$$D_P = \sum_{i=1}^{k} |x_i - P| f_i$$

Así pues, esta medida se obtiene como la media aritmética de las desviaciones

absolutas de los valores de la variable con respecto al promedio $P$.

En particular, por ejemplo, la *desviación absoluta media con respecto a la media aritmética* vendrá por la siguiente expresión:

$$D_{\bar{x}} = \sum_{i=1}^{k} |x_i - \bar{x}| f_i$$

**Definición 3.4.** Dada una variable estadística $X$, con distribución de frecuencias $(x_i, f_i)$, se define la *desviación cuadrática media* respecto a un promedio $P$, que denotamos por $D_P^2$, como la siguiente expresión:

$$D_P^2 = \sum_{i=1}^{k} (x_i - P)^2 f_i$$

Así pues, se trata de una medida de dispersión respecto al promedio $P$ que consiste en calcular la media aritmética de las desviaciones cuadráticas. En particular, si el promedio considerado es la media aritmética, se obtiene la varianza, que es la medida de dispersión más importante por su uso generalizado.

### 3.1.1. Varianza y desviación típica

**Definición 3.5.** Llamamos *varianza*, que denotamos por $S_X^2$, a la desviación cuadrática media respecto a la media aritmética de los valores de la variable:

$$S_X^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i \tag{3.1.3}$$

**Propiedad 3.1.** *La varianza toma valores no negativos, es decir,* $S_X^2 \geq 0$. *Además* $S_X^2 = 0$ *si y solo si todas las observaciones son iguales.*

*Demostración.* De la propia definición se desprende de forma inmediata que la varianza no puede ser negativa puesto que los sumandos que intervienen en ella solamente pueden tomar valores no negativos.

Por otra parte, si $S_X^2 = 0$ ha de cumplirse para cada uno de los sumandos que $(x_i - \bar{x})^2 f_i = 0$. Dado que $f_i \neq 0$, de la igualdad anterior se deriva que $x_i = \bar{x}$, $\forall i = 1, \ldots, k$, es decir, todas las observaciones son iguales.

En sentido contrario, si todas las observaciones son iguales entre sí, la media aritmética también será igual, de donde se deriva de forma inmediata que la varianza es nula. □

**Propiedad 3.2.** *La varianza de una distribución se puede expresar como:*

$$S_X^2 = \sum_{i=1}^{k} x_i^2 f_i - \bar{x}^2 \tag{3.1.4}$$

*Demostración.* En efecto, partiendo de la definición de la varianza y operando se llega a la siguiente expresión:

$$S_X^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 f_i = \sum_{i=1}^{k}(x_i^2 + \bar{x}^2 - 2x_i\bar{x})f_i = \sum_{i=1}^{k} x_i^2 f_i + \bar{x}^2 \overbrace{\sum_{i=1}^{k} f_i}^{=1} - 2\bar{x}\overbrace{\sum_{i=1}^{k} x_i f_i}^{=\bar{x}}$$
$$= \sum_{i=1}^{k} x_i^2 f_i - \bar{x}^2$$

□

Esta propiedad proporciona una expresión alternativa que tiene mucho interés desde el punto de vista práctico, pues facilita el cálculo de la varianza.

**Propiedad 3.3.** *Si todos los valores de una variable se incrementan en una misma cantidad c, la varianza no varía, es decir,* $S_{X+c}^2 = S_X^2$. *En otros términos, la varianza no se ve afectada por cambios de origen.*

*Si todos los valores de una variable se multiplican por una misma cantidad c, la varianza se multiplica por el cuadrado de esa constante, es decir,* $S_{cX}^2 = c^2 S_X^2$.

*Demostración.* Sea $(x_i, f_i)$ la distribución de la variable $X$; entonces la distribución de la variable $X+c$, resultante del cambio de origen, será $(x_i+c, f_i)$ cuya media aritmética, teniendo en cuenta el comportamiento de la media ante este tipo de cambio, será $\bar{x}+c$ (propiedad 2.2). Así pues se verificará que:

$$S_{X+c}^2 = \sum_{i=1}^{k}\left[(x_i + c) - (\bar{x} + c)\right]^2 f_i = \sum_{i=1}^{k}(x_i - \bar{x})^2 f_i = S_X^2$$

Análogamente, la distribución de la variable $cX$, resultante del cambio de escala, será $(cx_i, f_i)$ cuya media aritmética será $c\bar{x}$ (propiedad 2.2). En consecuencia se cumplirá que:

$$S_{cX}^2 = \sum_{i=1}^{k}(cx_i - c\bar{x})^2 f_i = \sum_{i=1}^{k} c^2(x_i - \bar{x})^2 f_i = c^2 \sum_{i=1}^{k}(x_i - \bar{x})^2 f_i = c^2 S_X^2$$

□

La desviación cuadrática media respecto a cualquier otro promedio P (mediana, moda) se comporta de igual modo que la varianza ante cambios de origen y de escala.

**Propiedad 3.4.** *La media de las desviaciones cuadráticas de los valores de una variable respecto a una constante cualquiera c se hace mínima cuando dicha constante es igual a la media aritmética, es decir,* $c = \bar{x}$*:*

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

*Demostración.* La función $G(c) = \sum_{i=1}^{k}(x_i - c)^2 f_i$ es de tipo parabólico y toma solamente valores no negativos, por tanto presenta un valor mínimo.

Atendiendo a la condición necesaria de óptimo, el valor de $c$ para el que se alcanza el mínimo ha de cumplir la condición siguiente:

$$\frac{\partial G(c)}{\partial c} = -2\sum_{i=1}^{k}(x_i - c)f_i = 0$$

de donde se obtiene:

$$\sum_{i=1}^{k}(x_i - c)f_i = 0 \iff \sum_{i=1}^{k} x_i f_i = c\sum_{i=1}^{k} f_i \iff c = \bar{x}$$

□

Esta propiedad significa que la varianza es la medida cuadrática de dispersión óptima, en el sentido de que la media de las desviaciones cuadráticas respecto a un promedio toma el valor más pequeño cuando se toma como referencia la media aritmética.

## La desviación típica

La varianza viene expresada en las mismas unidades que la variable pero elevadas al cuadrado, lo que supone una dificultad para su interpretación. Para salvar este inconveniente introduciremos otra medida que venga expresada en las mismas unidades que la variable.

**Definición 3.6.** Llamamos *desviación típica* o *desviación estándar*, que denotamos por $S_X$, a la raíz cuadrada de la varianza tomada con signo positivo:

$$S_X = +\sqrt{S_X^2} \tag{3.1.5}$$

El término "desviación típica" fue introducido por Karl Pearson (1857-1936), considerado padre de la ciencia de la Estadística en el siglo XX. Entre las numerosas aportaciones de Pearson podemos destacar algunos conceptos que estudiaremos posteriormente, como su coeficiente de asimetría, el coeficiente chi-cuadrado y la distribución asociada que tiene gran importancia en los análisis inferenciales.

**Propiedad 3.5.** *La desviación típica satisface las siguientes propiedades:*

*a)* $S_X \geq 0$.

*b) La desviación típica es invariante ante cambios de origen:* $S_{X+c} = S_X$.

*c) La desviación típica se modifica ante cambios de escala:* $S_{cX} = |c|\, S_X$.

*Demostración.* Estas propiedades se derivan de forma inmediata de las propiedades de la varianza 3.1 y 3.3. □

Determinada la desviación típica de una distribución se puede analizar el riesgo que supone sintetizar la información mediante $\bar{x}$.

En general, para cualquier distribución, se puede afirmar que al menos el 75 % de los datos se desvían de la media como mucho 2 desviaciones típicas; en otros términos, al menos el 75 % de los valores se encuentran en el intervalo $[\bar{x} - 2S_X, \bar{x} + 2S_X]$. Asimismo, más el 88,8 % lo están en el intervalo $[\bar{x} - 3S_X, \bar{x} + 3S_X]$. Estos resultados se derivan de un teorema, conocido como desigualdad de Chebyshev.

Para una *distribución normal*,, según refleja la figura 3.2, es posible medir aún con mayor precisión el porcentaje de valores situados en intervalos determinados por cierto número de desviaciones típicas.

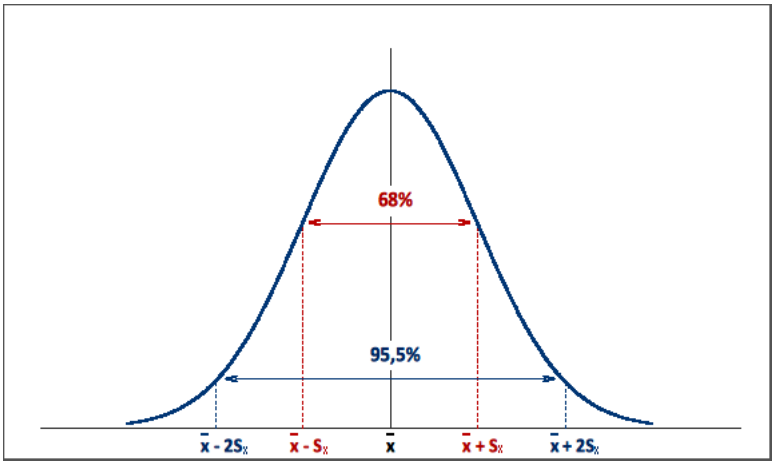



Figura 3.2: Distribución normal. Intervalos en torno a la media

En términos generales, decimos que un valor es atípico si se trata de una observación extrema, es decir, distante del resto de los datos. Ahora bien, teniendo en cuenta los comentarios anteriores, esta idea podemos precisarla del modo siguiente: un valor $x$ es atípico si su desviación respecto a la media de la distribución es superior a cierto número $k$ de desviaciones típicas, esto es, $|x - \bar{x}| > kS_X$, donde suele ser habitual tomar $k = 3$ o $k = 4$.

## 3.2. Medidas de dispersión relativas

El principal objetivo que perseguimos con las medidas de dispersión es analizar la representatividad de los promedios, finalidad que no podemos alcanzar en general con las medidas introducidas en el apartado anterior porque sus resultados dependen de las unidades de medida.

Consideremos dos embalses con tamaños muy diferentes, tales que el nivel medio de agua sea de 6 y 500 millones de metros cúbicos. Supongamos que en ambos casos el nivel de agua embalsada presenta la misma desviación típica, por ejemplo, 3 millones de metros cúbicos. A pesar de esta coincidencia, dado que la cantidad de agua embalsada será significativamente menor en el embalse pequeño que en el grande, no tiene la misma importancia que la dispersión sea de 3 millones de metros cúbicos respecto a un nivel medio de agua de 6 millones que con respecto a una media de 500 millones. Por tanto, el valor de la desviación típica no permite hacer afirmaciones sobre el nivel de dispersión ya que es necesario tener en cuenta las unidades de medida de la variable y el orden de magnitud de los valores.

Para solventar estos inconvenientes se emplearán una serie de coeficientes que miden la dispersión en términos relativos y que, como consecuencia, serán medidas adimensionales (es decir, que no están afectadas por las unidades de la variable), con las que se podrá comparar la representatividad de los promedios.

### 3.2.1. Coeficientes de variación basados en desviaciones cuadráticas

La idea de un coeficiente de dispersión que permita cuantificar la representatividad de un promedio deberá basarse en el promedio considerado y en una medida de dispersión absoluta que incluya en su definición a dicho promedio. Así, si queremos determinar la representatividad de $P$ (siendo $P \neq 0$), consideraremos la dispersión respecto a este parámetro, para lo cual podemos tomar la media de las desviaciones cuadráticas respecto a $P$.

**Definición 3.7.** Llamamos *coeficiente de variación* respecto a un promedio $P$, que denotamos por $V_P$, al resultado de la siguiente expresión:

$$V_P = \frac{\sqrt{D_P^2}}{|\,P\,|} = \frac{\sqrt{\sum_{i=1}^{k}(x_i - P)^2 f_i}}{|\,P\,|}$$

El resultado de esta fórmula carece de unidades puesto que numerador y denominador vienen expresados ambos en las mismas unidades que la variable, es decir, se trata de una medida adimensional.

El coeficiente $V_P$ puede utilizarse con cualquier tipo de promedio $P$, siempre que $P \neq 0$, pues si el promedio es nulo dicho coeficiente no está definido.

Si para una distribución el coeficiente de variación toma el valor 0,25 interpretaremos que la dispersión en torno al promedio considerado representa el 25 % del valor del mismo.

El coeficiente de variación toma solamente valores no negativos. Si una distribución tiene dispersión nula, el numerador del coeficiente será nulo y, por tanto, su valor también; por el contrario, dado un promedio concreto, cuanto mayor sea la dispersión absoluta, mayor será el coeficiente de variación.

Un caso especialmente interesante del coeficiente de variación $V_P$, por ser de uso generalizado, corresponde a la media aritmética y a la desviación típica como medida de dispersión.

**Definición 3.8.** Llamamos *coeficiente de variación de Pearson* al cociente entre la desviación típica y el valor absoluto de la media aritmética de una distribución:

$$V_{\bar{x}} = \frac{S_X}{\mid \bar{x} \mid} \tag{3.2.1}$$

**Propiedad 3.6.** *El coeficiente de variación respecto a un promedio $P$ es invariante ante cambios de escala.*

*El coeficiente de variación respecto a un promedio $P$ se modifica ante cambios de origen.*

*Demostración.* Representemos por $(x_i, f_i)$ la distribución inicial y por $(x'_i, f_i)$ la distribución resultante de la transformación proporcional, donde $x'_i = cx_i$. Considerado un promedio cualquiera $P$ de la variable $X$, sabemos que ante un cambio de escala su valor se transforma en $P' = cP$. Por otra parte, teniendo en cuenta la propiedad de la desviación cuadrática media ante cambios de escala, se obtiene que:

$$V_{P'} = \frac{\sqrt{D^2_{P'}}}{\mid P' \mid} = \frac{\sqrt{c^2 D^2_P}}{\mid cP \mid} = \frac{\sqrt{D^2_P}}{\mid P \mid} = V_P$$

Análogamente, sea $(x'_i, f_i)$ la distribución resultante de la transformación lineal, siendo $x'_i = x_i + c$. Considerado un promedio cualquiera $P$ de la variable $X$, sabemos que ante un cambio de origen su valor se transforma en $P' = P + c$. Si además tenemos en cuenta que la desviación cuadrática media respecto a $P$ es invariante ante cambios de origen, se deduce que:

$$V_{P'} = \frac{\sqrt{D^2_{P'}}}{\mid P' \mid} = \frac{\sqrt{D^2_P}}{\mid P + c \mid} \neq \frac{\sqrt{D^2_P}}{\mid P \mid} \quad \Rightarrow \quad V_{P'} \neq V_P$$

□

### 3.2.2. Coeficientes de variación basados en desviaciones absolutas

El procedimiento aplicado para definir los coeficientes de variación anteriores puede utilizarse también para construir otros coeficientes, tomando como medida de dispersión absoluta la desviación absoluta media respecto al promedio $P$ considerado. Bajo este nuevo planteamiento el coeficiente de variación vendrá dado por la siguiente expresión:

$$V_P = \frac{D_P}{|P|} = \frac{\sum_{i=1}^{k} |x_i - P| f_i}{|P|}$$

Este coeficiente puede utilizarse con cualquier tipo de promedio, y las interpretaciones comentadas en el apartado anterior para los coeficientes basados en desviaciones cuadráticas siguen siendo válidas.

### 3.2.3. Representatividad de los promedios

Los coeficientes de variación son medidas de dispersión que permiten analizar la representatividad de los promedios, bien sea para comparar la representatividad de varios promedios de una misma distribución, o para comparar la representatividad de un promedio en varias distribuciones.

En el tema 2 hemos analizado las ventajas e inconvenientes de los promedios. Ahora bien, si a priori fuera indiferente aplicar cualquiera de estas medidas, entonces podemos utilizar los coeficientes de variación para determinar el promedio más representativo: será aquel cuyo coeficiente de variación presente menor valor.

Así, dada una distribución, si calculamos los coeficientes de variación respecto a algunos promedios, podemos comparar el grado de representatividad de los mismos; si para la media se obtiene 0,2 y para la mediana 0,4 significa que la dispersión relativa respecto a la media es del 20 %, mientras que para la mediana tal dispersión es del 40 %; por tanto, la media representará mejor el conjunto de datos de la distribución que la mediana.

Razonablemente, para que los resultados anteriores puedan ser comparabIes, las medidas con que hayan sido obtenidos deben ser homogéneas; esto es, si hemos trabajado con coeficientes de tipo cuadrático, éstos deben ser considerados en todos los indicadores de dispersión empleados y lo mismo en el caso de medidas de tipo absoluto.

De la misma forma, para comparar la representatividad de un promedio en varias distribuciones, tendremos que calcular un coeficiente de variación respecto a ese promedio en cada una de ellas, siendo más representativo en aquélla para la que dicho coeficiente presente menor valor.

Hasta aquí hemos contemplado que la representatividad de un promedio depende de la dispersión. Si bien es cierto que éste es un factor determinante, es conveniente señalar también la importancia del número de datos en dicho análisis. Supongamos, por ejemplo, que al estudiar los salarios de los trabajadores en dos empresas, se obtiene el mismo resultado para el coeficiente de variación de Pearson. Si el número de trabajadores es significativamente distinto, el salario medio resultará más robusto en la empresa grande.

## 3.3. Variable tipificada

Los coeficientes de variación sirven para comparar la representatividad de los promedios; sin embargo, hasta el momento, no hemos diseñado ningún instrumento que nos permita comparar valores cualesquiera de dos distribuciones.

Para poder comparar dos distribuciones, o más concretamente algunos de sus valores, éstos deben trasladarse a una misma escala. Para ello debemos tener en cuenta

una medida de posición central y otra de dispersión; de esa forma, si calculamos la distancia de cada valor al promedio que se tome como referencia y luego dividimos entre una medida de dispersión respecto a dicho promedio, obtendremos la posición relativa del mismo. Existen numerosas formas de reducir a una misma escala las variables, no obstante la más habitual se basa en tomar como referencia la media aritmética y la desviación típica.

**Definición 3.9.** Llamamos *variable tipificada* a aquélla que tiene media cero y varianza uno.

Dada una variable estadística $X$, podemos obtener una tipificación de la misma mediante la siguiente transformación:

$$Z = \frac{X - \bar{x}}{S_X} \tag{3.3.1}$$

**Propiedad 3.7.** *La nueva variable $Z$ es una variable tipificada, esto es, su media es 0 y su varianza es 1.*

*Demostración.* En efecto, de las propiedades de cambio de origen y de escala de la media aritmética se deriva de forma inmediata que:

$$\bar{z} = \frac{\bar{x} - \bar{x}}{S_X} = 0$$

Análogamente, teniendo en cuenta el comportamiento de la varianza ante cambios de origen y de escala, se tiene que:

$$S_Z^2 = \frac{1}{S_X^2} S_X^2 = 1$$

□

El valor tipificado indica el número de desviaciones típicas que está por encima o por debajo de la media un valor determinado. Una vez tipificados, los valores superiores a la media tienen signo positivo mientras que el signo de los inferiores será negativo.

Retomemos el ejemplo de los embalses y supongamos ahora que en el embalse pequeño el nivel medio de agua es de 6 millones de metros cúbicos con una desviación típica de 2 millones, mientras que en el grande el nivel medio es de 500 millones con desviación típica de 150 millones. Si la primavera ha sido especialmente lluviosa y al finalizar esa estación el nivel de agua alcanzado fue de 10 millones de metros cúbicos en el embalse pequeño y de 665 millones en el grande, para determinar a cuál de los dos embalses ha beneficiado más la abundancia de lluvias debemos expresar el nivel de agua embalsado en la misma escala, es decir, debemos tipificar los valores. Para el embalse pequeño el valor tipificado resultante es $z = 2$ mientras que para el grande es $z = 1,1$. Dado que corresponde al embalse pequeño un valor tipificado mayor, podemos concluir que, en términos relativos, la abundancia de lluvias ha beneficiado más al embalse pequeño.

## 3.4. Medidas de forma

Las medidas de forma hacen referencia a la asimetría y al apuntamiento o kurtosis, que son las principales características de la representación gráfica de una distribución.

**Definición 3.10.** Una distribución se dice *simétrica* si su representación gráfica lo es respecto a la perpendicular trazada por su valor central.

En otras palabras, si tomamos esa perpendicular como eje de simetría, la distribución es simétrica si hay el mismo número de valores a ambos lados del eje, equidistantes dos a dos, y cada par de valores equidistantes tienen la misma frecuencia.

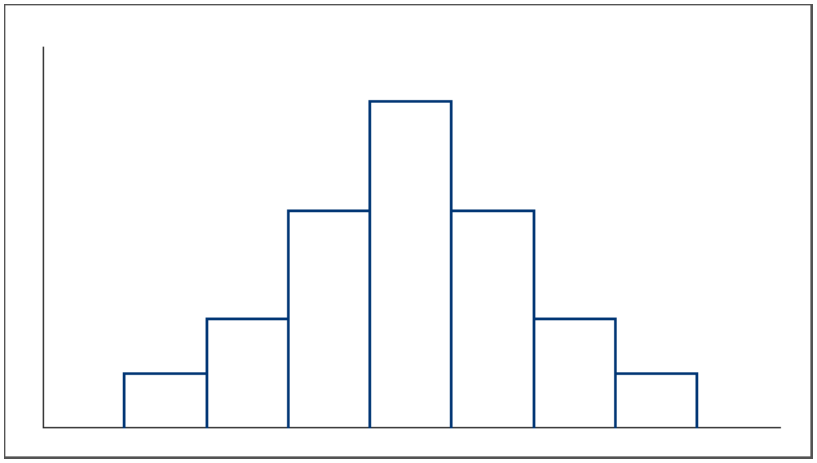


Figura 3.3: Distribución simétrica

Se denomina *asimetría* a la falta de simetría en la distribución. La asimetría puede ser *positiva o a la derecha* y *negativa o a la izquierda*, según que sea en la cola derecha o izquierda del eje donde se encuentre un mayor peso de la distribución.

En el ámbito económico encontramos magnitudes cuyo comportamiento es claramente asimétrico. En general, variables que representan "riqueza" (tales como renta, salario, beneficios, ...) suelen presentar asimetría positiva o a la derecha.

En una distribución simétrica unimodal todos los valores centrales (media, moda y mediana) coinciden. Dada una distribución arbitraria, partiendo de la posición de los promedios se puede indicar, aunque no siempre, el tipo de asimetría. Así, generalmente, se cumple que si $\bar{x} - Mo > 0$ la distribución es asimétrica a la derecha mientras que si $\bar{x} - Mo < 0$ es asimétrica a la izquierda.

En el caso de distribuciones unimodales, es frecuente que la mediana esté comprendida entre la media y la moda, cumpliéndose entonces que $Mo < Me < \bar{x}$ si la distribución es asimétrica a la derecha y $\bar{x} < Me < Mo$ cuando es asimétrica a la izquierda.

**Definición 3.11.** El *coeficiente de asimetría de Pearson* se define mediante la expre-

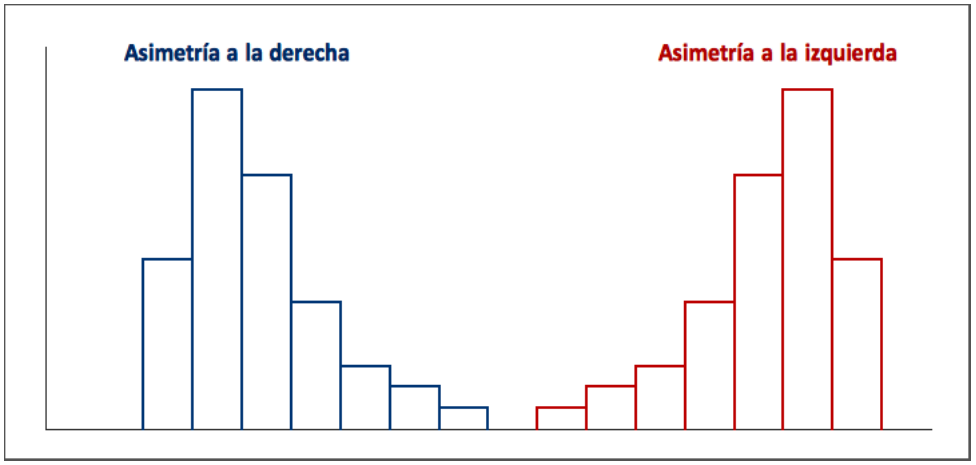



Figura 3.4: Tipos de asimetría

sión:

$$A_P = \frac{\bar{x} - Mo}{S_X} \tag{3.4.1}$$

cuyo resultado es nulo para distribuciones simétricas, positivo en el caso de distribuciones con asimetría a la derecha y negativo para distribuciones con asimetría a la izquierda.

Las principales ventajas de esta medida son su facilidad de cálculo e interpretación. Sin embargo, dado que el análisis de la forma se basa en la comparación de promedios, puede ser poco adecuada en el caso de asimetrías leves.

Otra alternativa para determinar el grado de asimetría que presenta una distribución es cuantificar a través de alguna medida el mayor o menor agrupamiento de los datos a cada lado de uno de sus promedios, a partir de las desviaciones $(x_i - P)$. En particular, el promedio de referencia más habitual es la media aritmética.

Ahora bien, dado que la suma de las desviaciones respecto a la media es siempre igual a cero, debemos tomar alguna potencia de $(x_i - \bar{x})$. La alternativa de considerar la suma de los cuadrados no es adecuada pues interesa tener en cuenta el signo de las desviaciones. Así pues, la posibilidad más sencilla consiste en utilizar la suma de las desviaciones elevadas al cubo. Si el resultado es positivo serán las desviaciones positivas las que tienen mayor peso (asimetría a la derecha), al revés si es negativo.

Nuestro objetivo es determinar el grado de asimetría de una distribución y comparar la asimetría de distintas distribuciones, por lo tanto es necesario construir una medida relativa, es decir, que no se vea afectada por las unidades de la variable. Para normalizar la expresión anterior tendremos que dividir por una medida de dispersión respecto a la media.

**Definición 3.12.** Dada una variable $X$, con distribución de frecuencias $(x_i, f_i)$, se define el *coeficiente de asimetría de Fisher*, que denotamos por $g_1$, como el valor de la siguiente expresión:

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^3 f_i}{S_X^3} \tag{3.4.2}$$

Si $g_1 = 0$, la distribución es simétrica; si $g_1 > 0$, la distribución es asimétrica a la derecha y si $g_1 < 0$ es asimétrica a la izquierda. Además, cuanto mayor sea la cuantía del coeficiente más marcada será la asimetría en el sentido indicado por su signo.

Sir Ronald Aylmer Fisher (1890-1962) es autor de aportaciones estadísticas de gran trascendencia, que contribuyeron en gran medida al desarrollo de esta disciplina. Entre ellas destacan los coeficientes de asimetría que llevan su nombre, el método de máxima verosimilitud y el desarrollo de la inferencia estadística.

El análisis del *apuntamiento o kurtosis* tiene por objeto básicamente comparar el de la distribución con el de la curva normal (de igual media y desviación típica), que se toma como patrón. Este tipo de análisis se aplica sobre todo a distribuciones unimodales de tipo campaniforme.

Si tratamos de medir el apuntamiento de una distribución debemos tener en cuenta que el mismo está inversamente relacionado con la dispersión. Además, ahora no interesa que se compensen desviaciones positivas y negativas, por lo tanto una alternativa puede consistir en sumar las desviaciones elevadas a una potencia par. Igualmente, como la medida debe ser adimensional, el resultado anterior debe ser normalizado mediante una medida de dispersión. Finalmente, se resta el valor 3 para asignar valor nulo al modelo normal, que se toma como referencia.

**Definición 3.13.** Dada una variable $X$, con distribución de frecuencias $(x_i, f_i)$, se define el *coeficiente de apuntamiento de Fisher*, que denotamos por $g_2$, como el valor de la siguiente expresión:

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 f_i}{S_X^4} - 3 \tag{3.4.3}$$

Si $g_2 = 0$ el grado de apuntamiento de la distribución coincide con el de la normal y se dice que la distribución es mesocúrtica. Si $g_2 > 0$, la distribución es más apuntada que la normal (leptocúrtica) y si $g_2 < 0$ es menos apuntada que la normal (platicúrtica).

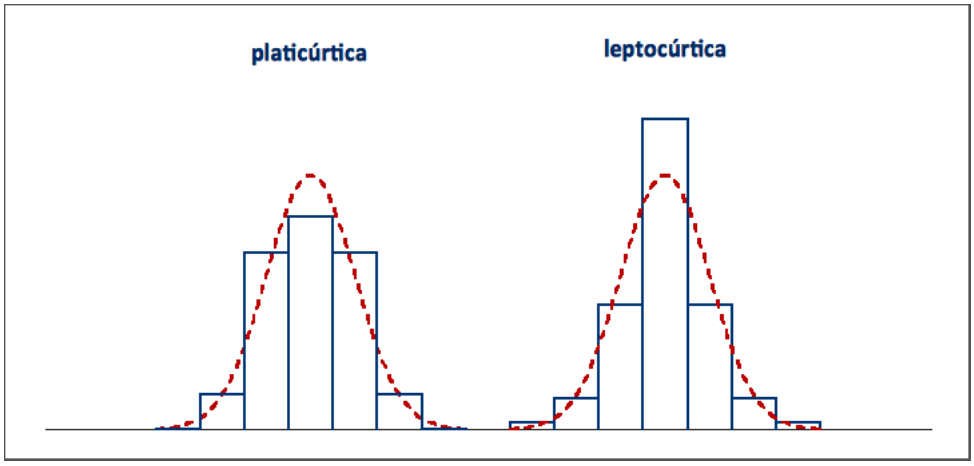



Figura 3.5: Tipos de apuntamiento

# 4 Desigualdad y pobreza

## 4.1. La desigualdad económica

La desigualdad en el reparto de la riqueza es uno de los problemas económicos más importantes en las sociedades actuales, por lo que su estudio viene siendo objeto de atención por parte de los gobiernos nacionales y también por distintos organismos internacionales. Los estudios orientados a conocer la situación de los más desfavorecidos permiten analizar la incidencia de la pobreza. A su vez, el estudio de la desigualdad en la distribución de la renta entre personas u hogares permite analizar el grado de bienestar de los ciudadanos de un país. Asimismo, el estudio de la desigualdad en el reparto de la riqueza entre países o regiones presenta gran interés de cara al establecimiento de medidas de política económica por parte de distintos organismos internacionales (Unión Europea, Fondo Monetario Internacional, Banco Mundial, ...) para tratar de corregir los desequilibrios territoriales.

> El Informe sobre desarrollo humano elaborado anualmente por Naciones Unidas, los estudios sobre pobreza y desigualdad de género realizados por Social Watch y los estudios sobre desigualdad en la distribución del ingreso del Banco Mundial son claros exponentes de la importancia que tiene la cuantificación de la desigualdad económica y la pobreza.
>
> El Índice de Desarrollo Humano (IDH), elaborado por el Programa de Desarrollo de Naciones Unidas desde 1970, es un indicador basado en tres dimensiones: esperanza de vida, educación e ingresos, y se calcula como media geométrica de los índices de las tres componentes.

Los análisis de la distribución de la renta no son estrictamente económicos sino que tienen también implicaciones sociológicas, políticas, éticas ... . No obstante, desde una perspectiva estadística, nos ocuparemos únicamente de cuantificar la desigualdad para lo cual utilizaremos medidas de desigualdad o concentración, evitando llevar a cabo juicios normativos sobre la equidad o justicia de las situaciones consideradas.

> Actualmente los análisis distributivos ocupan un papel muy relevante en los estudios económicos. De hecho, en la propia evolución de la teoría económica puede apreciarse que. frente a etapas anteriores en las que el objetivo prioritario era el crecimiento, a lo largo de la segunda mitad del siglo XX se fue consolidando una importante corriente dedicada al análisis sobre bienestar y equidad distributiva, en la que pasan a ocupar un papel dominante los análisis del desarrollo económico, término que incorpora al concepto de crecimiento connotaciones sobre equidad en la distribución.
>
> Las aproximaciones conceptuales al estudio de la desigualdad -cuyos antecedentes pueden encontrarse en algunos economistas clásicos como Adam Smith o Ricardo- aparecen con las obras de Lorenz (1880-1962) y Gini (1884-1965).

Más recientemente cabe destacar las importantes aportaciones del profesor Amartya Sen (India 1933-), a quien fue concedido el premio Nobel de Economía en el año 1998 por sus contribuciones a la economía del bienestar. En particular, el profesor Sen ha realizado valiosas aportaciones en el campo de la medición de la desigualdad y la pobreza y su conexión con las funciones de bienestar social.

En el contexto actual de la sociedad del conocimiento, resulta indudable el impacto que tienen las nuevas tecnologías de la información y comunicación sobre el crecimiento y el desarrollo, motivo por el cual el concepto de desigualdad ha rebasado el ámbito puramente económico para extenderse a otros ámbitos. En particular, el análisis de la desigualdad en cuanto a acceso y uso de las TIC ha dado lugar a lo que habitualmente se conoce como "brecha digital", término que según la OCDE, hace referencia a "la brecha entre individuos, hogares, negocios y áreas geográficas a diferentes niveles socioeconómicos, en relación tanto a sus oportunidades de acceso a las TIC como al uso de Internet para una amplia variedad de actividades".

En términos generales, las medidas de desigualdad o concentración indican el grado de desigualdad en el reparto del valor total de una variable entre los elementos que componen la población.

**Ejemplo 4.1.** Supongamos que un millonario deja una herencia de 110 millones de euros a repartir entre un total de 11 herederos. Puede resultar interesante conocer hasta qué punto la distribución que el fallecido hace de su fortuna beneficia a todos por igual, o bien discrimina a algunos parientes en favor de otros.

Una primera opción consistiría en repartir la herencia a partes iguales, lo que supondría asignar 10 millones de euros a cada uno de sus 11 herederos. Se trata entonces de una situación de equidistribución o concentración mínima. En el extremo opuesto, si uno solo de los herederos recibiese toda la herencia (110 millones) y los otros diez no heredasen nada, entonces el reparto presentaría máxima desigualdad o concentración.

Entre estos dos casos extremos, se pueden plantear muy diversas situaciones intermedias que implicarán diferentes grados de desigualdad en el reparto de la herencia. Así, un posible reparto podría ser el siguiente: 1, 1, 1, 1, 1, 1, 1, 2, 2, 2 y 97.

En este caso uno de los herederos recibe una cantidad -97 millones- muy superior a los restantes, que reciben 1 o 2 millones, es decir, se aprecia que existe cierto desequilibrio en el reparto de la herencia, aunque la situación no es de desigualdad máxima. Se trata por tanto de definir indicadores que nos permitan conocer el grado de desequilibrio que existe en el reparto.

Una primera aproximación al estudio de la desigualdad viene dada por la dispersión entre los cuantiles de la distribución. En particular, dicho análisis suele basarse habitualmente en el cociente entre el noveno y el primer decil, que cuantifica en qué medida la renta media del 10 % más rico de la población estudiada supera a la renta del 10 % más pobre. Como consecuencia, cuanto más elevado sea este resultado más desigualdad existirá en la distribución de renta.

Así, en el ejemplo anterior de las herencias los deciles primero y noveno serían respectivamente $D_1 = 1$ y $D_9 = 2$ con lo cual el cociente sería 2, indicando un elevado nivel de desigualdad (la cantidad heredada por el 10 % más beneficiado duplica la del 10 % menos favorecido).

## 4.2. La curva de Lorenz y el índice de Gini

Las medidas de desigualdad o concentración son indicadores que resumen el desequilibrio existente en el reparto del valor total de una magnitud económica que genéricamente denominamos riqueza o renta. Dado que la situación económica de las personas, empresas, etc. se aproxima habitualmente a través de salarios, rentas familiares, beneficios, ingresos ..., en la práctica aplicaremos dichas medidas a este tipo de variables económicas.

### 4.2.1. La curva de Lorenz

Una primera aproximación al estudio de la concentración viene dada por una representación gráfica denominada *curva de Lorenz*. Supongamos un conjunto de $N$ individuos, a cada uno de los cuales corresponde una renta $x_i$, y consideremos la distribución de rentas $(X)$ con los valores ordenados en sentido creciente, esto es, $x_1 \leq x_2 \leq \cdots \leq x_N$.

Para cada $i = 1, \ldots, N$ se definen los siguientes ratios:

- La *proporción acumulada de rentistas* $p_i$, como:

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

  En términos generales, el valor de $p_i$ nos indica la proporción que suponen respecto al total los $i$ rentistas que perciben rentas más pequeñas, esto es, cuya renta es menor o igual que $x_i$.

- La *proporción acumulada de rentas* $q_i$, dada por la siguiente expresión:

$$q_i = \frac{A_i}{A_N}$$

  donde $A_i = \sum_{j=1}^{i} x_j$ es la renta acumulada por los $i$ primeros individuos y $A_N = \sum_{j=1}^{N} x_j$ es la renta total.

  En general, el valor de $q_i$ nos indica la proporción sobre el valor total de la renta acumulada por los $i$ primeros individuos, es decir, recoge la participación que el grupo formado por los $i$ individuos con rentas inferiores o iguales a $x_i$ ($i$ individuos menos ricos) tiene sobre el valor total de la renta. Por su propio significado, la parte de renta $q_i$ acumulada por la proporción $p_i$ de individuos menos ricos nunca superará el valor $p_i$.

En el ejemplo 4.1 se obtiene $p_7 = 63,6\,\%, q_7 = 6,4\,\%$, lo que significa que el 6,4 % de la herencia es recibido conjuntamente por los 7 herederos que reciben individualmente las cantidades más pequeñas (1 millón cada uno), y que constituyen el 63,6 % de la

población. De forma análoga resulta $p_9 = 90,9\,\%, q_9 = 11,8\,\%$, lo que nos indica que en conjunto los herederos cuya herencia individual es menor o igual que 2 millones (90,9 % de la población) reciben un 11,8 % del valor total de la herencia.

**Propiedad 4.1.** *Los ratios $p_i$ y $q_i$ verifican la siguiente desigualdad:*

$$q_i \leq p_i,\ i = 1, \ldots, N$$

Además, por definición, se cumplirá que $p_N = q_N = 1$ (100 %) puesto que el 100 % de los rentistas percibirá el 100 % de la renta total. Asimismo, mayores diferencias entre ambas proporciones reflejarán mayores desequilibrios en el reparto.

**Definición 4.1.** Considerado un sistema de ejes cartesianos, se denomina *curva de Lorenz* a la línea que partiendo del origen de coordenadas une los pares $(p_i, q_i), i = 1, \ldots, N$.

Como consecuencia de la propiedad 4.1, la *curva de Lorenz* se situará siempre por debajo de la diagonal del cuadrado de lado unidad (recta $p = q$) y cuanto mayores sean las diferencias $p_i - q_i$ la curva estará más alejada de la diagonal.

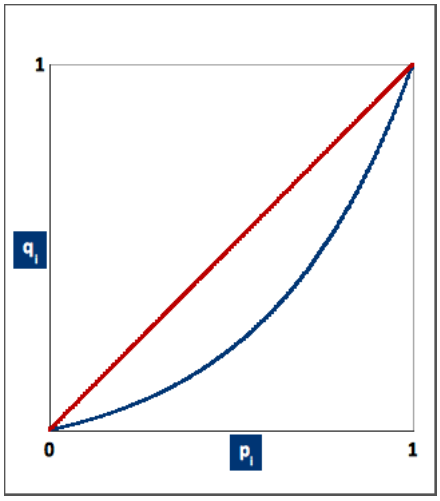



Figura 4.1: Curva de Lorenz

- En una situación de equidistribución (concentración mínima) todos los rentistas perciben la misma renta, por lo que al acumular resulta evidente que el 10 % de los rentistas percibirán el 10 % de la renta, el 20 % de los rentistas percibirán el 20 % de la renta etc., es decir, en tal caso se cumple que:

  $$q_i = p_i,\ i = 1, \ldots, N$$

  y, en consecuencia, la curva de Lorenz coincide con la diagonal del cuadrado de lado unidad (bisectriz del primer cuadrante), a la que por este motivo se conoce como *recta de equidistribución.*

- En el otro extremo, es decir, en situación de máxima desigualdad, la renta se concentraría en un único individuo, de modo que $q_i = 0$, $i = 1, \ldots, N-1$. Por consiguiente, en este caso la curva de Lorenz se aproximará a la curva formada por los lados OA y AB del cuadrado.

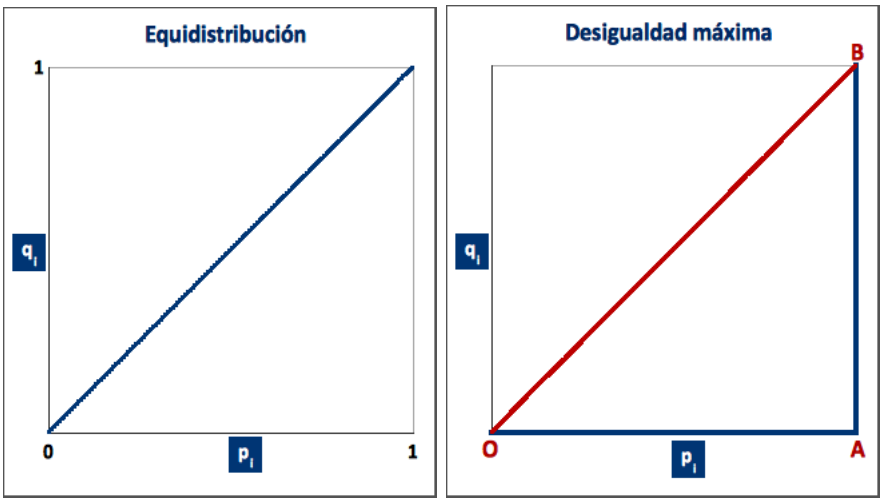



Figura 4.2: Concentración mínima y máxima

Además de analizar la desigualdad de una distribución, otro objetivo de interés es la comparación de la desigualdad de diferentes situaciones distributivas. Partiendo de las respectivas curvas de Lorenz, en casos como los representados en la figura 4.3 se dice que la distribución A es más igualitaria que B ya que en todo su recorrido la curva correspondiente a la distribución A se sitúa más cerca de la diagonal, lo que significa que para cualquier proporción de rentistas $p_i$ la proporción de renta acumulada $q_i$ es superior en el caso A que en B. En este tipo de situaciones, se dice que *A domina a B en el sentido de Lorenz*.

Ahora bien, existirán muchas situaciones en que el criterio de Lorenz no puede ser aplicado debido a que las curvas se entrecruzan y por lo tanto no es posible establecer entre ellas relaciones de dominación.

### 4.2.2. El índice de Gini

Además de representar las situaciones distributivas, la curva de Lorenz permite construir una medida, basada en las diferencias $p_i - q_i$, que resume el nivel de desigualdad o concentración de la distribución. Esta medida, conocida como índice de Gini, compara por cociente el área situada entre la recta de equidistribución y la curva de Lorenz de la distribución con el área correspondiente a la situación de máxima desigualdad (área del triángulo OAB):

**Definición 4.2.** Para una distribución de $N$ rentas $(X)$ con los valores ordenados en

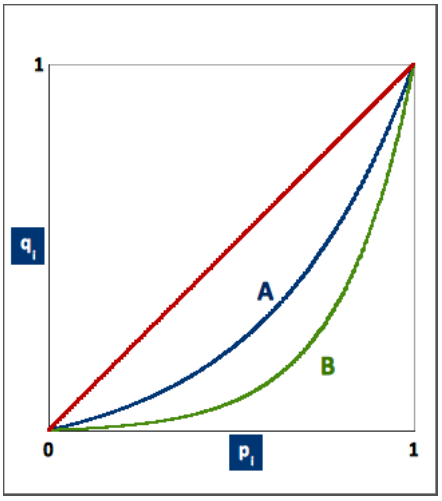



Figura 4.3: Criterio de dominación de Lorenz

sentido creciente, esto es, $x_1 \leq x_2 \leq \ldots \leq x_N$, el *índice de Gini*, que denotamos por $I_G$, viene dado por la siguiente expresión:

$$I_G = \frac{\sum_{i=1}^{N-1} (p_i - q_i)}{\sum_{i=1}^{N-1} p_i} \tag{4.2.1}$$

En su formulación original, el índice de Gini se basaba en los valores absolutos de las diferencias entre todos los pares de rentas $|x_i - x_j|$. Sin embargo, en la búsqueda de una interpretación gráfica para este indicador, se llegó posteriormente a establecer su relación con la curva de Lorenz y obtener su expresión en términos de las diferencias $p_i - q_i$, formulación que resulta más intuitiva.

El índice de Gini es uno de los indicadores aplicados habitualmente en los estudios de desigualdad realizados por distintos organismos internacionales tales como Banco Mundial, Naciones Unidas o Social Watch.

Asimismo, en la actualidad se aplica también al estudio de la desigualdad en otros ámbitos como, por ejemplo, la cuantificación de la brecha digital. Algunos organismos como Naciones Unidas utilizan la curva de Lorenz y el índice de Gini para analizar el nivel de desequilibrio en el reparto de diversas magnitudes relacionadas con las TIC (acceso a Internet, acceso mediante banda ancha, uso de teléfono móvil, ...) entre la población mundial.

**Propiedad 4.2.** *El índice de Gini está acotado entre 0 y 1, es decir, se cumple que:*

$$0 \leq I_G \leq 1$$

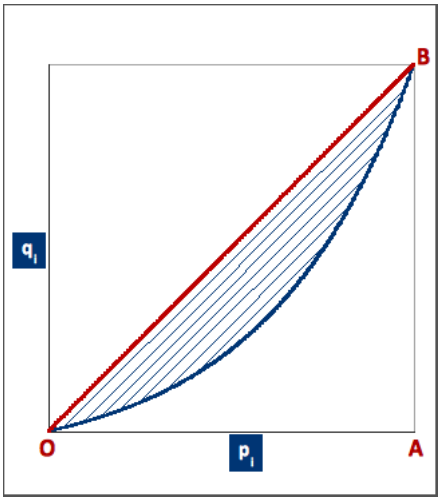



Figura 4.4: Área de concentración

*Además, el índice de Gini toma el valor mínimo (0) en caso de equidistribución (mínima concentración) y toma el valor máximo (1) en situación de máxima desigualdad o concentración.*

*Demostración.* Dado que $0 \leq p_i - q_i \leq p_i,\ i = 1, \ldots, N-1$, se verifican las siguientes desigualdades:

$$0 \leq \sum_{i=1}^{N-1} (p_i - q_i) \leq \sum_{i=1}^{N-1} p_i$$

de donde se deriva de forma inmediata la acotación propuesta.

Por otra parte, en situación de equidistribución se cumple que $q_i = p_i,\ i = 1, \ldots, N$, con lo cual el numerador de la expresión 4.2.1 toma el valor 0 y, en consecuencia, el índice también vale 0.

A su vez, en caso de desigualdad máxima se cumple que $q_i = 0, i = 1, \ldots, N-1$, con lo cual el numerador y el denominador de la expresión son iguales y, en consecuencia, el índice toma el valor 1. □

En general, para situaciones distributivas intermedias, el valor del índice de Gini estará más próximo a 1 cuanto mayor sea el nivel de concentración o desigualdad y viceversa.

En la ilustración de la herencia, el resultado del indicador para el reparto analizado es $I_G = 0,889$, lo que permite concluir que el grado de desigualdad de dicho reparto es elevado.

Además de las propiedades básicas anteriores, el índice de Gini satisface también otras propiedades que son especialmente relevantes por estar directamente relacionadas con los efectos que pueden tener sobre el nivel de desigualdad distintas políticas distributivas de la renta.

**Propiedad 4.3.** *El índice de Gini es invariante ante cambios proporcionales, es decir, si una variable X sufre una transformación proporcional, siendo la nueva variable X'=cX donde c es constante, se verifica que el índice de Gini toma el mismo valor en ambas situaciones:* $I_{X'} = I_X$.

*Demostración.* Designemos por $p'_i$ y $q'_i$ los ratios asociados a la nueva variable $X'$.

Los cambios de escala no afectan a las frecuencias, por tanto el ratio $p'_i$ es idéntico a $p_i,\ i = 1, \ldots, N$.

Por otra parte, dado que $x'_i = cx_i$, la renta acumulada $A'_i$ asociada a $x'_i$ está relacionada con $A_i$, renta acumulada correspondiente a $x_i$, a través de la siguiente expresión:

$$A'_i = \sum_{j=1}^{i} x'_j = c \sum_{j=1}^{i} x_j = cA_i$$

En consecuencia se obtiene que:

$$q'_i = \frac{A'_i}{A'_N} = \frac{cA_i}{cA_N} = q_i$$

Por tanto:

$$I_{X'} = \frac{\sum_{i=1}^{N-1} (p'_i - q'_i)}{\sum_{i=1}^{N-1} p'_i} = \frac{\sum_{i=1}^{N-1} (p_i - q_i)}{\sum_{i=1}^{N-1} p_i} = I_X$$

□

Esta propiedad significa que cambios proporcionales en la renta no alteran el nivel de desigualdad puesto que las participaciones relativas de los rentistas en el reparto se mantienen.

**Propiedad 4.4.** *El índice de Gini es decreciente ante aumentos constantes de renta, es decir, si todos los valores de la variable X aumentan en la misma cantidad* $c\,(c > 0)$*, siendo la nueva variable* $X' = X + c$*, se verifica que el índice de Gini asociado a la variable* $X'$ *es menor que el asociado a la variable* $X$*:* $I_{X'} < I_X$.

*Por el contrario, si la constante c es negativa* $(c < 0)$*, entonces el índice de Gini aumentará.*

*Demostración.* Designemos por $p'_i$ y $q'_i$ los ratios asociados a la nueva variable $X'$.

Los cambios de origen no afectan a las frecuencias, por tanto el ratio $p'_i$ es idéntico a $p_i,\ i = 1, \ldots, N$.

Por su parte, la renta acumulada $A'_i$ asociada a $x'_i$ está relacionada con $A_i$, renta acumulada correspondiente a $x_i$, a través de la siguiente expresión:

$$A'_i = \sum_{j=1}^{i} x'_j = \sum_{j=1}^{i} (x_j + c) = \sum_{j=1}^{i} x_j + ic = A_i + ic$$

de la cual se deriva que:

$$q'_i > q_i,\ i = 1, \ldots, N-1$$

Supongamos lo contrario, esto es, que $q'_i < q_i$; entonces se cumpliría que:

$$\frac{A'_i}{A'_N} < \frac{A_i}{A_N} \iff \frac{A_i + c\,i}{A_N + cN} < \frac{A_i}{A_N}$$

operando en la segunda desigualdad se llega a:

$$\frac{i}{N} < \frac{A_i}{A_N} \iff p_i < q_i$$

Sin embargo, por su propio significado, debe cumplirse siempre que $p_i \geq q_i$, y por tanto la hipótesis adoptada $q'_i < q_i$ no es válida. □

Esta propiedad pone de manifiesto el efecto positivo que tienen los aumentos constantes de renta ya que éstos benefician más a las rentas más bajas y, en consecuencia, el nivel de desigualdad relativa se reduce. Este tipo de efecto redistributivo se puede generalizar a situaciones en que se produce una transferencia de renta desde un individuo más privilegiado a otro en peor situación, siendo conocido como principio de transferencias progresivas o condición de Pigou-Dalton.

**Propiedad 4.5.** *El índice de Gini satisface la condición de Pigou-Dalton, es decir, si un individuo con renta $x_j$ transfiere una cantidad $\varepsilon$ a otro que percibe una renta inferior $x_i$, tal que $0 < \varepsilon < (x_j - x_i)/2$, se verifica que el índice de Gini asociado a la nueva situación distributiva ($Y$) es inferior al correspondiente a la situación inicial ($X$), esto es, $I_Y < I_X$.*

Además de las propiedades anteriores debemos destacar también ciertas limitaciones del índice de Gini. Una de ellas es que un mismo valor del indicador puede ir asociado a distribuciones con estructura muy diferente e interpretaciones distintas. De hecho, en cualquier situación de equidistribución el valor del índice es 0, con independencia del número de componentes de la población. Así, a modo de ejemplo, una vez que sabemos que en cierto mercado todas las empresas tienen igual cuota de participación (y existe, por tanto, concentración nula) resulta relevante la información sobre el número de empresas existentes, ya que podría ayudarnos a distinguir situaciones de oligopolio, un número reducido de empresas que se reparten la industria, de otras en que el número es muy elevado, tratándose entonces de una situación de competencia perfecta.

### Cálculo aproximado del índice de Gini

Hasta ahora, para obtener los ratios $p_i$ y $q_i$ hemos considerado la renta de cada uno de los $N$ individuos por separado calculando así un total de $N$ pares $(p_i, q_i)$. Si el número de rentas (N) es elevado, se suele partir de la tabla de frecuencias $(x_i, n_i)$, donde se recogen los diferentes valores de la renta ($k < N$), ordenados en sentido

creciente, $x_1 < x_2 < \ldots < x_k$ y sus respectivas frecuencias. Con respecto a este planteamiento conviene señalar que conduce a un cálculo "aproximado" del índice de Gini, pudiendo proporcionar resultados "bastante" alejados del valor exacto del indicador (la calidad de la aproximación depende del número de repeticiones).

Bajo este segundo planteamiento obtenemos un total de $k$ pares $(p_i, q_i)$, donde los ratios se calcularán como sigue:

$$p_i = \frac{N_i}{N}, \quad N_i = n_1 + n_2 + \ldots + n_i$$

es decir, $p_i$ representa la proporción de rentistas cuya renta es menor o igual que $x_i$.

Por otra parte:

$$q_i = \frac{A_i}{A_k}$$

donde $A_i = \sum_{j=1}^{i} x_j n_j$ es la renta acumulada por los $N_i$ primeros individuos y $A_k = \sum_{j=1}^{k} x_j n_j$ es la renta total.

En general, el valor de $q_i$ nos indica la proporción del valor total de la renta acumulada por los $N_i$ primeros individuos, es decir, recoge la participación que el grupo formado por los $N_i$ individuos con rentas inferiores o iguales a $x_i$ ($N_i$ individuos menos ricos) tiene sobre el valor total de la renta.

Finalmente, el índice de Gini se calcula como:

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} \tag{4.2.2}$$

Aplicando este procedimiento al reparto de la herencia, donde hay 7 herederos que reciben 1 millón y otros 3 que perciben 2 millones, el resultado aproximado del índice de Gini es 0,882 mientras que el valor exacto es 0,889.

> La utilización de algún procedimiento automático como una hoja de cálculo nos permitirá calcular con exactitud los ratios y el índice de Gini sin gran esfuerzo, con independencia del número de rentas observadas (N).

Finalmente, si la información disponible sobre la distribución de rentas viene dada en forma de tabla con datos agrupados en intervalos, entonces el cálculo del índice de Gini se basará en la marca de clase y la frecuencia de cada intervalo y, en consecuencia, se obtendrá una aproximación al verdadero valor del índice de Gini.

> Aunque, sin lugar a dudas, uno de los indicadores más utilizados en los estudios de desigualdad es el índice de Gini, en la literatura económica se han ido introduciendo otros tipos de medidas de desigualdad, basadas en funciones de bienestar social, en medidas de entropía, ... .

## 4.3. Medidas descomponibles

El fenómeno de la desigualdad, entendido como desequilibrio global en una población, aparece conectado con las distintas unidades que la componen. Cuando la población aparece dividida en unidades complementarias (estratos o subpoblaciones en general) resulta deseable obtener el valor de desigualdad poblacional a partir de los valores de desigualdad cuantificada en cada estrato.

Supongamos que la población está dividida en $p$ subpoblaciones. El índice $I(X)$ será aditivamente descomponible si la desigualdad global de la población se puede expresar como suma de dos componentes:

$$I(X) = \sum_{j=1}^{p} w_j I_j + I_0$$

donde $I_j$ representa el índice de desigualdad de la subpoblación $j$ y $w_j$ es un factor de ponderación, que depende del tamaño y de las rentas medias de las subpoblaciones y de la población; a su vez $I_0$ es la desigualdad entre las subpoblaciones, que es calculada considerando cada subpoblación como un individuo que percibe una renta igual a la media de su grupo. En síntesis, el primer componente resume las desigualdades dentro de las subpoblaciones (desigualdad intra-grupos) y el segundo recoge la desigualdad entre las diferentes subpoblaciones (desigualdad inter-grupos).

El requisito de descomponibilidad, deseable tanto desde el punto de vista conceptual como operativo, no es satisfecho por el índice de Gini, hecho que constituye uno de los principales inconvenientes de este indicador y que justifica el éxito alcanzado por la familia de medidas aditivamente descomponibles de aparición más reciente.

**Definición 4.3.** Dada una variable $X$, a la que genéricamente denominamos renta, con distribución $(x_i, f_i)$, en condiciones generales establecidas por D. Zagier (1983), una *medida de desigualdad aditivamente descomponible* viene dada por una expresión del tipo:

$$I_\beta(X) = \sum_{i=1}^{k} \Phi_\beta \left( \frac{x_i}{\bar{x}} \right) f_i$$

donde $\Phi_\beta(x)$ es una función definida para cada $\beta$ real como:

$$\Phi_\beta(x) = \begin{cases} x^\beta - 1, & \beta < 0 \\ -\log x, & \beta = 0 \\ 1 - x^\beta, & 0 < \beta < 1 \\ x \log x, & \beta = 1 \\ x^\beta - 1, & \beta > 1 \end{cases}$$

La expresión anterior toma siempre valores no negativos, aumentando a medida que la distribución presenta mayores desequilibrios en el reparto.

Entre los indicadores pertenecientes a esta familia destacaremos dos casos particulares, correspondientes a los valores del parámetro $\beta = 1$ (índice de Theil) y $\beta = -1$ (índice de desigualdad colectiva).

**Definición 4.4.** Dada una distribución de rentas $(x_i, f_i)$, con $x_i > 0$, $(i = 1, \ldots, k)$, se define el *índice de Theil*, que designamos por $T$, mediante la siguiente expresión:

$$T = \sum_{i=1}^{k} \left(\frac{x_i}{\bar{x}}\right) log \left(\frac{x_i}{\bar{x}}\right) f_i \tag{4.3.1}$$

Este indicador viene siendo aplicado desde hace años tanto en trabajos empíricos como en estadísticas oficiales. El origen de esta medida, propuesta por Theil en 1967, está en el concepto de entropía de Teoría de la Información.

**Definición 4.5.** Dada una distribución de rentas $(x_i, f_i)$, con $x_i > 0$, $(i = 1, \ldots, k)$, se define el *índice de desigualdad colectiva*, que designamos por $D$, mediante la siguiente expresión:

$$D = \sum_{i=1}^{k} \left(\frac{\bar{x}}{x_i} - 1\right) f_i \tag{4.3.2}$$

Este indicador, propuesto por López y Pérez (1991), puede ser obtenido como síntesis de desigualdades individuales.

## 4.4. La pobreza y su medición

La pobreza es un fenómeno complejo en el que influyen multitud de factores y cuyo análisis puede ser planteado desde diferentes enfoques. En particular, aquí nos referiremos al enfoque objetivo, dentro del cual se enmarcan aquellos estudios de pobreza que utilizan información recogida mediante variables que son directamente observables por el investigador (principalmente se trata de variables monetarias tales como ingresos, gastos, ...).

En los estudios sobre pobreza un problema básico es identificar el colectivo de pobres, para lo cual se introduce el concepto de línea o umbral de pobreza.

**Definición 4.6.** En términos generales, se entiende por *línea de pobreza* el nivel de renta requerido para cubrir las necesidades consideradas básicas. En consecuencia, un individuo es pobre si su renta es inferior a ese umbral.

Para establecer el umbral es necesario especificar el concepto de pobreza, para lo cual se plantean básicamente dos alternativas: pobreza absoluta y pobreza relativa.

- Se entiende por *pobreza absoluta* una situación en la que no están cubiertas las necesidades básicas del individuo, es decir, existe carencia de bienes y servicios básicos (normalmente relacionados con la alimentación, la vivienda y el vestido). En otros términos, se trata de una situación en que la persona no dispone de recursos indispensables para la subsistencia.

  Una *línea de pobreza absoluta* es aquélla que cuantifica en términos monetarios la cantidad mínima de subsistencia, es decir, se trata de un umbral basado en la satisfacción de las necesidades básicas exclusivamente. En el ámbito internacional suelen considerarse actualmente umbrales de 1$, 1,25$ ... 2$ diarios.

- A su vez, se entiende por *pobreza relativa* una situación en la que el individuo no tiene lo suficiente para vivir una vida que es considerada normal en la sociedad. En otros términos, desde esta perspectiva, se considera que una persona es pobre cuando se encuentra en una situación de clara desventaja respecto al resto de personas de su entorno, esto es, no dispone de recursos suficientes para alcanzar un nivel de vida mínimamente adecuado. Esta concepción de la pobreza está muy ligada a la noción de desigualdad.

  Para establecer una *línea de pobreza relativa* se considera una variable monetaria (ingresos, gastos, ...) y se fija el umbral como un porcentaje de un promedio de la distribución (generalmente la media o la mediana). Actualmente se utiliza la mediana ya que así se evita que los resultados se vean muy afectados por la presencia de valores extremos que no reflejan la realidad de la mayoría de la población. El umbral se fija en el 40, 50, 60 % ... de la mediana según el grado de severidad que se quiera contemplar; algunos organismos como EUROSTAT, OCDE, INE consideran el 60 % de la mediana.

**Definición 4.7.** Llamamos *tasa de pobreza*, que designamos por $H$, a la proporción de personas pobres, esto es, que se encuentran por debajo del umbral de pobreza en la población total considerada:

$$H = \frac{q}{N} \tag{4.4.1}$$

donde $q$ es el número de pobres y $N$ el tamaño poblacional.

Si se considera una línea de pobreza relativa el número de pobres depende de la posición relativa de cada individuo en la sociedad. Así pues, si se produce una variación proporcional de los ingresos de todos los individuos, la mediana, y por tanto también la línea de pobreza variarán en la misma proporción pero la tasa de pobreza no se modificará, ya que las frecuencias permanecen invariantes.

Se pueden calcular tasas de pobreza para diferentes grupos de población, según variables demográficas o socioeconómicas, tales como sexo, edad, nivel de estudios, etc.

> En la Encuesta de Condiciones de Vida elaborada por el INE se calculan diversas tasas de pobreza (por sexo, edad, etc.). En la figura 4.5 se recogen las tasas de pobreza por Comunidades Autónomas en 2009.

La tasa de pobreza es una medida muy sencilla; ahora bien refleja lo que se conoce como incidencia de la pobreza pero no su intensidad, es decir, proporciona información sobre la cantidad de personas que padecen pobreza pero no sobre el grado de pobreza que sufren los afectados. Para estudiar este aspecto es necesario definir otros indicadores.

**Definición 4.8.** Dado un umbral de pobreza $z$ y conocida la renta media del colectivo de pobres ($\bar{x}_q$), se define la *brecha de renta*, que designamos por $I$, como:

$$I = \frac{z - \bar{x}_q}{z} \tag{4.4.2}$$

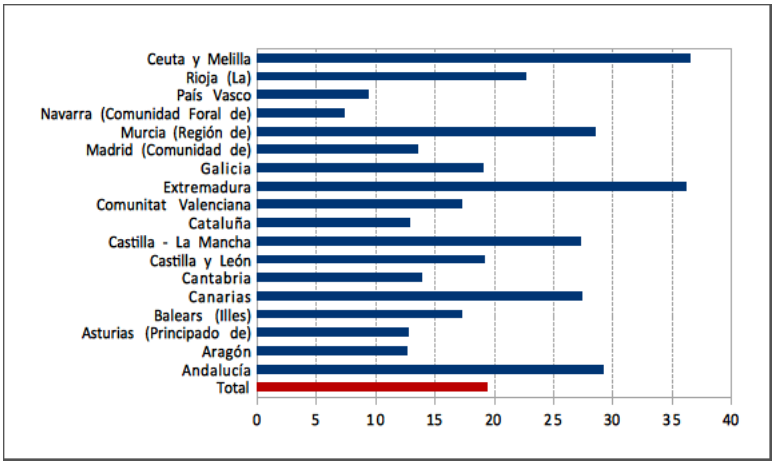



Figura 4.5: Tasas de pobreza, 2009

Dado que el numerador es la diferencia entre el umbral de pobreza y la renta media que perciben los pobres, la expresión anterior se puede interpretar como la proporción de la cantidad $z$ necesaria por término medio para que un individuo pobre deje de serlo, esto es, abandone la situación de pobreza. Así pues el cociente anterior es un indicador de la intensidad de la pobreza. Por ejemplo, un resultado $I = 0,25$ refleja que sería necesario aumentar por término medio la renta de una persona en una cantidad igual a la cuarta parte del umbral $z$ para que ésta abandonase la situación de pobreza.

A partir de la tasa de pobreza y de la brecha de renta se puede establecer un nuevo indicador que tiene en cuenta ambos aspectos, incidencia e intensidad de la pobreza.

**Definición 4.9.** Se define la *brecha de pobreza*, que designamos por $HI$, como el producto de la tasa de pobreza ($H$) y la brecha de renta ($I$), esto es:

$$HI = \frac{q}{N}\frac{z - \bar{x}_q}{z} \tag{4.4.3}$$

# 5 Análisis conjunto. Asociación y correlación

Al analizar la realidad socioeconómica encontramos información sobre características que generalmente están relacionadas. En algunos casos esta relación será directa, como sucede si por ejemplo estudiamos conjuntamente el Producto Interior Bruto (PIB) y el empleo, variables que habitualmente varían en el mismo sentido (crecen o decrecen a la vez). En cambio, en otras ocasiones podría existir una relación inversa, cuando las variables varían en sentido contrario (esto sucedería si por ejemplo sobre un conjunto de productos estudiamos conjuntamente sus precios y sus cantidades vendidas).

También podrían presentarse situaciones que no se corresponden con ninguno de los casos anteriores, por no existir relación entre los caracteres analizados, que podrían ser independientes. De hecho la idea de independencia tiene gran interés en el análisis estadístico.

En este tema vamos a examinar las relaciones existentes entre varios caracteres, para lo cual comenzaremos planteando un análisis conjunto de dos variables y la correspondiente notación.

## 5.1. Distribuciones bidimensionales

Consideremos la distribución conjunta $(X, Y)$ de dos variables o atributos $X$ e $Y$, que toma los valores o modalidades $(x_i, y_j)$, $(i = 1, ..., k;\ j = 1, ..., m)$. Denominaremos frecuencia absoluta conjunta $n_{ij}$ al número de repeticiones del par $(x_i, y_j)$. Si $N$ es el número total de observaciones, se verifica:

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

La frecuencia relativa conjunta $f_{ij}$, se obtiene como: $f_{ij} = \frac{n_{ij}}{N}$, $(i = 1, \ldots, k;\ j = 1, \ldots, m)$, y se interpreta como la proporción del total de las observaciones en las que se observan conjuntamente los valores $(x_i, y_j)$. Como consecuencia de esta definición se verifica: $\sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij} = 1$

**Definición 5.1.** Una *distribución bidimensional* viene dada por las observaciones conjuntas de dos caracteres con sus frecuencias correspondientes, que representaremos genéricamente por $(x_i, y_j, n_{ij})$ *o* $(x_i, y_j, f_{ij})$, $(i = 1, ..., k;\ j = 1, ..., m)$.

Las distribuciones bidimensionales suelen representarse mediante tablas del tipo:

| $X\backslash Y$ | $y_1$ | $y_2$ | $\cdots$ | $y_m$ |
|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1m}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2m}$ |
| $\vdots$ | $\ddots$ | $\ddots$ | $\cdots$ | $\vdots$ |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{km}$ |

o bien, como caso particular cuando las frecuencias son unitarias, por tablas de dos columnas:

| $X$ | Y |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $\vdots$ | $\vdots$ |
| $x_N$ | $y_N$ |

Si se trata de caracteres cuantitativos o variables las tablas anteriores se denominan *tablas de correlación*. La gráfica más habitual en este caso es la *nube de puntos*, que consiste en representar en un sistema de coordenadas cartesianas los pares de valores $(x_i, y_j)$.

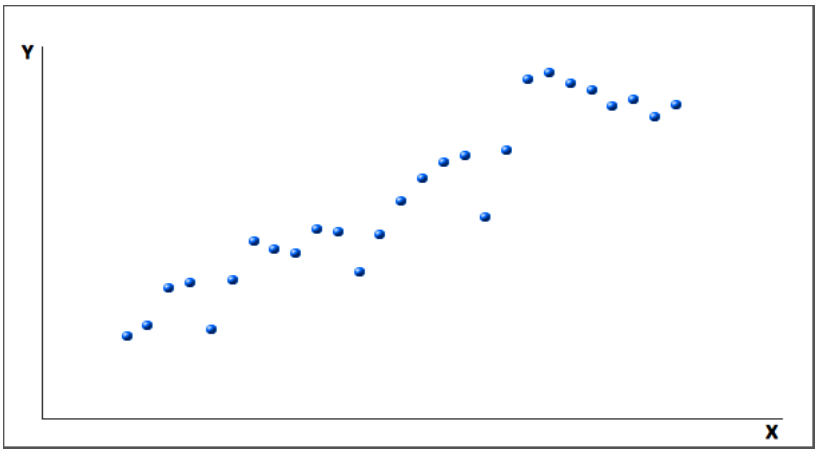



Figura 5.1: Nube de puntos X-Y

**Ejemplo 5.1.** Las variables renta familiar $(X)$ y gasto en viajes $(Y)$, ambas expresadas en miles de euros, pueden ser representadas mediante una tabla del siguiente tipo:

| $X\backslash Y$ | 2 | 4 | 10 |
|---|---|---|---|
| 24 | 4 | 1 | |
| 30 | 2 | 5 | 1 |
| 50 | | 1 | 6 |

En el caso del estudio conjunto de dos caracteres cualitativos o atributos, las tablas resultantes reciben el nombre de *tablas de contingencia*.

**Ejemplo 5.2.** La tabla siguiente representa la información relativa al sexo ($X$) y la situación de actividad ($Y$) de 30 personas:

| $X\backslash Y$ | Ocupados | Parados | Inactivos |
|---|---|---|---|
| Hombres | 12 | 4 | 2 |
| Mujeres | 2 | 4 | 6 |

También existen distribuciones mixtas, en las cuales una característica es cuantitativa y otra cualitativa, como la clasificación de los habitantes de un país según la edad y el sexo. Este tipo de representación resulta muy habitual por aparecer estrechamente ligado a las pirámides de población. La página web del Instituto Nacional de Estadística (www.ine.es) en el apartado dedicado a Censos de Población permite construir tablas como la siguiente:

| Edad en grandes grupos | TOTAL | Menos de 16 | 16-64 | 65 ó más |
|---|---|---|---|---|
| Sexo | | | | |
| TOTAL | 40.847.371 | 6.379.748 | 27.509.107 | 6.958.516 |
| Varón | 20.012.882 | 3.275.785 | 13.806.534 | 2.930.563 |
| Mujer | 20.834.489 | 3.103.963 | 13.702.573 | 4.027.953 |

En este tema nos centraremos principalmente en el estudio conjunto de caracteres cuantitativos o variables, que es el caso más habitual. Una visión intuitiva del comportamiento conjunto de dos variables puede obtenerse a partir de la correspondiente nube de puntos que, como hemos visto anteriormente, es una representación gráfica sencilla consistente en representar las variables en un sistema de ejes cartesianos. Tal y como puede verse en la figura 5.1, los pares de valores dan lugar a una serie de puntos sobre el plano que se corresponden con las observaciones, y la forma de la nube de puntos puede resultar de gran utilidad para analizar la forma en que se relacionan las variables estudiadas y la intensidad de la relación existente entre ellas.

## 5.2. Distribuciones marginales y condicionadas

Supongamos que a partir de la distribución conjunta sobre renta y gasto en viajes nos interesa estudiar aisladamente la renta o el gasto en viajes. En tal caso tendremos que considerar distribuciones unidimensionales que serán las distribuciones marginales de las variables $X$ e $Y$ respectivamente. Así, la distribución marginal de la renta ($X$) se obtiene considerando los valores que toma esa variable, así como sus respectivas frecuencias independientemente de los valores del gasto en viajes ($Y$) con los que aparece. Más concretamente, sobre el ejemplo anterior se tendría la siguiente distribución:

| Valores $X$ | Frecuencias |
|---|---|
| 24 | 5 |
| 30 | 8 |
| 50 | 7 |

De manera análoga se obtiene la distribución marginal del gasto en viajes ($Y$)

| Valores $Y$ | Frecuencias |
|---|---|
| 2 | 6 |
| 4 | 7 |
| 10 | 7 |

Como se puede observar, en ambos casos analizamos el comportamiento de una de las variables ignorando, en cierto modo, el de la otra.

**Definición 5.2.** Dada una distribución bidimensional $(x_i, y_j, n_{ij});\ i = 1, \ldots, k,\ j = 1, \ldots, m$, la *distribución marginal* de $X$ es una distribución unidimensional que vendrá dada por los pares $(x_i, n_{i.}), i = 1, \cdots, k$ donde $n_{i.}$ es la *frecuencia marginal* del valor $x_i$, obtenida como suma de la i-ésima fila de frecuencias absolutas conjuntas, es decir $n_{i.} = \sum_{j=1}^{m} n_{ij}$.

Siguiendo esta definición, en la tabla de correlación aparecen las frecuencias marginales $n_{1.}, n_{2.}, \cdots, n_{k.}$

| $X \backslash Y$ | $y_1$ | $y_2$ | $\cdots$ | $y_m$ | $n_{i.}$ |
|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1m}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2m}$ | $n_{2.}$ |
| $\vdots$ | $\ddots$ | $\ddots$ | $\ldots$ | $\ddots$ | $\vdots$ |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{km}$ | $n_{k.}$ |
| $n_{.j}$ | $n_{.1}$ | $n_{.2}$ | | $n_{.m}$ | N |

La distribución marginal de $X$ puede venir también dada por $(x_i, f_{i.})$, donde la frecuencia relativa $f_{i.}$ es el cociente entre el número de veces que aparece el valor $x_i$ y el número total de observaciones ($N$), es decir: $f_{i.} = \dfrac{n_{i.}}{N}$.

Dado que las distribuciones marginales son distribuciones unidimensionales, es posible calcular sobre ellas todas las medidas estudiadas en temas anteriores, tales como media, mediana, moda, varianza, medidas de desigualdad, etc. La *media* y la *varianza marginal* de la variable $X$ vienen dadas por las expresiones:

$$\bar{x} = \sum_{i=1}^{k} x_i f_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{m} x_i f_{ij}$$

$$S_X^2 = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{m} (x_i - \bar{x})^2 f_{ij}$$

De modo análogo se podría definir la distribución marginal de $Y$ como $(y_j, n_{.j})$ o $(y_j, f_{.j})$ con $n_{.j} = \sum_{i}^{k} n_{ij}$, calculando sobre la misma la media marginal u otros promedios y la varianza marginal.

**Ejemplo 5.3.** A modo de ilustración, partiendo de la tabla de correlación entre renta y gasto en viajes (Ejemplo 5.1) se obtienen las siguientes medidas marginales:

$$\bar{x} = 35,5 \quad ; \qquad S_X^2 = 118,75$$

$$\bar{y} = 5,5 \quad ; \qquad S_Y^2 = 11,55$$

Además del análisis marginal anteriormente descrito, a partir de una distribución bidimensional es posible llevar a cabo un análisis condicionado, estudiando de qué modo una de las variables estudiadas influye o condiciona a la otra.

**Ejemplo 5.4.** Supongamos que a partir de los datos anteriores sobre renta y gasto en viajes queremos conocer el comportamiento del gasto en viajes para las familias con determinado nivel de renta. Este planteamiento nos conduce a la distribución de $Y$ condicionada a un valor de $X$. Así por ejemplo, la tabla nos muestra los valores y frecuencias del gasto en viajes condicionado a una renta de 24.000 euros (es decir, la distribución de $Y$ condicionada a $X = 24$, en el Ejemplo 5.1):

| Valores Y/X=24 | Frecuencias |
|---|---|
| 2 | 4 |
| 4 | 1 |
| 10 | 0 |

De esta manera tenemos la distribución del gasto en viajes para una subpoblación de la población global: el colectivo de familias que perciben una renta de 24.000 euros. Análogamente, se podría obtener la distribución de la renta condicionada a algún valor del gasto en viajes.

**Definición 5.3.** Dada una distribución bidimensional $(x_i, y_j, n_{ij})$, $i = 1, \ldots, k$ , $j = 1, \ldots, m$, la *distribución de $X$ condicionada a $y_j$* es una distribución unidimensional que se denota por $X/y_j$, $\forall j = 1, \cdots, m$ y viene dada por:

$$(x_i, n_{ij}),\ \forall i = 1, \ldots, k$$

o bien en términos de frecuencias relativas por: $(x_i, f_{i/j})$, $\forall i = 1, \ldots, k$, donde la frecuencia relativa $f_{i/j}$ que también se denota por $f(x_i/y_j)$ se obtiene como cociente entre la frecuencia conjunta del par $(x_i, y_j)$, y la frecuencia marginal del valor $y_j$, es decir:

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

**Definición 5.4.** Dada una distribución bidimensional $(x_i, y_j, n_{ij})$, $i = 1, \ldots, k$, $j = 1, \ldots, m$ la distribución de $Y$ condicionada a $x_i, \forall i = 1, \ldots, k$, viene dada por:

$$(y_j, n_{ij}),\ \forall j = 1, \ldots, m$$

o bien, en términos de frecuencias relativas, por: $(y_j, f_{j/i})$, $\forall j = 1, \ldots, m$ donde $f_{j/i} = \dfrac{n_{ij}}{n_{i.}} = \dfrac{f_{ij}}{f_{i.}}$, expresión que puede también denotarse como $f(y_j/x_i)$.

Las distribuciones condicionadas, al igual que las marginales, son distribuciones unidimensionales para las cuales se pueden calcular todas las medidas definidas en temas previos.

Las medias y las varianzas de las distribuciones condicionadas anteriores vienen dadas por las expresiones siguientes:

| Media condicionada | Varianza condicionada |
|---|---|
| $\bar{x}/y_j = \sum_{i=1}^{k} x_i f_{i/j}$ | $S^2_{X/y_j} = \sum_{i=1}^{k} (x_i - \bar{x}/y_j)^2 f_{i/j}$ |
| $\bar{y}/x_i = \sum_{j=1}^{m} y_j f_{j/i}$ | $S^2_{Y/x_i} = \sum_{j=1}^{m} (y_j - \bar{y}/x_i)^2 f_{j/i}$ |

**Propiedad 5.1.** *La media marginal de $X$ puede ser obtenida como media de las medias de $X$ condicionadas a los distintos valores de $Y$.*

*Demostración.*

$$\sum_{j=1}^{m} (\bar{x}/y_j)\, f_{.j} = \sum_{i=1}^{k}\sum_{j=1}^{m} x_i f_{i/j} f_{.j} = \sum_{i=1}^{k}\sum_{j=1}^{m} x_i \frac{n_{ij}}{n_{.j}} \frac{n_{.j}}{N} = \sum_{i=1}^{k} x_i \sum_{j=1}^{m} \frac{n_{ij}}{N} = \sum_{i=1}^{k} x_i f_{i.} = \bar{x}$$

□

En otros términos, si consideramos las medias de $X$ condicionadas a todos los posibles valores de $Y$, $\bar{x}/y_1, \bar{x}/y_2, \ldots, \bar{x}/y_m$, la situación sería equivalente a considerar distintas subpoblaciones de una población total. Como consecuencia podríamos aplicar la propiedad ya estudiada que permite obtener la media global a partir de las medias de subpoblaciones.

Análogamente se puede demostrar que la media marginal de $Y$ puede obtenerse a partir de las medias de $Y$ condicionadas a distintos valores de $X$:

$$\sum_{i=1}^{k} (\bar{y}/x_i)\, f_{i.} = \bar{y}$$

**Ejemplo 5.5.** A modo de ilustración, podemos calcular las medias de gasto en viajes condicionadas a los distintos valores de la renta:

$$(\bar{y}/X = 24) = 2,40$$

$$(\bar{y}/X = 30) = 4,25$$

$$(\bar{y}/X = 50) = 9,14$$

Y a partir de todas ellas es posible obtener la media marginal de $Y$:

$$\bar{y} = \sum_{i=1}^{k} (\bar{y}/x_i)\, f_{i.} = 2,4 \times 0,25 + 4,25 \times 0,4 + 9,14 \times 0,35 = 5,5$$

## 5.3. Dependencia e independencia estadística

Al llevar a cabo un análisis conjunto de variables resulta interesante examinar las posibles relaciones de dependencia existentes entre ellas, que pueden ser de muy diverso tipo, tal y como ilustran los siguientes ejemplos:

**Ejemplo 5.6.** En un surtidor de gasolina se observa determinado día el número de litros de cierto tipo de combustible ($X$) y el importe pagado por los clientes ($Y$), cuya distribución conjunta se representa en una nube de puntos. Como se puede apreciar en la figura 5.2 es posible trazar una línea recta que pasa por todos esos puntos, y que explica el importe pagado en función de la cantidad de combustible.

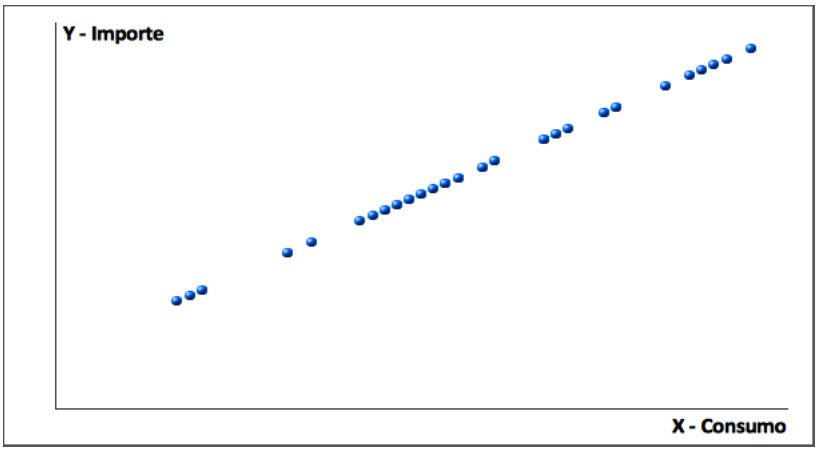



Figura 5.2: Nube de puntos consumo de combustible-importe pagado

Así pues, éste es un caso de *dependencia funcional*: la variable $Y$ depende funcionalmente de $X$ puesto que existe una aplicación unívoca que expresa la relación entre ambas variables y permite obtener los importes pagados a partir de la cantidad de combustible ($Y = kX$, siendo $k$ el precio del litro de combustible el día en que realizamos las observaciones).

Se observa además que en este ejemplo también $X$ depende funcionalmente de $Y$, puesto que existe una función que permite obtener la cantidad de combustible a partir del importe abonado ($X = \frac{Y}{k}$).

**Ejemplo 5.7.** La información relativa a renta familiar y gasto en viajes puede ser representada mediante una nube de puntos como en la figura 5.3, donde se aprecia que si bien no existe una función que explique de forma exacta el gasto en viajes a partir

de la renta, las observaciones tienden a agruparse alrededor de una línea y podemos aproximarlas mediante una recta, pudiendo resultar de interés tratar de obtener aquella recta que mejor explique, basándose en estos datos, el gasto en viajes a partir de la renta. Es necesario notar que conceptualmente no tendría sentido considerar en este caso una recta para explicar la renta a partir del gasto en viajes.

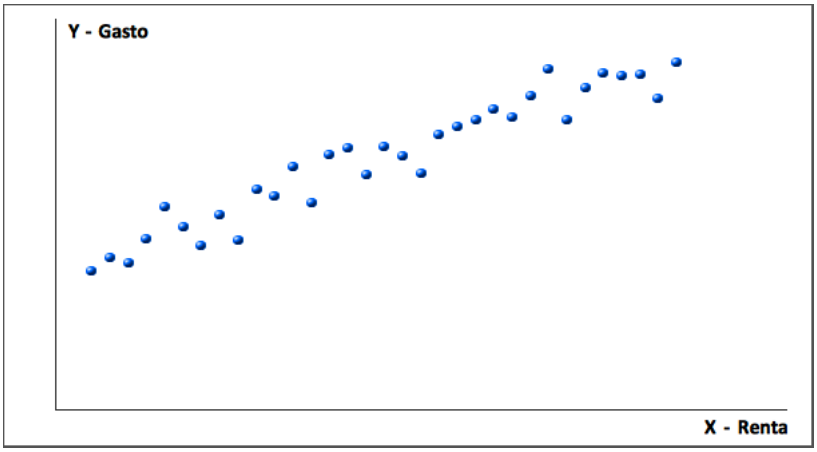



Figura 5.3: Nube de puntos renta familiar-gasto en viajes

**Ejemplo 5.8.** De modo similar, si consideramos información para los técnicos de una empresa sobre su experiencia ($X$, en años) y el tiempo medio tardado en realizar una tarea ($Y$, en minutos), la nube de puntos de la figura 5.4 muestra que tampoco en este caso existe una función que exprese de manera exacta el tiempo medio empleado a partir de la experiencia, pero estos datos pueden ser aproximados por una hipérbola, resultando de interés tratar de obtener, basándonos en las observaciones, aquélla que mejor explique el tiempo medio tardado a partir de la experiencia.

**Ejemplo 5.9.** Por último, si representamos la altura ($X$, en m) y la renta disponible anual ($Y$, en miles de euros), la distribución conjunta da lugar a la nube de puntos de la figura 5.5, donde las observaciones no presentan relación alguna y no es posible encontrar una función que aproxime los datos y nos permita obtener la renta disponible a partir de la altura.

En síntesis, los ejemplos anteriores recogen los distintos tipos de relación existente entre variables, desde la dependencia funcional (cuando existe una función que expresa de manera exacta la relación entre las variables, como en el caso del consumo de combustible y su importe), hasta la independencia (como el caso de altura y renta), pasando por las situaciones de dependencia estadística. La independencia estadística se presentará cuando entre las variables consideradas no exista ningún tipo de relación. Intuitivamente diremos que dos variables son independientes si los valores que adopta cada una de ellas no están influenciados por los valores que toma la otra.

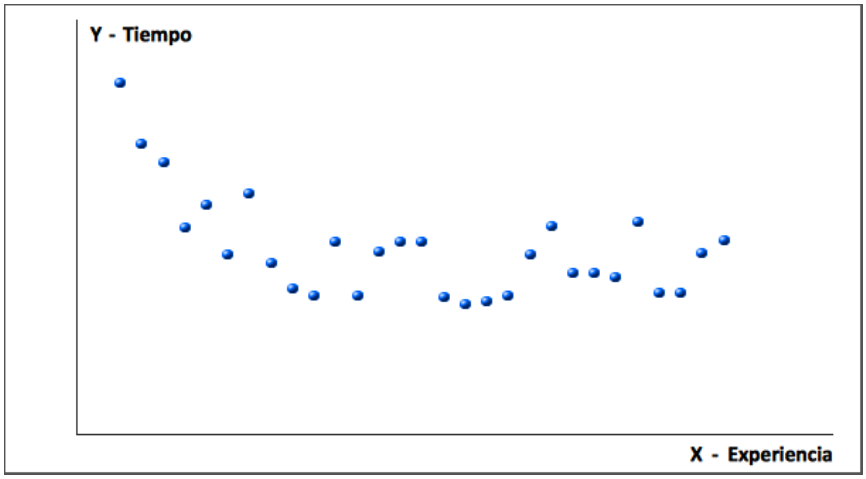



Figura 5.4: Nube de puntos experiencia-tiempo para realizar una tarea

**Definición 5.5.** Dada una variable bidimensional $(X, Y)$ se dice que $X$ es *independiente* de $Y$ cuando las distribuciones de $X$ condicionadas a cualquier valor de $Y$ coinciden, siendo además iguales a la distribución marginal de $X$, esto es, cuando:

$$f(x_i/y_j) = f(x_i/y_h) = f_{i.},\ \forall i = 1, \ldots, k;\ \forall j, h = 1, \ldots, m$$

La definición anterior implica que las frecuencias no se ven afectadas por los valores a los que se condiciona la variable $X$.

De manera análoga, la variable $Y$ es independiente de $X$ cuando las distribuciones de $Y$ condicionadas a cualquier valor de $X$ coinciden con la distribución marginal de $Y$, es decir, cuando:

$$f(y_j/x_i) = f(y_j/x_r) = f_{.j},\ \forall j = 1, \ldots, m;\ \forall i, r = 1, \ldots, k$$

La independencia es un concepto relativo, por lo que si se habla de variables independientes conviene preguntarse respecto a qué. El concepto que acabamos de introducir se refiere a la independencia en términos de frecuencias, también conocido como *independencia estadística* pero pueden plantearse otras definiciones: independencia en probabilidad, independencia en información, etc.

**Propiedad 5.2.** *(*Condición de independencia*) Si una variable $X$ es independiente de otra variable $Y$, la frecuencia relativa conjunta de todos los pares de valores será igual al producto de las frecuencias relativas marginales de los valores correspondientes, es decir, $X$ es independiente de $Y$, si y solo si, se cumple:*

$$f_{ij} = f_{i.}f_{.j},\ \forall i = 1, \ldots, k;\ \forall j = 1, \ldots, m$$

*Demostración.* Dado que $X$ es independiente de $Y$, se tiene: $f_{i/j} = f_{i.}$ y, por definición, $f_{i/j} = \frac{f_{ij}}{f_{.j}}$, por tanto, $f_{ij} = f_{i.}f_{.j}$, $\forall i = 1, \ldots, k$; $\forall j = 1, \ldots, m$.

Procediendo de forma análoga en sentido contrario deducimos el recíproco. □

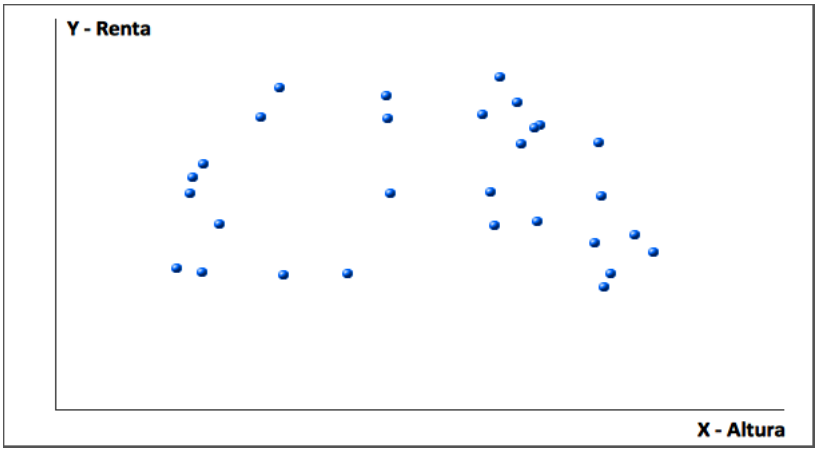



Figura 5.5: Nube de puntos altura-renta disponible

La *condición de independencia* resulta de gran utilidad en la práctica ya que, si se quiere comprobar a partir de una tabla de correlación si dos variables son independientes, en general resultará más cómodo realizar la comprobación de dicha igualdad que comprobar si todas las distribuciones condicionadas de una variable coinciden con la marginal. La condición de independencia es una relación simétrica por lo que si $X$ es independiente de $Y$, se verifica la condición permutando las variables, es decir, se cumple también que $Y$ es independiente de $X$. La independencia entre variables es, pues, un concepto recíproco sin implicaciones en el orden de independencia.

En las tablas de correlación aparecen a menudo frecuencias absolutas, resultando conveniente expresar la condición de independencia en términos de éstas, es decir:

$$f_{ij} = f_{i.}f_{.j} \;\Leftrightarrow\; n_{ij} = \frac{n_{i.}n_{.j}}{N};\; \forall i = 1,\ldots,k;\; \forall j = 1,\ldots,m$$

Entre los dos casos extremos de dependencia funcional e independencia estadística pueden existir muchas situaciones intermedias, caracterizadas porque existe una cierta relación entre las variables que no puede ser expresada mediante una función; son casos de *dependencia estadística*. En general, los fenómenos económicos no suelen ser deterministas, no vienen especificados por leyes rígidas que conducen a resultados exactos; por el contrario, más bien podemos afirmar que existe una importante conexión entre diferentes fenómenos económicos, pero que no admite una expresión exacta o funcional; por tanto, la forma de representar esa conexión será mediante una dependencia estadística.

## 5.4. Medidas de asociación

Cuando disponemos de información referida a dos caracteres cualitativos recogida en una tabla de contingencia podemos determinar si estos dos atributos son o no

independientes mediante la condición de independencia, puesto que en dicha condición no intervienen los valores, sino únicamente las frecuencias con las que éstos aparecen.

Una vez comprobado que dos atributos son dependientes podemos cuantificar el grado de asociación o dependencia entre ellos a través de diversas medidas de asociación entre las que destacan los coeficientes de Pearson y Kendall.

El coeficiente chi-cuadrado de Pearson compara la distribución observada con la que se habría obtenido en el supuesto de que los atributos fuesen estadísticamente independientes.

Como ya hemos visto, en condiciones de independencia las frecuencias conjuntas $n_{ij}$ se obtendrían a partir de las marginales, según la condición de independencia, como: $\frac{n_{i.}n_{.j}}{N}$.

**Definición 5.6.** El *coeficiente chi-cuadrado de Pearson* es una medida de la distancia entre dos distribuciones, que viene dada por la expresión siguiente:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left[n_{ij} - \frac{n_{i.}n_{.j}}{N}\right]^2}{\frac{n_{i.}n_{.j}}{N}} \tag{5.4.1}$$

Este coeficiente mide, en términos relativos, cuánto dista la distribución conjunta de los atributos de la situación de independencia; por tanto, un valor más elevado de $\chi^2$ indica un mayor grado de asociación entre los atributos mientras que en caso de independencia este coeficiente será nulo.

**Propiedad 5.3.** *$\chi^2 = 0$ si y sólo si $X$ e $Y$ son independientes*

*Demostración.* Dado que todos los sumandos que intervienen en el coeficiente chi-cuadrado son no negativos, si $\chi^2 = 0$ entonces todos y cada uno de dichos términos deben ser nulos y, en consecuencia, los respectivos numeradores también, con lo cual se cumplirá la condición de independencia. Análogamente, en sentido contrario, partiendo de la condición de independencia se deducirá que en tal situación el coeficiente chi-cuadrado toma valor nulo. □

Esta medida no presenta una cota absoluta sino que su valor máximo depende del número de filas y columnas de la distribución sobre la que se calcule. No obstante, su gran ventaja es que para tamaños grandes desde un enfoque probabilístico se conoce su distribución (denominada chi-cuadrado de Pearson), que nos permite asignar una fiabilidad en términos de probabilidad a nuestras conclusiones sobre la dependencia de las variables o atributos.

**Definición 5.7.** Llamamos *coeficiente de contingencia de Pearson*, $C$, al valor positivo de la siguiente expresión:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{5.4.2}$$

El resultado de $C$ está acotado entre 0 y 1, correspondiendo el valor nulo al caso de independencia entre los atributos, y aumentando el valor de $C$ con la intensidad de la asociación entre atributos. Puede comprobarse que el valor 1 no se alcanza nunca y para una tabla cuadrada de dimensión $m \times m$ la cota superior que se puede alcanzar es $\sqrt{\frac{m-1}{m}}$.

En el caso de que sea posible establecer una ordenación natural en las modalidades de los atributos, para medir el grado de asociación entre los mismos se pueden emplear medidas de correlación por rangos entre las que destaca el coeficiente de Kendall.

En general, si se parte de una distribución conjunta como la representada en la tabla:

| X | Y |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $\vdots$ | $\vdots$ |
| $x_N$ | $y_N$ |

la medida de Kendall se construye analizando los pares de observaciones $(x_i, y_i)$ $(x_j, y_j)$ que se clasifican en concordantes y discordantes. Se dice que dos pares de observaciones $(x_i, y_i)$ $(x_j, y_j)$ son concordantes si los rangos de sus dos elementos coinciden, es decir si se cumple $x_i > x_j$, $y_i > y_j$ o bien $x_i < x_j$, $y_i < y_j$. En cambio, se dice que los pares son discordantes si se cumple $x_i > x_j$, $y_i < y_j$ o bien $x_i < x_j$, $y_i > y_j$. Por último, en caso de igualdad $x_i = x_j$ o $y_i = y_j$ los pares no son ni concordantes ni discordantes (se dice que hay un empate).

**Definición 5.8.** El *coeficiente $\tau$ de Kendall* se define mediante la expresión:

$$\tau = \frac{\text{(número de pares concordantes)-(número de pares discordantes)}}{\frac{N(N-1)}{2}} \tag{5.4.3}$$

El coeficiente $\tau$ de Kendall está acotado entre -1 y 1, tomando el valor 0 en caso de independencia, el valor 1 cuando la asociación es máxima (es decir, coinciden las dos ordenaciones) y -1 cuando las dos ordenaciones son inversas.

## 5.5. La correlación y su medida

La dependencia estadística, que es el tipo de relación más habitual entre variables económicas, admite diferentes grados, ya que puede presentarse una asociación más o menos intensa entre variables. De ahí el interés de definir medidas que cuantifican el grado de dependencia entre dos variables.

Si consideramos la nube de puntos asociada a la distribución de renta y gasto en viajes (figura 5.4), se puede apreciar que las variables $X$ e $Y$ están relacionadas positivamente, pues a valores altos de renta corresponden valores altos de gasto en viajes.

Las comparaciones sobre las variables no suelen plantearse en términos absolutos, pues si se produce un cambio de dimensionalidad en las unidades de las variables, ésta puede distorsionar el verdadero grado de asociación existente entre ellas.

Como primera etapa en la construcción de una medida, tomaremos los valores normalizados respecto a sus medias; esto es, en vez de considerar los valores originales de las variables los consideraremos centrados respecto a su media. Si consideramos ahora el producto de estas desviaciones: $(x_i - \bar{x})(y_j - \bar{y})$, un signo positivo indica que los valores de ambas variables se sitúan por debajo o por encima de la media, mientras un signo negativo señala un cambio de sentido de modo que mientras una de las variables se sitúa por encima de la media, la otra está por debajo.

Este producto de desviaciones nos permite hacer comparaciones para cada par de valores de la distribución. Sin embargo, lo que pretendemos es construir una medida que facilite un valor único para el conjunto de todos los datos y con este objetivo podemos construir un promedio del producto de desviaciones teniendo en cuenta sus correspondientes frecuencias.

**Definición 5.9.** Dada una variable bidimensional $(X, Y)$, se denomina *covarianza*, que se denota por $S_{XY}$, al valor de la expresión:

$$S_{XY} = \sum_{i=1}^{k} \sum_{j=1}^{m} (x_i - \bar{x})\,(y_j - \bar{y})\,f_{ij} \tag{5.5.1}$$

La covarianza es una medida de variación conjunta de dos variables que indica únicamente el signo de su relación lineal, ya que como observamos en su expresión relaciona las desviaciones de orden 1 en $X$ con las del mismo orden en $Y$.

Si consideramos la covarianza de una variable consigo misma, obtenemos la varianza de esa variable, que adopta siempre un valor no negativo (puesto que al comparar una variable consigo misma siempre varían en la misma dirección). Sin embargo la covarianza entre dos variables distintas puede tener signo positivo o negativo indicando dicho signo la dirección de la relación.

Más concretamente, teniendo en cuenta la definición anterior, un signo positivo de la covarianza significa que al promediar pesan más los sumandos positivos que los negativos, de modo que cuando una variable aumenta por encima de su media, la otra lo hace también, situación representada en la nube de puntos de la figura 5.6.

Teniendo en cuenta los cuadrantes definidos por las medias marginales $\bar{x}$ e $\bar{y}$, en la figura 5.6, se observa que un punto $(x_i, y_j)$ situado en el cuadrante superior derecho tiene desviaciones positivas respecto a las medias $\bar{x}$ e $\bar{y}$, por lo que el producto de ambas desviaciones será también positivo. De modo análogo, para los pares de valores situados en el cuadrante inferior izquierdo las desviaciones respecto de las medias marginales serán ambas negativas, y el producto de ambas adoptará nuevamente signo positivo. Como consecuencia, si -como sucede en la figura 5.6- la mayoría de los puntos de la nube están en los cuadrantes superior derecho e inferior izquierdo entonces la covarianza será positiva porque la mayoría de los sumandos tendrán ese signo.

**Ejemplo 5.10.** A modo de ejemplo, ésta sería la situación observada al analizar conjuntamente la renta y el gasto en viajes, ya que generalmente cuando la renta supera al valor medio también el gasto en viajes se encontrará por encima del gasto medio. Como consecuencia, la covarianza entre ambas variables presentará signo positivo. Más concretamente, partiendo de la tabla de correlación de la renta y el gasto en viajes del ejemplo 5.1 se llega al resultado $S_{XY} = 30,15$.

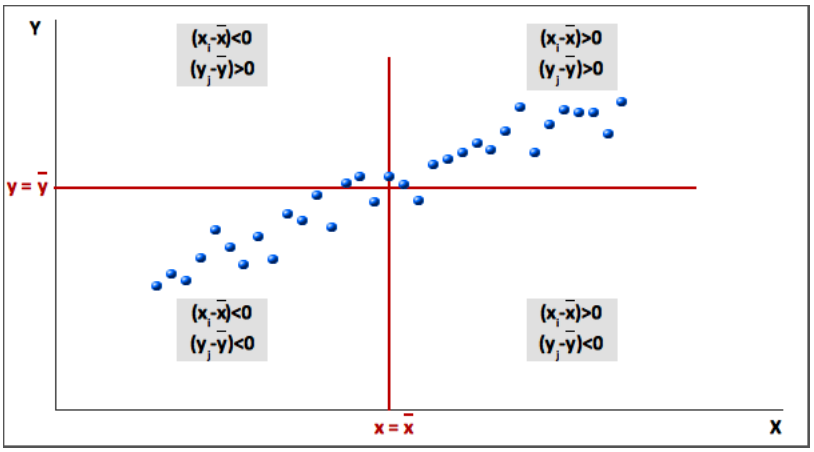



Figura 5.6: Análisis del signo de la covarianza

La situación opuesta se presenta cuando la mayor parte de las observaciones corresponden a los cuadrantes superior izquierdo e inferior derecho, en los que una de las variables adopta valores superiores a su media marginal mientras la otra se sitúa en valores inferiores a la media. En este caso, las desviaciones tienen signos opuestos y su producto dará un resultado negativo. Como consecuencia, la covarianza será negativa, indicando la existencia de una relación lineal inversa entre las variables. Este tipo de situaciones se presentará por ejemplo al analizar conjuntamente la experiencia y el tiempo empleado en realizar una tarea, como se puede observar en la nube de puntos representada en la figura 5.4, que presenta pendiente negativa. Por último, en el caso de que la nube de puntos observada se repartiese homogéneamente entre los cuatro cuadrantes, la covarianza adoptaría un valor aproximadamente nulo (ya que los productos de desviaciones de signo positivo se compensarían con los de signo negativo). Esta situación (similar a la representada en la nube de puntos de la figura 5.5) estaría indicando que no existe relación lineal entre las variables (como por ejemplo sucedería al analizar conjuntamente la renta y la altura).

**Propiedad 5.4.** *Si dos variables $X$ e $Y$ son independientes entonces la covarianza entre $X$ e $Y$ es nula*

*Demostración.* Para $X$ e $Y$ independientes se verifica $f_{ij} = f_{i.}f_{.j}$ y, en consecuencia:

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})f_{ij} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})f_{i.}f_{.j} \\ &= \underbrace{\sum_{i=1}^{k}(x_i-\bar{x})f_{i.}}_{=0}\underbrace{\sum_{j=1}^{m}(y_j-\bar{y})f_{.j}}_{=0} = 0 \end{aligned}$$

donde la última igualdad es consecuencia de la propiedad 2.1 de la media aritmética, aplicada a las distribuciones marginales de $X$ e $Y$. □

Conviene señalar que sin embargo, el recíproco de esta propiedad no es cierto, esto es, existen variables cuya covarianza es nula y que no son independientes. Este hecho se debe a que la covarianza es una medida de dependencia lineal, y por tanto cuando proporciona resultados nulos permite únicamente asegurar que las variables analizadas no presentan relación lineal entre sí, es decir, son *incorreladas*.

**Ejemplo 5.11.** Si el precio ($X$, en euros) y la demanda ($Y$, en miles de unidades) de cierto producto presentan la siguiente distribución conjunta:

| X | Y |
|---|---|
| 20 | 1 |
| 30 | 5 |
| 50 | 4 |
| 70 | 2 |

se puede comprobar fácilmente que $S_{XY} = 0$ y sin embargo no se cumple la condición de independencia ya que se tiene $n_{11}N = 4$ y en cambio $n_{1.}n_{.1} = 1$.

**Propiedad 5.5.** *La covarianza se puede obtener como diferencia entre la media del producto de las variables y el producto de las medias, esto es:*

$$S_{XY} = \sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j f_{ij} - \bar{x}\bar{y}$$

*Demostración.* Partiendo de la definición de la covarianza se obtiene:

$$
\begin{aligned}
S_{XY} &= \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})\,(y_j-\bar{y})\,f_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{m}x_iy_jf_{ij} - \bar{x}\sum_{i=1}^{k}\sum_{j=1}^{m}y_jf_{ij} - \bar{y}\sum_{i=1}^{k}\sum_{j=1}^{m}x_if_{ij} + \bar{x}\bar{y}\sum_{i=1}^{k}\sum_{j=1}^{m}f_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{m}x_iy_jf_{ij} - \bar{x}\sum_{j=1}^{m}y_jf_{.j} - \bar{y}\sum_{i=1}^{k}x_if_{i.} + \bar{x}\bar{y}\sum_{i=1}^{k}\sum_{j=1}^{m}f_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{m}x_iy_jf_{ij} - \bar{x}\bar{y}
\end{aligned}
$$

□

Esta expresión resulta de gran utilidad práctica, pues facilita los cálculos al permitir trabajar con las variables originales, sin necesidad de realizar una transformación en desviaciones.

**Propiedad 5.6.** *La covarianza es invariante ante cambios de origen, es decir, dadas* $X' = X + a$ *e* $Y' = Y + b$ *se cumple* $S_{X'Y'} = S_{XY}$

*Demostración.*

$$
\begin{aligned}
S_{X'Y'} &= S_{X+a,Y+b} = \sum_{i=1}^{k}\sum_{j=1}^{m}\left[x_i+a-(\bar{x}+a)\right]\left[y_j+b-(\bar{y}+b)\right]f_{ij} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})f_{ij} = S_{XY}
\end{aligned}
$$

□

Esta propiedad indica que un cambio de origen en una o ambas variables supone una traslación de los datos que no afecta a la relación lineal, en cambio no sucede lo mismo con los cambios de escala, ya que éstos afectan a las unidades de las variables y por tanto a su covarianza.

**Propiedad 5.7.** *La covarianza viene afectada por cambios de escala. Si se tiene* $X' = cX$ *e* $Y' = dY$ *la covarianza de las nuevas variables vendrá dada por la expresión:* $S_{X'Y'} = (cd)S_{XY}$

*Demostración.*

$$
\begin{aligned}
S_{X'Y'} &= S_{cX,dY} = \sum_{i=1}^{k}\sum_{j=1}^{m}(cx_i-c\bar{x})(dy_j-d\bar{y})f_{ij} \\
&= cd\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i-\bar{x})(y_j-\bar{y})f_{ij} = cdS_{XY}
\end{aligned}
$$

□

A pesar de sus muchas ventajas, la covarianza tiene también algunas limitaciones importantes. Por una parte, como ya hemos comentado esta medida recoge únicamente la dependencia que afecta a la componente lineal de la relación entre variables. Por tanto, es posible obtener distribuciones de variables que guardan relación exacta y presentan covarianza nula (éste sería el caso por ejemplo, en situaciones del tipo $y = x^2$, donde $x$ adopta valores positivos y negativos simétricos).

Además, la covarianza presenta el inconveniente de ser una medida absoluta, ya que depende de las unidades de medida de las variables y por lo tanto no permite comparar la intensidad de la dependencia lineal de distribuciones que vienen expresadas en unidades diferentes. Como consecuencia, la covarianza no está acotada y por tanto no indica el grado de dependencia lineal entre dos variables, sino únicamente su signo. De ahí la conveniencia de definir un coeficiente que, respetando las ventajas de la covarianza, solucione esta limitación.

**Definición 5.10.** El *coeficiente de correlación lineal de Pearson* se define como el cociente entre la covarianza de $X$ e $Y$ y el producto de las desviaciones típicas de ambas variables, es decir:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \tag{5.5.2}$$

Como consecuencia de su definición, el coeficiente de correlación lineal mantiene el signo de la covarianza que lleva en su numerador (ya que el denominador es siempre positivo al ser producto de desviaciones típicas). Así pues, para variables con relación lineal directa el coeficiente de correlación lineal de Pearson será positivo y para variables con relación lineal inversa dicho coeficiente adoptará signo negativo. Además, ya hemos comprobado que en caso de independencia la covarianza adopta valor nulo, y por tanto también será nulo en ese caso el coeficiente de correlación lineal de Pearson.

**Propiedad 5.8.** *El coeficiente de correlación lineal es una medida acotada entre -1 y 1:*

$$-1 \leq r_{XY} \leq 1$$

Al tratarse de una medida acotada entre -1 y 1, el coeficiente de correlación lineal de Pearson permite medir el grado de dependencia lineal entre dos variables. Así, en caso de *relación lineal directa*, el coeficiente se aproxima a 1 a medida que aumenta la intensidad de dicha relación, presentando valor unitario en el caso de dependencia funcional lineal directa (como en el ejemplo 5.6 del consumo de combustible y el gasto asociado). Obviamente se cumple también $r_{XX} = 1$ ya que cada variable presenta correlación exacta consigo misma. De modo análogo, en el caso contrario (*correlación inversa*) a medida que aumenta la intensidad el coeficiente se acerca a -1 y se obtendría $r_{XY} = -1$ en caso de dependencia funcional lineal inversa.

**Ejemplo 5.12.** El coeficiente de correlación lineal de Pearson entre las variables renta y gasto puede ser calculado a partir de su covarianza y las desviaciones típicas

marginales:

$$S_{XY} = 30,15$$

$$S_X = 10,9$$

$$S_Y = 3,4$$

Así se llega al resultado

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = 0,81$$

que detecta un alto nivel de correlación positiva (81 %) entre la renta y el gasto en viajes.

Conviene tener presente que la existencia de un elevado nivel de correlación lineal no siempre indica la existencia de relaciones de dependencia entre las variables analizadas, ya que a menudo se presentan correlaciones espurias, en las que dos variables que no tienen relación de dependencia entre ellas presentan un elevado nivel de correlación, como consecuencia de su conexión con una tercera variable que en ocasiones se denomina variable escondida o factor de confusión. Este sería el caso si por ejemplo analizamos las ventas de helados y los ingresos hospitalarios en unidades respiratorias, ya que ambas variables podrían verse afectadas por las temperaturas.

**Propiedad 5.9.** *El coeficiente de correlación lineal no se ve afectado por cambios de origen en las variables, es decir, dadas* $X' = X + a$ *e* $Y' = Y + b$ *se cumple* $r_{X'Y'} = r_{XY}$.

**Propiedad 5.10.** *El coeficiente de correlación lineal viene afectado por cambios proporcionales. Más concretamente, si se produce un cambio de escala en las variables, esto es, si* $X' = cX$ *e* $Y' = dY$ *entonces se cumple:*

$$\begin{cases} r_{X'Y'} = r_{XY} \text{ si } cd > 0 \\ \\ r_{X'Y'} = -r_{XY} \text{ si } cd < 0 \end{cases}$$

Ambas propiedades son consecuencia directa de las propiedades de la covarianza (propiedades 5.6 y 5.7) y la desviación típica (propiedad 3.5).

# 6 Regresión lineal simple

Como ya hemos visto en el tema anterior, las relaciones entre variables estadísticas pueden mostrar distintos niveles de intensidad y ser representadas mediante diferentes formas funcionales (lineal, parabólica, hiperbólica, ...).

En este tema nuestro objetivo son las técnicas de regresión simple, que permiten construir modelos para representar la relación existente entre dos variables. Así, nos planteamos buscar la línea que mejor explique el comportamiento de una variable dependiente ($Y$) a partir de una variable explicativa ($X$) que suponemos causa de $Y$. Esta línea, que denominaremos línea de regresión de $Y$ sobre $X$ ($Y/X$), corresponde a un concepto ideal, al que trataremos de aproximamos con la información estadística disponible y sobre la base de algún criterio de optimalidad.

## 6.1. Correlación y regresión

La existencia de un alto nivel de correlación entre dos variables puede detectarse a través de la correspondiente nube de puntos y cuantificarse a partir del coeficiente de correlación lineal, que como ya hemos visto permite conocer la intensidad y el signo de la correlación existente entre $X$ e $Y$. Sin embargo es posible que se observen altos niveles de correlación en distintos tipos de situaciones y como consecuencia de diferentes motivos, tal y como estudiaremos en los apartados que siguen.

> El punto de arranque de los estudios sobre regresión y correlación, está asociado al nombre de *Francis Galton* (1822-1911), cuyas aportaciones a la estadística surgieron en conexión con sus estudios sobre herencia natural, tema muy de actualidad a finales del siglo XIX a raíz de la publicación en 1859 de la obra de su primo Darwin, "*El origen de las especies*".
>
> La noción de correlación, que hemos estudiado en el tema anterior, fue introducida por Galton a raíz de sus investigaciones sobre la identificación de criminales según las relaciones entre diversas características antropométricas, como la altura y la longitud del antebrazo o de los dedos.
>
> Su otra gran aportación, la idea de línea de regresión surgió al medir el tamaño de las semillas de plantas de guisantes madres y de sus descendientes y observar la estabilidad de dicho tamaño. Galton encontró la justificación en que el tamaño de las semillas hijas "revertía" al tamaño promedio (que él cuantificaba a través de la mediana).

**Definición 6.1.** Dada una variable bidimensional $(X, Y)$, se denomina *línea de regresión* a la función que asigna a cada valor $x_i$ de $X$, la correspondiente media condicionada de $Y$, $f(x_i) = \bar{y}/x_i$.

**Ejemplo 6.1.** A modo de ilustración, si consideramos las variables Renta ($X$) y Gasto en viajes ($Y$), es evidente que para familias con una misma renta $x_i$ se pueden observar

niveles muy distintos del gasto en viajes. De ahí que la línea de regresión asigne a cada valor de la renta la media de todos los gastos en viaje observados en ese caso, es decir, la correspondiente media condicionada.

**Propiedad 6.1.** *La línea de regresión es óptima en el sentido mínimo cuadrático, es decir, de todas las posibles funciones de $Y$ respecto a $X$, la que minimiza la suma de los cuadrados de los errores es la que pasa por las medias condicionadas: $f(x_i) = \bar{y}/x_i$.*

Según la propiedad 3.4, la varianza es una medida óptima de dispersión cuadrática. Por tanto si nos limitásemos a un solo valor $x_i$ y los errores se midiesen mediante las desviaciones cuadráticas respecto al valor ideal por donde debe pasar esa línea, obtendríamos que ese óptimo se alcanza en la media condicionada a $x_i$.

Dado que en la práctica únicamente se dispone de un conjunto de observaciones aisladas de $X$, asignando a cada valor $x_i$ la media de $Y$ condicionada al mismo $\bar{y}/x_i$, obtendríamos la línea de regresión para esos valores concretos. En cambio la *línea de regresión* es un concepto teórico que resulta inalcanzable desde una óptica real y empírica, pero podemos aproximarnos a ella mediante el *ajuste mínimo cuadrático*, cuyo planteamiento conlleva, en primer lugar, decidir cuál será la forma más adecuada de la función, para seguidamente obtener los parámetros que la caracterizan.

La representación gráfica mediante la nube de puntos nos servirá de orientación sobre la forma de la función que mejor aproxima las observaciones y así por ejemplo la información sobre la renta $(X)$ y el gasto en viajes $(Y)$ de un grupo de individuos daría lugar a una nube de puntos creciente de forma aproximadamente lineal, que sugiere el ajuste mediante una recta. En otros casos la información disponible y la correspondiente representación gráfica podrán aconsejar ajustes mediante funciones parabólicas, hiperbólicas, exponenciales, etc.

## 6.2. Rectas de regresión mínimo cuadráticas

La regresión de mínimos cuadrados es el método de utilización más generalizada, ya que otros procedimientos como el de ajuste ortogonal o el de los momentos no garantizan las mismas propiedades que el ajuste minimo cuadrático.

Dada una variable bidimensional $(X, Y)$ que toma valores $(x_i, y_j)$ con frecuencias $n_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, m$, la nube de puntos que representa su distribución nos permite decidir -o al menos intuir- cuál puede ser la forma de la función que ajusta esos datos. Esa función genérica dependerá de una serie de parámetros desconocidos y nuestro objetivo será obtener, a partir de los datos disponibles, una estimación de esos parámetros de manera que la función obtenida sea la que mejor aproxime las observaciones.

Así, para cada valor observado de la variable independiente $X\,(x_i)$ podemos considerar dos valores de la variable dependiente $Y$: el valor observado $y_j$ y el valor teórico $y_{ti}$, que se obtiene mediante la función de ajuste.

La diferencia entre el valor observado y el valor teórico recibe el nombre de *error o residuo*, que denotaremos por $e_{ij} = y_j - y_{ti}$ y como puede apreciarse en la figura 6.1

nos proporciona la equivocación cometida al estimar mediante la función de ajuste el valor de la variable $Y$ correspondiente a $x_i$.

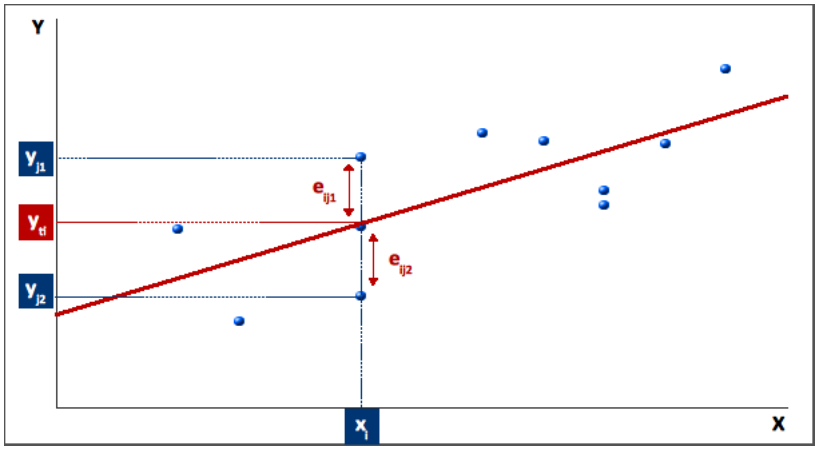



Figura 6.1: Recta de regresión

Parece razonable considerar como función de regresión o ajuste aquélla que proporciona los errores más pequeños, puesto que buscamos la línea que mejor aproxima los datos. Por tanto, los parámetros que la caracterizan deberán ser los que minimicen los errores de ajuste.

Una primera posibilidad podría ser minimizar la suma de los residuos, es decir:

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

Sin embargo siguiendo este camino surgen algunos inconvenientes, puesto que los errores pueden ser positivos o negativos de modo que al sumarlos pueden cancelarse unos con otros proporcionando una idea falsa sobre las equivocaciones realmente cometidas.

Una alternativa a esta situación podría ser minimizar la suma de errores absolutos, es decir:

$$\sum_{i=1}^{k}\sum_{j=1}^{m} |e_{ij}|\, n_{ij}$$

De esta forma se considera únicamente la cuantía de los errores, eliminando su signo, lo cual impide la cancelación de errores opuestos. Sin embargo, este método presenta algunos inconvenientes, ya que no distingue una situación en la que haya muchos errores pequeños de otra situación en la que se presenten pocos errores pero de gran magnitud. Además el método de mínimos errores absolutos presenta dificultades desde el punto de vista matemático por requerir aplicar el cálculo diferencial a expresiones con valores absolutos y, al igual que en la minimización de la suma de errores, no proporcionar al problema una solución única.

El método utilizado será el *ajuste por mínimos cuadrados* o *mínimo cuadrático* y consistirá en minimizar la suma de los cuadrados de los errores, es decir:

$$\sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}^2 n_{ij}$$

En este caso, al elevar al cuadrado los errores individuales, se elimina el signo de los residuos no pudiendo éstos cancelarse y se penalizan más aquéllos que tienen mayor cuantía. Además, este sistema no presenta dificultades de cálculo y proporciona una solución única del problema.

> Introducido por Carl Friedrich Gauss (1777-1855) en 1795 en el marco de sus estudios sobre la distribución de los errores, este método fue también formulado de forma independiente por Adrien Marie Legendre (1752-1833) en 1805.
>
> Aunque inicialmente este procedimiento fue diseñado por Gauss para minimizar los errores de sus estudios astronómicos en la práctica es de aplicación generalizada en ciencias sociales y más concretamente en el ámbito económico, gracias a la interpretación intuitiva de la que es susceptible y a las propiedades que lleva asociadas.

En el caso de que entre dos variables exista una relación lineal, la función de ajuste vendrá dada por una recta $y_{ti} = b_0 + b_1 x_i$ y por tanto el método de mínimos cuadrados nos llevaría a determinar los coeficientes $b_0$ y $b_1$ que minimizan la expresión:

$$E(b_0, b_1) = \sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}^2 n_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m} (y_j - b_0 - b_1 x_i)^2 n_{ij}$$

La condición necesaria de extremo exige que la derivada de la expresión respecto a los parámetros sea nula; en este caso igualamos a cero las derivadas parciales de $E(b_0, b_1)$ respecto a los parámetros $b_0$ y $b_1$, obteniendo las *ecuaciones normales*:

$$\begin{cases} \dfrac{\delta E(b_0, b_1)}{\delta b_0} = -2 \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} (y_j - b_0 - b_1 x_i) n_{ij} = 0 \\ \\ \dfrac{\delta E(b_0, b_1)}{\delta b_1} = -2 \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} (y_j - b_0 - b_1 x_i)\, x_i n_{ij} = 0 \end{cases}$$

Aplicando propiedades del operador suma, el sistema de ecuaciones normales puede ser expresado:

$$\begin{cases} \sum\limits_{j=1}^{m} y_j \overbrace{\sum\limits_{i=1}^{k} n_{ij}}^{n_{\cdot j}} = b_0 \overbrace{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} n_{ij}}^{N} + b_1 \sum\limits_{i=1}^{k} x_i \overbrace{\sum\limits_{j=1}^{m} n_{ij}}^{n_{i\cdot}} \\ \\ \sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} x_i y_j n_{ij} = b_0 \sum\limits_{i=1}^{k} x_i \sum\limits_{j=1}^{m} n_{ij} + b_1 \sum\limits_{i=1}^{k} x_i^2 \sum\limits_{j=1}^{m} n_{ij} \end{cases}$$

$$\begin{cases} \sum_{j=1}^{m} y_j n_{\cdot j} = b_0 N + b_1 \sum_{i=1}^{k} x_i n_{i\cdot} \\ \\ \sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij} = b_0 \sum_{i=1}^{k} x_i n_{i\cdot} + b_1 \sum_{i=1}^{k} x_i^2 n_{i\cdot} \end{cases}$$

y a partir de este sistema, dividiendo ambas ecuaciones por $N$ y teniendo en cuenta las expresiones de cálculo abreviado de $S_X^2$ y $S_{XY}$ (propiedades 3.2 y 5.5), se obtienen los coeficientes $b_0$ y $b_1$ de la recta:

$$b_1 = \frac{S_{XY}}{S_X^2}; \quad b_0 = \bar{y} - b_1 \bar{x}$$

Como consecuencia, la recta de regresión mínimo cuadrática viene dada por la expresión:

$$y - \bar{y} = \frac{S_{XY}}{S_X^2}(x - \bar{x}) \tag{6.2.1}$$

**Ejemplo 6.2.** A partir de la información sobre renta y gasto en viajes disponible en el tema anterior podemos obtener la recta de regresión mínimo cuadrática del gasto respecto a la renta, teniendo en cuenta que:

$$\begin{array}{ll} \bar{x} = 35,5 & \bar{y} = 5,5 \\ S_{XY} = 30,15 & S_X^2 = 118,75 \end{array}$$

Así pues, aplicando las expresiones mínimo cuadráticas anteriormente deducidas se obtienen los coeficientes

$$b_1 = \frac{S_{XY}}{S_X^2} = 0,25; \quad b_0 = 5,5 - 0,25 \times 35,3 = -3,5$$

o equivalentemente a la recta mínimo cuadrática de $Y$ respecto a $X$

$$Y = -3,5 + 0,25X$$

cuya representación gráfica aparece en la figura 6.2.

Resulta interesante interpretar los coeficientes de la recta de regresión: el coeficiente $b_0$ es la ordenada en el origen, mientras el coeficiente $b_1$ es la pendiente de la recta, esto es: $b_1 = \dfrac{\partial Y}{\partial X}$ y por tanto indica la variación producida en la variable $Y$ ante un incremento unitario en la variable $X$.

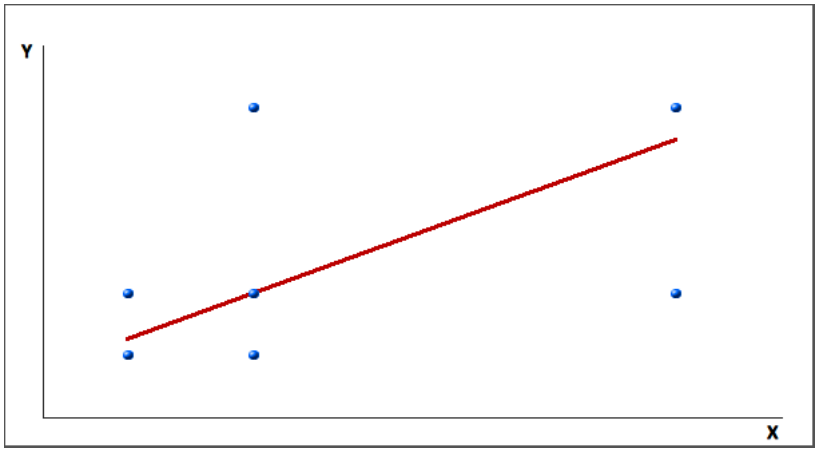



Figura 6.2: Recta de regresión Renta/Gasto en viajes

**Ejemplo 6.3.** El coeficiente de la variable $X$ resulta especialmente interesante por su interpretación económica, que habitualmente puede realizarse en términos de un efecto marginal. Así, si planteamos una recta de regresión del consumo $C$ respecto a la renta $R$, el coeficiente $b_1$ se interpreta como la Propensión Marginal al Consumo, es decir, el incremento que se produce en el Consumo ante un aumento unitario en la Renta disponible.

Por su parte, $b_0$ es el término independiente de la recta que en ocasiones puede ser interpretado como un efecto fijo (por ejemplo, en la recta de Consumo se correspondería con el Consumo autónomo).

En nuestro ejemplo sobre renta y gasto en viajes, el coeficiente estimado $b_1 = 0,25$ indica que por cada euro adicional de renta las familias dedican a gasto en viajes 0,25. Se observa además que el término independiente estimado adopta signo negativo, ya que no tiene sentido plantearse un gasto fijo en viajes para familias de rentas nulas.

**Definición 6.2.** La pendiente de la recta de regresión mínimo cuadrática de $Y$ sobre $X$ se denomina *coeficiente de regresión* de $Y$ sobre $X$, $r_{Y/X}$.

Este coeficiente puede ser interpretado como una medida de la sensibilidad de la variable $Y$ ante cambios unitarios de $X$, y está relacionado con la correlación existente entre $X$ e $Y$.

**Propiedad 6.2.** *El coeficiente de regresión de $Y$ sobre $X$ puede ser expresado como:* $r_{Y/X} = r_{XY}\dfrac{S_Y}{S_X}$

*Demostración.* Partiendo de la definición del coeficiente de regresión basta multiplicar y dividir por la desviación típica de $Y$ para a llegar la relación anterior:

$$r_{Y/X} = \frac{S_{XY}}{S_X^2} = \frac{S_{XY}}{S_X^2}\frac{S_Y}{S_Y} = \frac{S_{XY}}{S_X S_Y}\frac{S_Y}{S_X} = r_{XY}\frac{S_Y}{S_X}$$

□

Esta expresión muestra que existe una estrecha relación entre el coeficiente de regresión y el de correlación lineal, aunque ambas medidas no representan lo mismo. En realidad el coeficiente de correlación lineal trata a las dos variables de forma simétrica mientras el coeficiente de regresión de $Y$ sobre $X$ analiza la respuesta de una variable ante cambios en la otra. Así pues, esta conexión entre el coeficiente de regresión y el de correlación permite interpretar el primero como una extensión del segundo, que añade a la idea de relación lineal una aproximación a la "explicabilidad". Teniendo en cuenta la interpretación del coeficiente de regresión como efecto marginal de $X$ sobre $Y$ podemos establecer una conexión de este término con la *elasticidad de $Y$ respecto a $X$*, que cuantificará el cambio porcentual en la variable $Y$ cuando el valor de $X$ se incrementa un 1 %. Teniendo en cuenta que la elasticidad punto viene dada por la expresión:

$$E_{y/x} = \frac{\partial Y}{\partial X}\frac{x}{y}$$

se observa que el primer factor coincide con el coeficiente de regresión y bastaría sustituir los valores del par $(x, y)$ en el que deseamos evaluar la elasticidad-punto para obtener el correspondiente resultado. Como consecuencia, a lo largo de la recta de regresión la pendiente permanece constante, en cambio la elasticidad varía en cada punto.

**Propiedad 6.3.** *La suma de errores mínimo cuadráticos es nula:*

$$\sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}n_{ij} = 0$$

Esta propiedad garantiza que al realizar un ajuste mínimo cuadrático los errores de estimación globalmente considerados se compensan unos con otros.

*Demostración.* Su comprobación es directa a partir de la primera ecuación normal.

$$\sum_{i=1}^{k}\sum_{j=1}^{m} (y_j - b_0 - b_1 x_i)\, n_{ij} = 0$$

□

**Propiedad 6.4.** *La recta de ajuste mínimo cuadrática pasa por el centro de gravedad de la distribución bidimensional, es decir, por el punto $(\bar{x}, \bar{y})$.*

*Demostración.* Esta propiedad conduce a la igualdad: $\bar{y} = b_0 + b_1\bar{x}$, que se demuestra dividiendo por $N$ la primera ecuación normal. □

Este resultado garantiza que, aunque existan errores en la función ajustada, ésta asocia al valor medio marginal de $X$ la correspondiente media marginal de $Y$, conclusión que claramente aparece relacionada con la "ausencia de error promedio" enunciada en la propiedad anterior.

Teniendo en cuenta esta propiedad, la elasticidad media de $Y$ respecto a $X$ podrá ser obtenida como:

$$E_{\bar{y}/\bar{x}} = \frac{\partial Y}{\partial X}\frac{\bar{x}}{\bar{y}}$$

expresión que, partiendo del centro de gravedad de la recta $(\bar{x}, \bar{y})$, cuantifica el efecto porcentual que se produce en $Y$ ante un incremento de un 1 % en $X$. Así, a modo de ejemplo si a partir de la información anterior deseamos calcular la elasticidad media del gasto en viajes respecto a la renta, se obtiene el resultado: $E_{\bar{y}/\bar{x}} = 25 \times \frac{35,5}{5,5} = 1,6$.

**Propiedad 6.5.** *Si dos variables $X$ e $Y$ son incorreladas, la recta de regresión es paralela al eje de abscisas: $y = \bar{y}$.*

*Demostración.* La demostración de esta propiedad es inmediata teniendo en cuenta que si las variables $X$ e $Y$ son incorreladas su covarianza será nula y por tanto también será nulo el coeficiente de regresión $r_{Y/X}$, obteniéndose una recta de regresión paralela al eje de abscisas.

$$r_{XY} = 0 \Rightarrow y = \bar{y}$$

□

Esta propiedad resulta lógica teniendo en cuenta que la línea de regresión se basa en las medias condicionadas. Ahora bien, si la variable $Y$ no está correlacionada con $X$ las medias condicionadas coincidirían con la marginal.

Si bien este tema va referido al modelo de regresión lineal, el planteamiento anterior podría ser extendido a situaciones en las que se desea explicar una variable dependiente $Y$ a partir de una variable explicativa $X$ mediante una función no lineal. Así, si nos interesa explicar la producción de un output $Y$ a partir de algún input $X$ es posible que la nube de puntos no muestre una forma lineal, sugiriendo por ejemplo el ajuste a una función potencial tipo Cobb-Douglas. En este caso, el modelo podría ser formulado como:

$$Y = b_0 X^{b_1}$$

expresión que puede ser linealizada mediante una transformación logarítmica, dando lugar a:

$$\log(Y) = \log(b_0) + b_1 \log(X)$$

siendo posible aplicar sobre este modelo linealizado las expresiones deducidas para la recta de regresión mínimo cuadrática.

Además, es interesante destacar que en este caso la elasticidad es constante y adopta valor $b_1$, es decir, existe una respuesta porcentual constante de magnitud $b_1$ en el output $Y$ ante un incremento de 1 % en el input $X$.

## 6.3. Análisis de la bondad de modelos

Una vez aplicadas las técnicas del análisis de regresión dispondremos de una función que explica en cierta medida la variación de la variable dependiente $Y$ según el

comportamiento de $X$. Si ahora nos preguntamos hasta qué punto esta función proporciona una descripción adecuada del comportamiento de las variables, para analizar la bondad de ese modelo sólo disponemos del conjunto de las observaciones $(x_i, y_j)$.

En el peor de los casos, esto es, en aquellas situaciones en las que no se dispone de información sobre ninguna variable relevante que nos ayude a explicar el comportamiento de $Y$, podemos tomar como mejor explicación de esta variable su media aritmética $\bar{y}$. El error en que incurrimos como consecuencia de esta elección vendría dado en función de las desviaciones de los valores observados respecto a la media: $(y_j - \bar{y})$ y tomando este extremo como referencia, la consideración de un modelo explicativo $y_t = f(x)$ nos servirá para reducir el error. Además, esta reducción se produce en mayor medida cuanto mayor sea la bondad del ajuste.

Designando por $y_{ti}$ a los valores teóricos sobre la función de ajuste, $y_{ti} = f(x_i)$, tendremos:

$$y_j - \bar{y} = \overbrace{(y_j - y_{ti})}^{e_{ij}} + (y_{ti} - \bar{y})$$

es decir, la desviación respecto a la media para cada observación puede descomponerse como suma del error que queda tras efectuar la regresión $e_{ij} = (y_j - y_{ti})$ y la desviación que es explicada por la regresión $(y_{ti} - \bar{y})$.

Partiendo de la igualdad anterior y elevando al cuadrado los dos miembros se obtiene al promediar:

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(y_j - \bar{y})^2 f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m}\left[(y_j - y_{ti}) + (y_{ti} - \bar{y})\right]^2 f_{ij}$$

Desarrollando el cuadrado del segundo miembro se llega a:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{m}(y_j - \bar{y})^2 f_{ij} \quad &= \sum_{i=1}^{k}\sum_{j=1}^{m}(y_j - y_{ti})^2 f_{ij} + \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ti} - \bar{y})^2 f_{ij} \\ &\quad + 2\sum_{i=1}^{k}\sum_{j=1}^{m}(y_j - y_{ti})(y_{ti} - \bar{y}) f_{ij}\end{aligned}$$

$$S_Y^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}^2 f_{ij} + \sum_{i=1}^{k}(y_{ti} - \bar{y})^2 f_{i.} + 2\sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ti} - \bar{y})\, e_{ij} f_{ij}$$

Tal y como hemos visto, el criterio mínimo-cuadrático de estimación de parámetros trata de evitar incurrir en errores sistemáticos, ya sean positivos o negativos, por lo que habitualmente la media de los errores es nula. De hecho, siempre que el modelo ajustado incluya término independiente, la primera ecuación normal garantiza que $\bar{e} = 0$.

Teniendo en cuenta esta consideración, el primer sumando de la expresión obtenida corresponde a la varianza del error; el segundo se identifica con la varianza de la variable teórica puesto que $\bar{y}_t = \bar{y} - \bar{e} = \bar{y}$, y el tercer sumando contiene a la covarianza entre el error y la variable teórica:

$$S_Y^2 = S_e^2 + S_{Y_t}^2 + 2S_{Y_t,e} \tag{6.3.1}$$

En los ajustes habituales la covarianza $S_{Y_t,e}$ será nula como consecuencia del propio concepto de error, que no puede guardar relación con la variable a explicar.

**Definición 6.3.** Se denomina *varianza residual* y se denota por $S_e^2$, al valor de la expresión:

$$S_e^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}^2 f_{ij}$$

**Definición 6.4.** Se denomina *varianza explicada* o debida a la regresión y se denota por $S_{Y_t}^2$, al valor de la expresión:

$$S_{Y_t}^2 = \sum_{i=1}^{k} (y_{ti} - \bar{y})^2 f_{i.} \tag{6.3.2}$$

**Propiedad 6.6.** *En el caso de un ajuste lineal, la varianza de la variable dependiente $Y$ puede ser expresada como suma de la varianza explicada y la varianza residual, es decir:*

$$S_Y^2 = S_{Y_t}^2 + S_e^2$$

*Esta igualdad se obtiene teniendo en cuenta que en el caso lineal $S_{Y_t,e} = 0$.*

*Demostración.*

$$S_{Y_t,e} = \sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}\,(y_{ti} - \bar{y})\,f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{m} [e_{ij}\,(b_0 + b_1 x_i - \bar{y})]\,f_{ij} = b_1 \sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij} x_i f_{ij} = 0$$

puesto que la última suma equivale a la segunda ecuación normal del ajuste por mínimos cuadrados. □

Este resultado parece lógico puesto que en caso contrario nos encontraríamos con que aún permanece cierta relación lineal entre los errores y la variable teórica, lo que haría sospechar que la función lineal estimada podría ser mejorada.

La relación obtenida nos permite interpretar la variación de $Y$ respecto a su valor medio a través de dos componentes, uno que nos indica la variación de $Y$ que es capaz de explicar el modelo, y otro que representa la parte de variación de $Y$ que permanece sin explicar.

Partiendo de la descomposición anterior de la varianza de Y:

$$S_Y^2 = S_{Y_t}^2 + S_e^2$$

bastaría dividir los dos miembros de la igualdad por $S_Y^2$ para obtener:

$$1 = \frac{S_{Y_t}^2}{S_Y^2} + \frac{S_e^2}{S_Y^2} \tag{6.3.3}$$

**Definición 6.5.** El *coeficiente de determinación* $R^2$ se define como la proporción de variación de $Y$ explicada por el modelo teórico y viene dado por la expresión:

$$R^2 = \frac{S^2_{Y_t}}{S^2_Y} = 1 - \frac{S^2_e}{S^2_Y} \tag{6.3.4}$$

**Propiedad 6.7.** *El coeficiente de determinación está acotado entre 0 y 1:*

$$0 \leq R^2 \leq 1$$

*Demostración.* Dado que el coeficiente de determinación puede ser expresado como un cociente de varianzas, su resultado es siempre no negativo.

Por otra parte, la relación 6.3.3 indica que la suma de dos cuadrados -uno de los cuales es $R^2$- es unitaria, y en consecuencia el valor de dicho coeficiente está acotado superiormente por 1. □

**Ejemplo 6.4.** A partir de nuestra información sobre renta y gasto en viajes, la varianza total de $Y$ que adoptaba valor 11,55 puede descomponerse en varianza explicada y residual cuyos resultados serían

$$S^2_Y = 11,55, \quad S^2_{Y_t} = 7,65, \quad S^2_e = 3,9$$

Como consecuencia se obtiene el coeficiente de determinación $R^2 = 1 - \dfrac{3,9}{11,55} = 0,66$, que permite afirmar que un 66 % de las variaciones del gasto en viajes se explican mediante la recta mínimo cuadrática a partir de la renta. Así pues, podemos concluir que la recta estimada tiene una capacidad explicativa aceptable.

**Propiedad 6.8.** *El coeficiente de determinación es nulo cuando el modelo no aporta ninguna explicación sobre el comportamiento de $Y$.*

*Demostración.* La propiedad se comprueba teniendo en cuenta que en este caso la mejor explicación de $Y$ es su media $y_{ti} = \bar{y}$, de donde $S^2_{Y_t} = 0$. □

**Propiedad 6.9.** *Cuando el ajuste es perfecto se obtiene $R^2 = 1$.*

*Demostración.* En este caso, todos los errores serían nulos y por tanto $S^2_e = 0$. □

Gracias a su interpretación, el coeficiente de determinación permite evaluar la capacidad explicativa o bondad de un modelo, aproximándose este coeficiente a 1 a medida que aumenta la proporción de cambios en $Y$ que son explicados por la variable $X$. No obstante, debemos ser prudentes a la hora de realizar afirmaciones relativas a la causalidad a partir de los resultados de este coeficiente, ya que como hemos señalado en el tema anterior, en ocasiones pueden observarse relaciones espurias entre $X$ e $Y$, que no se deben a la existencia de causalidad sino a la presencia de una tercera variable relacionada con $X$ e $Y$.

Un caso anecdótico de relación espuria entre variables se debe a Jerzy Neyman (1894-1981). Este estadístico de origen polaco analizó en 1952 la tasa de nacimientos y la población de cigüeñas en varias regiones, y encontró un alto coeficiente de correlación entre estas variables. Lógicamente, esta elevada correlación no permite afirmar que las cigüeñas sean causa de los nacimientos, sino que se debe a la conexión de ambas características con la renta, que actuaría como "variable oculta".

Otras ilustraciones conocidas se deben a George Udny Yule (1871-1951) quien estudió en Inglaterra y Gales la evolución conjunta de la tasa de mortalidad y el porcentaje de matrimonios, obteniendo coeficientes de correlación lineal de 0,95.

**Propiedad 6.10.** *En el caso lineal el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación lineal; es decir, se cumple* $R^2 = r_{XY}^2$

*Demostración.* Teniendo en cuenta que en el caso lineal los valores teóricos obtenidos mediante la aplicación del método mínimo cuadrático vienen dados por la expresión:

$$y_{ti} = f(x_i) = \bar{y} + \frac{S_{XY}}{S_X^2}(x_i - \bar{x})$$

el coeficiente de determinación vendrá dado por:

$$R^2 = \frac{\sum_{i=1}^{k} \left(\frac{S_{XY}}{S_X^2}\right)^2 (x_i - \bar{x})^2 f_{i.}}{S_Y^2} = \frac{S_{XY}^2}{S_X^2 S_Y^2} = r_{XY}^2$$

□

Cuando tenemos únicamente dos observaciones y llevamos a cabo un ajuste lineal, $R^2$ será igual a 1 cualesquiera que sean las variables implicadas. Por tanto, a la hora de extraer conclusiones sobre la validez de una recta de regresión a partir del coeficiente de determinación, debemos tener en cuenta que el valor de este coeficiente tendrá más fiabilidad cuanto mayor sea el número disponible de datos.

Este aspecto enlaza con el concepto de *grados de libertad* en la determinación de modelos, que va referido al margen de elección de los valores de los parámetros a partir del conjunto de datos disponible. En el caso extremo planteado, el número de grados de libertad es 0, ya que se han estimado dos parámetros a partir de dos datos; en general, el número de grados de libertad vendrá determinado por la diferencia entre el número de observaciones y el número de parámetros a estimar a partir de ellas.

**Ejemplo 6.5.** Supongamos que calificamos un trabajo realizado en equipo por cuatro estudiantes, siendo la nota media del trabajo de 7,5. Podemos preguntarnos cuántos grados de libertad tenemos para asignar las notas individuales. Supongamos que asignamos al primer estudiante un 8, al segundo un 7 y al tercero un 9; la nota del cuarto ya no la podemos elegir, porque como su media es 7,5, se tendrá:

$$7,5 = \frac{8 + 7 + 9 + x}{4}$$

de donde despejando, se tiene: $x = 6$, por lo tanto solo tenemos 3=4-1 grados de libertad.

En general, por cada ecuación o parámetro que tengamos que despejar, perderemos un grado de libertad, de ahí la frase anterior.

Parece lógico, por tanto, que cuanto mayor sea el número de grados de libertad más informativo sea el resultado aportado por el coeficiente de determinación $R^2$.

En el contexto del análisis de la bondad del modelo tiene interés dar una medida del margen de error cometido al estimar los valores observados a través de los teóricos y este objetivo puede alcanzarse también considerando la varianza residual $S_e^2$ o su raíz cuadrada. Cuanto más alejadas estén las observaciones de la línea estimada, o equivalentemente, cuanto mayores sean los errores cometidos, mayor será la varianza residual.

> Dado que esta medida se corresponde con la varianza del error, según la interpretación de la desviación típica podemos establecer bandas en torno a la línea estimada de radio $kS_e$, de modo que es posible asegurar que una cierta proporción de observaciones quedan dentro de esas bandas (por ejemplo, en la de radio $2S_e$ se hallarán al menos el 75 % de las observaciones).

## 6.4. Predicción con modelos causales

La capacidad de efectuar pronósticos es un elemento clave tanto en el desarrollo de políticas socioeconómicas como en la toma de decisiones empresariales ya que, en la medida en que seamos capaces de anticipar comportamientos futuros, incrementaremos nuestras posibilidades de éxito. Éste sería el caso del gobierno, que necesita disponer de previsiones sobre la demanda sanitaria y de educación con el fin de efectuar las inversiones oportunas o de la dirección de una empresa que decidirá reforzar los recursos destinados a la producción ante unas previsiones de fuertes alzas en las ventas.

La carencia de predicciones, o la existencia de fallos acusados en las predicciones disponibles pueden tener consecuencias graves tales como inversiones innecesarias o excedentes de stock con las consiguientes pérdidas económicas. De ahí la importancia de cuidar la calidad de las predicciones, que dependerá tanto de la información estadística en la que se basan como de los instrumentos y técnicas empleados.

Adoptando como referencia un modelo de regresión y asumiendo que la variable explicativa $X$ adopta un valor $x_0$, la predicción de $Y$ se obtiene como el valor teórico de la variable $Y$ condicionado a dicho valor de la variable explicativa, es decir, $y_{t0} = b_0 + b_1 x_0$.

La existencia de relaciones causales entre varios caracteres proporciona un soporte adecuado para la realización de predicciones, al permitirnos estimar el valor de la variable efecto $Y$ a partir de información referida a la causa $X$. En realidad, la predicción aúna dos ópticas: la estimación de los parámetros del modelo considerado y la cuantificación del valor previsto, ya que los niveles en los que se sitúen las variables explicativas condicionarán el resultado -y el riesgo- de la predicción obtenida.

De ahí que existan algunos rasgos diferenciales entre la regresión mínimo cuadrática que hemos visto en apartados anteriores y la predicción asociada a estos modelos, en la que debemos contemplar varias fuentes de riesgo:

- Por una parte, la distancia que separa el modelo de la realidad (los errores de la regresión $e_{ij}$).
- Por otra parte, los supuestos en los que se basa el método de estimación (mínimo cuadrática en general).
- Por último, la asignación de un valor numérico determinado a la variable explicativa $X$ (ya que la predicción irá condicionada a ese valor).

Así pues, si nos planteamos qué garantías tienen nuestras previsiones, aunque somos conscientes de que nunca podremos "adivinar" el verdadero valor de la variable $Y$, parece lógico pensar que, si asumimos modelos teóricos adecuados y realizamos ajustes fiables de los mismos a partir de información estadística de buena calidad, podremos llegar a predicciones que nos merezcan un alto nivel de confianza.

Comenzando por el primero de estos aspectos, la validez conceptual, nos interesarán desde el punto de vista estadístico únicamente las predicciones basadas en modelos, de modo que exista algún supuesto teórico o hipótesis que avale nuestras actuaciones. Si la información de la que disponemos parece confirmar alguna hipótesis económica, plantearíamos el correspondiente modelo para proceder a su estimación. Sin embargo, somos conscientes de que los errores y lagunas de la información económica pueden llegar a afectar a este planteamiento, encontrándonos en ocasiones con que no podemos acceder a las cifras de la hipotética causa o bien que éstas son escasamente fiables.

Por lo que se refiere a la bondad de ajuste, hemos visto anteriormente que la medida más habitual de la bondad es el coeficiente de determinación $R^2$, que será por tanto un buen indicador de la fiabilidad de las predicciones.

Sin embargo, debemos tener presente que esta medida ha sido calculada a partir de la información disponible en la muestra, por lo que no puede garantizarse su validez cuando nos alejemos del recorrido de nuestra distribución. Así, en la figura 6.3, la recta ha sido estimada con observaciones comprendidas entre 12 y 22, mientras las predicciones se realizan para el valor $X = 40$ que se sitúa lejos de este recorrido y podría en consecuencia no adaptarse al mismo patrón de comportamiento estimado (lo mismo podría suceder si las predicciones van referidas a un valor bajo de $X$, pudiendo incluso obtenerse predicciones negativas para $Y$).

En definitiva, la utilización de un modelo perfecto desde el punto de vista teórico y con coeficiente de determinación elevado no nos autoriza a su utilización indiscriminada para fines predictivos.

> Además de acompañar cada predicción de una medida de su fiabilidad, es aconsejable proporcionar márgenes de error o equivalentemente bandas de confianza entre cuyos extremos se encontrará "casi con total seguridad" el valor verdadero. Este objetivo puede conseguirse en el caso de los modelos causales utilizando $S_e$, que es la raíz cuadrada de la varianza residual y puede ser interpretada como una medida de dispersión.

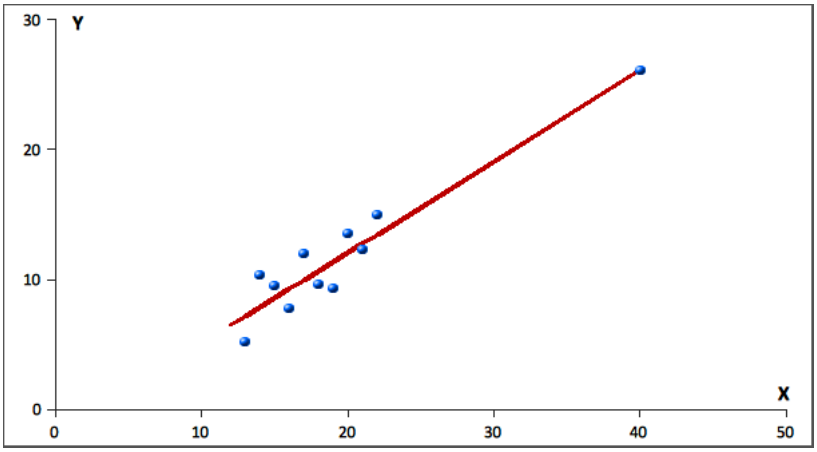



Figura 6.3: Predicción

# 7 Regresión lineal múltiple

El ámbito socioeconómico muestra una amplia diversidad de variables como renta, gasto, precios, salarios, ... que presentan relaciones entre sí. Como consecuencia resulta prácticamente imposible aislar comportamientos ya que todo tiende a depender -en mayor o menor grado- de todo.

Esta consideración nos lleva a efectuar un planteamiento más general de las técnicas de regresión y correlación que permita tratar situaciones en las que haya más de dos variables implicadas. Así, por ejemplo, el consumo no sólo dependerá de la renta sino también de los precios del bien analizado y otros bienes (tanto complementarios como sustitutivos); las ventas de una empresa estarán relacionadas no sólo con la inversión en I+D sino también con la política de precios o el gasto en publicidad. De hecho, parte del error que aparecía en las regresiones lineales simples estudiadas en el tema anterior podría deberse a la omisión de otras variables explicativas, y por tanto la incorporación de nuevas variables causales al modelo debería suponer una mejora sustancial en su capacidad explicativa.

El procedimiento que seguiremos en este tema es una extensión del empleado en la regresión simple, en el que surgen nuevos conceptos y problemas específicos, derivados fundamentalmente de la existencia de correlación entre las variables explicativas. Este último punto dará lugar a la introducción de diferentes indicadores de correlación dependiendo de las variables consideradas.

> Los métodos de regresión y correlación múltiple fueron introducidos por Karl Pearson, y desarrollados posteriormente por su discípulo G. Udney Yule (1871-1951), a quien se debe la conexión entre el concepto de regresión y la técnica de ajuste por mínimos cuadrados así como la definición de los coeficientes de correlación parcial y múltiple.

A lo largo del capítulo nos centraremos en el desarrollo de la regresión múltiple con dos variables explicativas. La consideración de un mayor número de variables, si bien no supone ninguna diferencia sustancial desde el punto de vista conceptual, conlleva un notable aumento en la complejidad de los cálculos, que hace imprescindible el empleo de notación matricial.

## 7.1. Planteamiento de la regresión múltiple

El objetivo de la regresión múltiple será construir un modelo explicativo de una variable $Y$ en términos de un conjunto de variables causales que, sin pérdida de generalidad, reduciremos a dos: $X_1$, $X_2$.

Nuestro planteamiento en este apartado constituye una generalización del efectuado al introducir el análisis de regresión simple: obtener el valor promedio de la variable

dependiente $Y$ condicionado a ciertos comportamientos de las variables explicativas $X_1$ y $X_2$. Ello dará lugar a la definición de la función de regresión mínimo-cuadrática $Y = f(X_1, X_2)$ como aquélla que aproxima el verdadero valor de $Y$ cuando $X_1 = x_{1i}$ y $X_2 = x_{2i}$ a través de la media condicionada de las observaciones de $Y$: $\bar{y}/x_{1i}, x_{2i}$.

Como puede observarse en la figura 7.1, esta definición de $f$ daría lugar a una superficie en $\mathbb{R}^3$ y si introducimos el supuesto de que las medias condicionadas se encuentran sobre un plano, es decir, que la contribución de cada variable independiente a la explicación de Y es de tipo lineal, la función de regresión será de la forma:

$$y_{ti} = f(x_{1i}, x_{2i}) = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

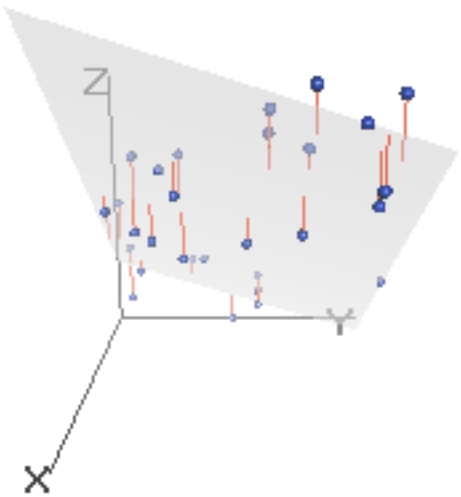



Figura 7.1: Hiperplano de regresión

De aquí en adelante nos centraremos en este tipo de funciones, ya sea porque la regresión es lineal o porque mediante una transformación de las variables es reducible a lineal. De modo similar al visto en el tema anterior para la recta de regresión mínimo cuadrática, la forma de obtener esta función o plano de regresión consistirá en buscar aquellos parámetros $b_0, b_1$ y $b_2$ que minimicen la suma de los errores cuadráticos.

Si en vez de tener dos variables independientes $X_1$ y $X_2$ se tienen $m$ variables $(X_1, X_2, ..., X_m)$ la función lineal de ajuste, que recibe el nombre de *hiperplano* de ajuste, será de la forma:

$$y_{ti} = f(x_{1i}, x_{2i}, \cdots, x_{mi}) = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_m x_{mi}$$

y se obtendrá por el método de mínimos cuadrados al igual que en los casos anteriores.

## 7.2. Plano de regresión mínimo cuadrático

La función lineal que aproxima las observaciones de $Y$ a partir de dos variables explicativas $X_1$ y $X_2$ mediante la expresión $y_{ti} = b_0 + b_1 x_{1i} + b_2 x_{2i}$ recibe el nombre de plano de ajuste. Para calcular los parámetros $b_0$, $b_1$, $b_2$ que determinan la ecuación del plano de regresión a partir de un conjunto de $N$ observaciones con frecuencias unitarias $(x_{1i}, x_{2i}, y_i)$, $i = 1, ..., N$ recurriremos a la técnica de ajuste por mínimos cuadrados. Los errores cometidos al aproximar los valores observados $y_i$ por los teóricos $y_{ti}$ vendrán dados por las desviaciones entre ambos:

$$e_i = y_i - y_{ti} = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i})$$

Por tanto, el valor de los parámetros será el resultado de minimizar:

$$E(b_0, b_1, b_2) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i})^2$$

problema que equivale a resolver el sistema de ecuaciones normales:

$$\frac{\partial E(b_0, b_1, b_2)}{\partial b_0} = -2 \sum_{i=1}^{N} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) = 0$$

$$\frac{\partial E(b_0, b_1, b_2)}{\partial b_1} = -2 \sum_{i=1}^{N} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i}) x_{1i} = 0$$

$$\frac{\partial E(b_0, b_1, b_2)}{\partial b_2} = -2 \sum_{i=1}^{N} (y_j - b_0 - b_1 x_{1i} - b_2 x_{2i}) x_{2i} = 0$$

Sin más que dividir por $N$ las ecuaciones anteriores, la primera nos garantiza que el error medio es nulo, $\bar{e} = 0$, y nos permite obtener el valor del término independiente una vez calculados $b_1$ y $b_2$:

$$b_0 = \bar{y} - b_1 \bar{x_1} - b_2 \bar{x_2}$$

Sustituyendo este valor en las dos ecuaciones restantes se llega a las relaciones:

$$b_1 S_{X_1}^2 + b_2 S_{X_1, X_2} = S_{Y, X_1}$$

$$b_1 S_{X_1, X_2} + b_2 S_{X_2}^2 = S_{Y, X_2}$$

de donde se obtienen los coeficientes $b_1$ y $b_2$.

**Definición 7.1.** Los *coeficientes de regresión parcial* $b_1$ y $b_2$ vienen dados por las expresiones:

$$b_1 = \frac{S^2_{X_2} S_{Y,X_1} - S_{X_1,X_2} S_{Y,X_2}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1,X_2}}$$

$$b_2 = \frac{S^2_{X_1} S_{Y,X_2} - S_{X_1,X_2} S_{Y,X_1}}{S^2_{X_1} S^2_{X_2} - S^2_{X_1,X_2}}$$

La interpretacion de estos coeficientes de regresión parcial resulta de interés, ya que nos indican cuál es el efecto marginal de cada variable explicativa sobre la variable dependiente. Así, si calculamos las derivadas parciales de $Y$ respecto a las variables explicativas $X_1$ y $X_2$ se tiene:

$$\frac{\partial Y}{\partial X_1} = b_1; \quad \frac{\partial Y}{\partial X_2} = b_2$$

**Ejemplo 7.1.** Si realizamos una regresión mínimo cuadrática de los gastos en función de la renta $X_1$ y los precios $X_2$, es de esperar que el coeficiente $b_1$ presente signo positivo (dado que la relación entre gasto y renta es directa), mientras para $b_2$ se espera signo negativo (recogiendo así la existencia de una relación inversa entre gasto y precios).

Más concretamente, si la estimación mínimo cuadrática proporciona como resultado los coeficientes $b_1 = 0,7$ y $b_2 = -0,3$ podemos afirmar que ante un aumento unitario en la renta (ceteris paribus los precios) se espera que el gasto aumente en 0,7 unidades. De modo análogo, se observa que un aumento unitario en los precios (ceteris paribus la renta) origina una reducción de 0,3 unidades en el gasto.

**Propiedad 7.1.** *Los coeficientes de regresión parcial pueden ser expresados como:*

$$b_1 = \frac{S_Y}{S_{X_1}} \left( \frac{r_{Y,X_1} - r_{Y,X_2} r_{X_1,X_2}}{1 - r^2_{X_1,X_2}} \right)$$

$$b_2 = \frac{S_Y}{S_{X_2}} \left( \frac{r_{Y,X_2} - r_{Y,X_1} r_{X_1,X_2}}{1 - r^2_{X_1,X_2}} \right)$$

*donde $r_{Y,X_1}$ y $r_{Y,X_2}$ son los coeficientes de correlación lineal entre $Y$ y las variables explicativas de la regresión:*

$$r_{Y,X_1} = \frac{S_{YX_1}}{S_Y S_{X_1}}; \; r_{Y,X_2} = \frac{S_{YX_2}}{S_Y S_{X_2}}$$

Los coeficientes de regresión parcial reflejan, por tanto, la transformación de las unidades de cada variable explicativa $X_i$ a la escala de variación de la variable explicada corregida por un factor indicador de la intensidad del efecto que cada $X_i$ tiene sobre $Y$ una vez eliminada la parte de influencia común con la otra variable independiente. Estas expresiones justifican también que los coeficientes de regresión obtenidos en el modelo múltiple no coincidan con los que se obtendrían en los modelos simples que

incluyeran como explicativas las variables $X_1$ y $X_2$ respectivamente. Este resultado es lógico, puesto que la omisión de una de las variables explicativas induce un sesgo en el valor de los coeficientes de regresión simple; únicamente daría lugar a resultados coincidentes si la covarianza entre las variables explicativas fuera nula.

**Propiedad 7.2.** *Si las variables explicativas no están correlacionadas, los coeficientes de regresión parcial del modelo múltiple coinciden con los coeficientes de regresión del modelo lineal simple.*

*Demostración.* A partir de la expresión de los parámetros $b_i$ en términos de los coeficientes de correlación lineal se comprueba fácilmente que, si las dos variables explicativas $X_1$ y $X_2$ están incorreladas, entonces:

$$r_{X_1,X_2} = 0 \;\Rightarrow\; b_1 = \frac{S_Y}{S_{X_1}}\left(\frac{r_{Y,X_1}}{1}\right) = \frac{S_{Y,X_1}}{S^2_{X_1}}$$

expresión que coincide con la vista en el tema anterior para el coeficiente de regresión en un modelo lineal simple. □

Es necesario tener presente que, una vez obtenidos los dos coeficientes de regresión del modelo $y_{ti} = b_0 + b_1x_{1i} + b_2x_{2i}$ los resultados numéricos de estos coeficientes no pueden ser comparados entre sí, ya que las variables explicativas vendrán expresadas generalmente en unidades diferentes. Para llevar a cabo este tipo de comparaciones se introducen los coeficientes beta mediante una estandarización de los coeficientes de regresión parcial:

$$\beta_i = \frac{b_i S_{X_i}}{S_Y}$$

## 7.3. Análisis de la bondad de modelos múltiples

El hecho de que en el análisis de regresión múltiple se consideren modelos explicativos en los que intervienen dos o más causas da lugar a que puedan confundirse las contribuciones de cada una de ellas a la explicación del efecto o variable dependiente. De ahí que distingamos tres tipos de coeficientes de determinación asociados al análisis de regresión múltiple.

Comenzaremos introduciendo el coeficiente de determinación múltiple como una generalización del ya conocido para la regresión simple. Se trata de buscar una medida del nivel de explicación alcanzado a través del modelo lineal gracias a la aportación de todas las variables explicativas consideradas conjuntamente. Así, mediante un desarrollo análogo al efectuado en el tema anterior podemos descomponer la variación total de $Y$ como suma de la parte explicada por el plano de regresión y la parte residual que queda sin explicar:

$$S^2_Y = S^2_{Yt} + S^2_e$$

donde ahora los valores teóricos son los obtenidos sobre el plano: $y_{ti} = b_0 + b_1 x_{1i} + b_2 x_{2i}$ y la varianza residual es la varianza de la variable error, que toma valores : $e_i = y_i - y_{ti}$.

**Definición 7.2.** El *coeficiente de determinación múltiple* $R^2$ es la proporción de variación de $Y$ explicada por el modelo y viene definido por la expresión:

$$R^2_{Y,X_1,X_2} = \frac{S^2_{Yt}}{S^2_Y} \tag{7.3.1}$$

cuyo resultado está acotado entre 0 y 1, acercándose a este último valor cuanto mayor sea la parte de variabilidad de $Y$ que explica el modelo.

Las propiedades del coeficiente de determinación son similares a las vistas para el modelo lineal simple. Al mismo tiempo $R^2_{Y,X_1,X_2}$ puede ser considerado como una medida de correlación entre $Y$ y las variables explicativas $X_1$ y $X_2$ en su conjunto.

Al introducir los coeficientes de regresión parcial planteamos la comparación entre ambos y la definición de los coeficientes beta con el fin de analizar qué variable tenía mayor importancia relativa a la hora de explicar el comportamiento de $Y$. Trataremos ahora de cuantificar la aportación de cada una de ellas.

Mediante el coeficiente de determinación múltiple anteriormente definido se cuantifica la aportación conjunta de la totalidad de variables explicativas sin distinguir qué parte es debida a cada una de ellas. En este sentido, es posible también considerar los *coeficientes de determinación simples* definidos en el capítulo anterior.

$$R^2_{Y,X_1} = r^2_{Y,X_1} = \frac{S^2_{Y,X_1}}{S^2_Y S^2_{X_1}}$$

$$R^2_{Y,X_2} = r^2_{Y,X_2} = \frac{S^2_{Y,X_2}}{S^2_Y S^2_{X_2}}$$

Sin embargo, debemos tener presente que estos coeficientes no aprovechan las ventajas de la consideración conjunta de las dos variables explicativas y que puede haber una parte "conjunta" de aportación de $X_1$ y $X_2$ a la explicación de $Y$. Esto justifica que si sumamos los dos coeficientes de determinación simples el resultado será generalmente superior al 100 %.

Los coeficientes de determinación simples sólo reflejarán la parte de variación de $Y$ explicada exclusivamente por cada $X_i$ cuando no exista relación lineal alguna entre las variables causales, es decir, cuando $r_{X_1,X_2} = 0$ , ya que cuando las variables son incorreladas, se obtiene:

$$R^2_{Y,X_1,X_2} = R^2_{Y,X_1} + R^2_{Y,X_2}$$

Sin embargo, ésta no es la situación más frecuente, ya que por lo general existirá correlación entre las variables independientes y parte de su contribución a la explicación de $Y$ será común. Ello nos lleva a introducir los coeficientes de determinación parcial, que nos permitirán conocer el nivel de explicación de cada una de las $X_i$ habiendo eliminado la influencia del resto.

**Definición 7.3.** El *coeficiente de determinación parcial* entre $Y$ y $X_2$ se define como:

$$R^2(Y, X_2/X_1) = \frac{R^2_{Y,X_1,X_2} - R^2_{Y,X_1}}{1 - R^2_{Y,X_1}} \tag{7.3.2}$$

que representa la proporción de variación residual del modelo de regresión simple de $Y$ sobre $X_1$ que es explicada gracias a la inclusión de $X_2$, y adopta valores comprendidos entre 0 y 1.

Esta medida adopta como punto de partida la regresión simple de $Y$ sobre $X_1$ y analiza la ganancia de explicación que se obtiene cuando adicionalmente se introduce $X_2$, esto es, $S_{Y_t^2}(X_2/X_1)$, que se obtendrá como diferencia entre la varianza explicada del modelo que incluye a las dos variables y la correspondiente al modelo que únicamente incluye a $X_1$:

$$S^2_{Y_t}(X_2/X_1) = S^2_{Y_t}(X_1, X_2) - S^2_{Y_t}(X_1)$$

Puesto que la parte de variación de $Y$ no explicada por $X_1$ viene dada por la varianza residual correspondiente $S^2_e(X_1)$, podríamos expresar la ganancia de explicación en términos relativos mediante el cociente:

$$\frac{S^2_{Y_t}(X_1, X_2) - S^2_{Y_t}(X_1)}{S^2_e(X_1)}$$

expresión en la que, dividiendo numerador y denominador por $S^2_Y$, se llega al coeficiente de determinación parcial anteriormente definido.

Por lo que se refiere a la acotación, debemos tener en cuenta que lo peor que puede ocurrir al incorporar $X_2$ al modelo de regresión es que esta nueva variable explicativa no aporte nada nuevo a la explicación de $Y$ dada por $X_1$ con lo que $S_{Y_t^2}(X_1) \leq S_{Y_t^2}(X_1, X_2)$ y por consiguiente $R^2(Y, X_2/X_1) \geq 0$. Por otra parte, acudiendo a la descomposición de la variación de $Y$ en parte explicada y parte residual, se tiene:

$$S^2_{Y_t}(X_1, X_2) - S^2_{Y_t}(X_1) = S^2_Y - S^2_e(X_1, X_2) - S^2_Y + S^2_e(X_1) \leq S^2_e(X_1)$$

es decir, no podemos ganar en explicación más de lo que quedaba por explicar y, por tanto, $R^2(Y, X_2/X_1) \leq 1$.

Análogamente se define el coeficiente de determinación parcial entre Y y $X_1$ cuando se elimina la influencia de $X_2$ como:

$$R^2(Y, X_1/X_2) = \frac{R^2_{Y,X_1,X_2} - R^2_{Y,X_2}}{1 - R^2_{Y,X_2}}$$

# 8 Números índices y tasas

El seguimiento de una economía se basa habitualmente en índices, que recogen distintos aspectos de la actividad económica: el Índice de Precios de Consumo (IPC), que con periodicidad mensual recoge la evolución de los precios de los principales productos que consumen los hogares españoles; el Índice de Producción Industrial (IPI), referido a la evolución mensual del volumen de producción de los establecimientos industriales o el IBEX-35, que muestra la marcha diaria de la cotización bursátil de las acciones de las mayores empresas españolas. Estos indicadores permiten hacer comparaciones de los valores de una magnitud (precios, producción, cotizaciones bursátiles,...) entre distintos periodos de tiempo y su rasgo distintivo es que efectúan estas comparaciones en términos relativos, mediante cociente, permitiendo hacer afirmaciones del tipo: "en el último mes los precios de consumo aumentaron un 0,2 %" o "en el día de ayer las cotizaciones bursátiles disminuyeron en un 0,5 %".

## 8.1. Índices simples y tasas

**Definición 8.1.** Un *número índice* es una medida estadística de la variación relativa de una magnitud en el tiempo o en el espacio.

Los indicadores que cuantifican las variaciones de una magnitud entre dos periodos de tiempo reciben el nombre de *índices temporales*, mientras que denominaremos *índices espaciales* a aquellos que permiten analizar variaciones entre áreas geográficas.

**Definición 8.2.** Dada una magnitud $X$ y dos periodos temporales, denominados periodo base $(0)$ y periodo actual $(t)$, en los que $X$ toma los valores $x_0$ y $x_t$ (con $x_0 \neq 0$), el *índice simple temporal* $I_{t,0}$ se define como:

$$I_{t,0} = \frac{x_t}{x_0} \tag{8.1.1}$$

El resultado de $I_{t,0}$ permite estudiar la variación relativa de $X$ en el periodo actual con respecto al periodo base:

- $I_{t,0} > 1$ indica que la magnitud ha aumentado en el periodo $t$ con respecto al periodo base 0.
- $I_{t,0} = 1$ indica que la magnitud no ha experimentado cambios entre los periodos 0 y $t$.
- $I_{t,0} < 1$ indica que la magnitud ha disminuido en el periodo $t$ con respecto al periodo base 0.

En este texto se presentarán los resultados de los números índices expresados en tantos por uno, puesto que ello facilita su utilización en aplicaciones posteriores. No obstante, es habitual publicar los resultados de los números índices expresados en porcentaje, resultados que se obtienen sin más que multiplicar por 100 los valores en tantos por uno.

**Definición 8.3.** La *tasa de variación* en términos porcentuales $r_{t,0}$, asociada a la variación de una magnitud $X$ en un periodo $t$ con respecto al periodo de referencia 0, viene dada por el valor de la expresión:

$$r_{t,0} = \left(\frac{x_t - x_0}{x_0}\right) 100 \tag{8.1.2}$$

**Propiedad 8.1.** *La tasa de variación $r_{t,0}$ puede calcularse a partir del índice simple $I_{t,0}$ mediante la expresión: $r_{t,0} = (I_{t,0} - 1)100$.*

**Ejemplo 8.1.** Según las cifras de la Encuesta Trimestral de Coste Laboral elaborada por el INE, las cifras medias del coste laboral ($c_t$) fueron de 2.431,92€ en 2008 y de 2.516,82€ en 2009. Por tanto, el índice simple de 2009 tomando como referencia el año 2008 vendrá dado por $I_{09,08} = \dfrac{c_{09}}{c_{08}} = 1,035$, lo que nos lleva a afirmar que el coste laboral medio por trabajador y mes en España en el año 2009 aumentó un 3,5 % respecto al año 2008.

**Definición 8.4.** Dada una magnitud $X$, una zona geográfica $h$ y una zona de referencia $z$, en las que $X$ toma los valores $x_h$ y $x_z$ (con $x_z \neq 0$), el *índice simple espacial* $I_{h,z}$ se define como:

$$I_{h,z} = \frac{x_h}{x_z} \tag{8.1.3}$$

El índice espacial $I_{h,z}$ permite analizar la situación relativa de la magnitud $X$ en una zona geográfica en relación a una zona de referencia, siendo su interpretación equivalente a la de los índices temporales: resultados unitarios indican valores coincidentes en ambas zonas geográficas y resultados mayores (inferiores) que 1 reflejan que el valor de $X$ en la zona $h$ supera (es inferior) al de la zona de referencia.

**Ejemplo 8.2.** En el contexto del ejemplo 8.1, podemos calcular índices espaciales para estudiar la situación relativa de los costes laborales de las comunidades autónomas en relación a la media nacional. Así, teniendo en cuenta que en el año 2009 el coste laboral medio en el País Vasco fue de 2.893,38€, el índice espacial correspondiente vendría dado por el cociente $\frac{c_{PV}}{c_{Nac}} = 1,15$, lo que supone que el coste laboral medio en el País Vasco fue un 15 % superior a la media nacional. De forma análoga, y teniendo en cuenta que el coste laboral medio en Asturias fue de 2.502,68€, se obtiene un índice espacial 0,99, por lo que en el caso de Asturias el coste laboral se situó un 1 % por debajo de la media nacional.

Un aspecto importante en el cálculo de índices temporales y espaciales es la elección de la referencia. Una alternativa habitual es tomar una referencia fija, siempre y cuando el valor correspondiente a dicho periodo o zona no sea atípico.

El dato tomado como referencia en el cálculo de índices y tasas puede ofrecer resultados engañosos. Es bien conocida la falacia basada en una noticia de prensa según la cual "los homicidios cometidos en una ciudad habían aumentado un 60 % respecto al año anterior". El problema en este caso estaba en que el dato base era pequeño pues, tal y como aclaró posteriormente el periódico, el número de homicidios había pasado en un año de 5 a 8.

También es interesante señalar que, como consecuencia de su carácter multiplicativo, una misma tasa aplicada a niveles diferentes de una magnitud da lugar a variaciones de distinta cuantía. Un caso particular de esta situación es el correspondiente a aplicar una tasa positiva para a continuación aplicar la misma tasa con signo negativo: pensemos, por ejemplo, en un trabajador con un salario de 1.000€ que el último año se incrementó en un 10 %, pasando a ganar 1.100€ y que, para el próximo año disminuirá en un 10 %, lo que supondrá que su nuevo salario pasará a ser de 990€, inferior a los 1.000€ iniciales.

En el caso de datos temporales es habitual calcular tasas con periodo de referencia variable, alternativa de especial interés cuando se dispone de datos con periodicidad inferior a un año (trimestres, meses, etc.). Así, si se dispone de datos trimestrales, siendo $x_t$ el valor de una magnitud $X$ en el trimestre $t$, la aplicación de la expresión 8.1.2 a los datos de dos trimestres consecutivos permite estudiar la variación relativa respecto al trimestre anterior a través de las *tasas intertrimestrales* $r_{t,t-1} = \left(\frac{x_t - x_{t-1}}{x_{t-1}}\right) 100$. Asimismo, es posible estudiar la variación relativa respecto al mismo trimestre del año anterior (en un año) a través de las *tasas interanuales* $r_{t,t-4} = \left(\frac{x_t - x_{t-4}}{x_{t-4}}\right) 100$. De forma análoga pueden calcularse tasas asociadas a series mensuales, en cuyo caso se obtienen *tasas intermensuales*, comparando las cifras del mes $t$ con las del mes inmediatamente anterior $(t-1)$ y tasas interanuales, mediante la comparación de las cifras del mes $t$ con las del mismo mes del año anterior $t-12$.

**Ejemplo 8.3.** Consideremos ahora las cifras trimestrales procedentes de la Encuesta Trimestral de Coste Laboral del INE reflejadas en el cuadro adjunto. Comparando la cifra del primer trimestre de 2009 con la correspondiente al cuarto trimestre de 2008, se obtiene una tasa intertrimestral negativa de -5,5 %, mientras que si se efectúa la comparación respecto al primer trimestre de 2008, se obtiene una tasa interanual positiva del 4,2 %.

| Trimestre | Coste laboral (en €) |
|---|---|
| 2008.I | 2.342,28 |
| 2008.II | 2.451,40 |
| 2008.III | 2.350,17 |
| 2008.IV | 2.583,82 |
| 2009.I | 2.440,54 |

La información que ofrecen las tasas interanuales e intertrimestrales (intermensuales) es complementaria. En general, para conocer la verdadera evolución en el tiempo de una magnitud suelen preferirse las cifras interanuales, que es el criterio adoptado en el cuadro "Datos principales" de la web del INE. Pensemos, por ejemplo, en una serie trimestral de ventas de helados: la tasa intertrimestral del tercer trimestre de cada año ofrecerá siempre tasas positivas como consecuencia del incremento de las

ventas en la temporada estival, mientras que el descenso del consumo tras el verano justifica que los cuartos trimestres ofrezcan siempre tasas intertrimestrales negativas; en cambio, las tasas interanuales no se verán afectadas por la estación del año, dado que comparan cifras de un mismo trimestre de años diferentes.

Los índices y tasas temporales calculados en esta sección cuantifican variaciones "exactas" entre pares de periodos. Así, en el ejemplo 8.3, se ha obtenido que los costes laborales aumentaron un 4,2 % en un año, desde el primer trimestre de 2008 al primer trimestre de 2009. Si ahora nos preguntamos respecto a la tasa promedio de crecimiento trimestral en dicho año, ¿podríamos decir que ha sido la cuarta parte, es decir, del 1,05 % cada trimestre? La respuesta es negativa y ello es debido a que los crecimientos trimestrales son acumulativos, es decir, si el primer trimestre los costes salariales crecen un 1,05 %, la cifra base del coste salarial del trimestre siguiente sobre la que se aplicaría el 1,05 % sería mayor, lo que daría lugar a que la tasa final resultante para 2009.I fuera superior al 4,2 %.

**Propiedad 8.2.** *El índice de crecimiento medio acumulativo entre los periodos 0 y $t$, $I_{t,0}(m)$, se calcula como la media geométrica de los índices entre periodos consecutivos a través de la expresión:*

$$I_{t,0}(m) = \sqrt[t]{\prod_{j=1}^{t} I_{j,j-1}}$$

*Demostración.* El índice de crecimiento medio acumulativo entre los periodos 0 y $t$, $I_{t,0}(m)$, debe verificar: $I_{t,0} = (I_{t,0}(m))^t$. Por otra parte, si $I_{j,j-1}$ es el índice del periodo $j$ con base $j-1$ $(j = 0, \ldots, t)$, aplicando la propiedad de circularidad de los índices simples se tiene: $I_{t,0} = I_{t,t-1} I_{t-1,t-2} \cdots I_{1,0} = \prod_{j=1}^{t} I_{j,j-1}$. Finalmente, igualando ambas expresiones se obtiene:

$$I_{t,0}(m) = \sqrt[t]{I_{t,0}} = \sqrt[t]{\prod_{j=1}^{t} I_{j,j-1}}$$

□

**Propiedad 8.3.** *La tasa media de crecimiento acumulativo entre los periodos 0 y $t$, $r_{t,0}(m)$, se calcula como el valor de la expresión:*

$$r_{t,0}(m) = \left( \sqrt[t]{\prod_{j=1}^{t} \left(1 + \frac{r_{j,j-1}}{100}\right)} - 1 \right) 100$$

*Demostración.* La comprobación es inmediata a partir de la propiedad anterior, sin más que tener en cuenta la relación entre índices y tasas de la propiedad 8.1. □

Aplicando los resultados de la proposición anterior a las cifras del ejemplo 8.3 se obtiene una tasa media de crecimiento trimestral acumulativo del 1,03 %.

## 8.2. Índices sintéticos

El supuesto desarrollado en el apartado anterior no resuelve gran parte de las situaciones que se presentan en los estudios económicos, en los que más que estudiar la evolución de una única magnitud (el precio de un bien, el salario de un sector de actividad, ...) se precisa analizar la evolución de magnitudes complejas (los precios de un conjunto de $n$ bienes, los salarios para un conjunto de $n$ sectores productivos, ...). Esta consideración justifica la necesidad de definir índices sintéticos o complejos, que permitan resumir las variaciones temporales de los distintos componentes de una magnitud compleja en una única cifra para cada periodo.

Sea $X$ una magnitud con $n$ componentes $X_i$ $(i = 1, ..., n)$, cuyos valores en los periodos base y actual se designan respectivamente por $x_{i0}$ y $x_{it}$ $(i = 1, ..., n)$. En estas condiciones, es posible calcular un índice simple asociado a cada componente que permitirá estudiar la variación temporal del i-ésimo componente $(i = 1, ..., n)$ según se resume en el cuadro 8.1.

| Componente | 1 | $\cdots$ | $i$ | $\cdots$ | n |
|---|---|---|---|---|---|
| Valores periodo base 0 | $x_{10}$ | $\cdots$ | $x_{i0}$ | $\cdots$ | $x_{n0}$ |
| Valores periodo actual $t$ | $x_{1t}$ | $\cdots$ | $x_{it}$ | $\cdots$ | $x_{nt}$ |
| Índices simples | $I_{t,0}(1) = \dfrac{x_{1t}}{x_{10}}$ | $\cdots$ | $I_{t,0}(i) = \dfrac{x_{it}}{x_{i0}}$ | $\cdots$ | $I_{t,0}(n) = \dfrac{x_{nt}}{x_{n0}}$ |

Tabla 8.1: Información para el cálculo de índices sintéticos

El método más habitual para sintetizar estos $n$ índices simples consiste en aplicar un promedio, preferentemente la media aritmética debido a sus ventajas operativas, ya analizadas en el tema 2. De hecho, este promedio vendrá justificado teóricamente cuando los $n$ componentes de un índice actúen de forma aditiva, superponiéndose las distintas variaciones independientes de éstos. Sin embargo, hay que tener presente que los distintos componentes pueden no tener la misma importancia y, por consiguiente, sus variaciones no tendrán la misma influencia en la variación global. Ello justifica que en la mayor parte de las situaciones sea necesario asignar ponderaciones, que cuantifiquen la importancia relativa de cada componente, y utilizar como promedio la media ponderada.

**Definición 8.5.** Dada una magnitud $X$ con $n$ componentes, cada uno de los cuales lleva asociado un índice simple $I_{t,0}(i)$ $(i = 1, ..., n)$ y un sistema de ponderaciones $w_i$ $(i = 1, ..., n)$, se define el *índice sintético media ponderada* del periodo $t$ con base en el periodo 0 como la media ponderada de los índices simples calculada a través de la expresión:

$$I_{t,0} = \frac{\sum_{i=1}^{n} I_{t,0}(i) w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} \frac{x_{it}}{x_{i0}} w_i}{\sum_{i=1}^{n} w_i} \tag{8.2.1}$$

**Ejemplo 8.4.** Si se dispone de cifras de salarios por sectores, pueden calcularse índices que muestren la evolución global de los salarios aplicando un índice media ponderada de los índices simples de cada sector. En la tabla adjunta se resumen los resultados correspondientes a la aplicación de la expresión 8.2.1, tomando como base el año 2008 y como ponderaciones los pesos relativos del empleo en cada sector:

$$I_{t,08} = \frac{\sum_{i=1}^{3} I_{t,08}(i)w_i}{\sum_{i=1}^{3} w_i} = \frac{I_{t,08}(Ind)w_{IND} + I_{t,08}(C)w_C + I_{t,08}(S)w_S}{w_{Ind} + w_C + w_S}$$

| | Salarios (€) | | | |
|---|---|---|---|---|
| Año | Industria (Ind) | Construcción (C) | Servicios (S) | |
| 2008 | 1.800 | 1.550 | 1.630 | |
| 2009 | 1.850 | 1.580 | 1.650 | |
| 2010 | 1.900 | 1.600 | 1.700 | |
| Empleo ($w_i$) | 20 % | 10 % | 70 % | |
| | Índices simples salarios | | | Índ.media ponderada |
| Año | $I_{t,08}(Ind)$ | $I_{t,08}(C)$ | $I_{t,08}(S)$ | $I_{t,08}$ |
| 2008 | 1 | 1 | 1 | 1 |
| 2009 | 1,028 | 1,019 | 1,012 | 1,020 |
| 2010 | 1,056 | 1,032 | 1,043 | 1,044 |

Así, por ejemplo, en el año 2010, se observa que los salarios de todos los sectores no agrarios han aumentado respecto a 2008: un 5,6 % en el caso de la industria, un 3,2 % en construcción y un 4,3 % en servicios. Tras calcular el índice media ponderada de los índices simples, se obtiene una cifra promedio de los aumentos salariales en el conjunto de los sectores: la subida global de los salarios no agrarios en 2010 respecto a 2008 fue del 4,4 %.

A la vista de la ilustración surgen al menos dos interrogantes vinculados a la elección de las ponderaciones en un índice media ponderada. En primer lugar, la variable adecuada para su determinación, que dependerá del tipo de magnitud analizada. En el estudio de salarios por sectores, una opción adecuada es el peso relativo del empleo de cada sector; en los temas siguientes se presentarán las ponderaciones utilizadas habitualmente en los estudios de precios y cantidades, que se corresponden con el valor de cada producto.

El segundo interrogante está relacionado con el periodo de referencia de las ponderaciones: ¿son constantes en todo el ámbito temporal del estudio o varían a lo largo del tiempo? En función de la respuesta a esta pregunta se distingue entre *índices de base fija*, que son aquéllos para los que la estructura de ponderaciones no depende del periodo temporal considerado y se mantiene constante a lo largo del tiempo ($w_i = w_{i0}$), e *índices de base móvil*, para los que dicha estructura de ponderaciones va cambiando

a lo largo del periodo temporal considerado ($w_i = w_{it}$). La fórmula utilizada en el ejemplo es de base fija, puesto que se mantiene la misma distribución de empleo por sectores para los tres años; para construir un índice de base móvil se precisaría disponer de información sobre la distribución sectorial del empleo para cada uno de los años incluidos en el estudio. La opción de utilizar índices de base fija tiene la ventaja de ser menos exigente en cuanto a las necesidades de información para determinar las ponderaciones, si bien este aspecto conlleva el inconveniente de que el sistema de ponderaciones puede quedar obsoleto si se consideran periodos temporales muy amplios. La alternativa será utilizar índices de base móvil, opción que resulta ventajosa por su flexibilidad y capacidad de adaptación a la situación de cada periodo, aunque es de menor uso en la práctica como consecuencia del mayor volumen de información estadística requerida para su cálculo.

Si bien los índices sintéticos tipo media ponderada son los más utilizados, cabe la posibilidad de definir otras fórmulas, tales como los índices agregativos, que presentan un buen comportamiento en la práctica.

**Definición 8.6.** Dada una magnitud $X$ con $n$ componentes, que toma valores $x_{i0}$ y $x_{it}$ $(i = 1, ..., n)$ en los periodos base y actual, respectivamente, con un sistema de ponderaciones asociado $w_i$ $(i = 1, ..., n)$, se define el *índice sintético agregativo* del periodo $t$ con base en el periodo 0 como el cociente entre los valores agregados ponderados correspondientes a ambos periodos, calculado a través de la expresión:

$$I_{t,0} = \frac{\sum_{i=1}^{n} x_{it} w_i}{\sum_{i=1}^{n} x_{i0} w_i} \tag{8.2.2}$$

## 8.3. Propiedades de los índices

Con el fin de establecer un marco de referencia para elegir entre distintas alternativas de fórmulas de cálculo de índices sintéticos, se proponen a continuación un conjunto de cinco propiedades deseables, basadas en la propia interpretación de los números índices: identidad, inversión o reversión temporal, circularidad, proporcionalidad e independencia de la escala de medida. Además de la definición de las cinco propiedades básicas, se añade la deducción de una sexta propiedad, la de cambio de base, vinculada directamente a las propiedades de inversión y circularidad.

> La definición de las propiedades deseables de los números índices tiene como punto de partida los trabajos del economista estadounidense Irving Fisher (1867-1947). Fisher fue pionero en los estudios sobre números índices, tanto en los aspectos teóricos como empíricos. En su obra de 1922 "The Making of Index Numbers. A study of their varieties, tests and reliability"[4], Fisher estableció el marco teórico para la selección de fórmulas de cálculo de números índices y propuso un indicador de precios, denominado Índice "ideal" de Fisher, que será introducido en el tema siguiente. En lo que se refiere a los aspectos empíricos, Fisher creó en 1923 el Instituto de Números Índices, que se

convirtió en el primer organismo en publicar de forma regular datos económicos en forma de números índices.

**Definición 8.7.** *Identidad*: Si coinciden los periodos base y actual, el valor del índice debe ser la unidad (o 100 %): $I_{0,0} = I_{t,t} = 1$.

La lógica de esta propiedad se desprende del propio objetivo de los índices ya que se trata de medir variaciones entre dos periodos de tiempo y, por consiguiente, si los periodos son coincidentes el índice no debe reflejar ninguna variación.

**Definición 8.8.** *Inversión (reversión temporal)*: Si se permutan los periodos actual y base de un índice, el resultado debe ser el inverso del valor inicial: $I_{0,t} = \frac{1}{I_{t,0}}$.

Esta propiedad se basa en la idea de que la variación de una magnitud entre dos periodos debe ser única, independientemente de la óptica adoptada. Consideremos, por ejemplo, un producto que valía 10€ en el periodo 0 y 20€ en el periodo 1. El índice del periodo 1 respecto al periodo base 0 indicará que el precio se ha duplicado ($I_{1,0} = 2$) y el índice del periodo 0 tomando como base el periodo 1 debe indicar que el precio es la mitad ($I_{0,1} = 1/2$).

**Definición 8.9.** *Circularidad:* Dado un periodo de tiempo $t'$ ($0 < t' < t$), el índice entre los periodos 0 y $t$ debe coincidir con el producto de los índices calculados a través del periodo intermedio $t$', es decir, $I_{t,0} = I_{t,t'} I_{t',0}$.

Esta propiedad se basa en la misma idea que la propiedad de inversión, en el sentido de que la variación de una magnitud entre dos periodos debe ser única, en este caso ante la consideración de periodos intermedios. En el ejemplo anterior, si el producto también duplica su precio en el periodo 2, pasando a valer 40€, el índice del periodo 2 tomando como base el periodo 0 indicará que el precio se ha cuadruplicado, que será el resultado de multiplicar el índice del periodo 1 base 0 (el precio se duplica entre esos dos años) por el índice del periodo 2 con base 1 (periodos entre los que también se duplica el precio del producto).

**Definición 8.10.** *Proporcionalidad*: Si en el periodo actual todos los componentes del índice varían en una proporción, el índice debe experimentar la misma variación proporcional. Es decir, si $x'_t = kx_t$ ($k \epsilon \mathbb{R}$), entonces $I'_{t,0} = kI_{t,0}$.

Dado que los índices cuantifican variaciones relativas, cabe esperar que reflejen cualquier variación proporcional de la magnitud analizada en idéntica proporción. Supongamos, por ejemplo, que los salarios de todos los sectores se incrementan en un 3 % respecto a la situación del periodo actual, $x'_{i,t} = 1,03 x_{i,t} \, \forall i$, entonces el nuevo índice global de salarios deberá ser también un 3 % superior al valor del índice antes de la subida.

**Definición 8.11.** *Independencia de la escala de medida (homogeneidad)*: El valor de un índice es invariante ante cambios en las unidades de medida, es decir, si $Y = kX$ ($k \epsilon \mathbb{R}$), entonces $I^Y_{t,0} = I^X_{t,0}$.

Por tratarse de medidas relativas, los números índices son adimensionales y por lo tanto no deben verse afectados por cambios en las unidades de medida (cambios de escala). Como consecuencia de esta propiedad el resultado de un índice que muestre la evolución de la cantidad producida de acero será el mismo si la producción se expresa en kilogramos o en toneladas. Análogamente, el resultado de un índice de salarios sería el mismo si éstos se expresaran en euros o en miles de euros, pero podría verse modificado si se expresaran en dólares. Los cambios de divisa son habituales en economía, pero no se trata de simples cambios de escala, puesto que no son constantes a lo largo del tiempo, por lo que no sería aplicable la propiedad de independencia de la escala de medida.

En ocasiones interesa cambiar la óptica de referencia del estudio y se precisa transformar la serie de índices disponible de modo que las variaciones se midan respecto a un periodo base diferente. La operación requerida en este caso se denomina *cambio de base*.

**Propiedad 8.4.** *Dada una serie de índices $I_{t,0}$, calculada a partir de una fórmula $I$ que verifica las propiedades de inversión y circularidad, la serie referida a un nuevo periodo base $b$ puede calcularse a partir de la serie inicial, con base en el periodo 0, a través de la expresión: $I_{t,b} = \frac{I_{t,0}}{I_{b,0}}$.*

*Demostración.* Se efectuará la demostración en dos casos, según el intervalo temporal al que pertenezca la nueva base b:

Caso 1.- Supongamos que $0 < b < t$. Entonces, aplicando la propiedad de circularidad se tiene: $I_{t,0} = I_{t,b}I_{b,0}$, de donde despejando el valor de $I_{t,b}$, se obtiene la expresión propuesta en el enunciado.

Caso 2.- Supongamos que $b > t$. Entonces por la propiedad de circularidad se tiene: $I_{b,0} = I_{b,t}I_{t,0}$, expresión en la que es posible despejar el valor de $I_{b,t} = \frac{I_{b,0}}{I_{t,0}}$. Teniendo en cuenta que, además, por la propiedad de inversión $I_{b,t} = \frac{1}{I_{t,b}}$ e igualando ambas expresiones se obtiene la expresión de cálculo buscada para $I_{t,b}$. □

Una vez establecidas las propiedades deseables, la pregunta relevante se referirá a su verificación por parte de las fórmulas introducidas en este tema. La respuesta es afirmativa para los índices simples, según se demuestra en la proposición siguiente. Sin embargo, la respuesta ya no es tan satisfactoria para otras fórmulas; así, por ejemplo, los índices media ponderada no cumplen las propiedades de inversión y circular ni, por consiguiente, la de cambio de base. A esta cuestión nos referiremos de manera más extensa en el tema siguiente, una vez se hayan introducido las fórmulas de cálculo habituales para precios y cantidades.

**Propiedad 8.5.** *Los índices simples cumplen las propiedades de identidad, inversión, circularidad, proporcionalidad e independencia de la escala de medida.*

*Demostración.* Sea $I_{t,0} = \frac{x_t}{x_0}$ el índice simple del periodo $t$ con base en el periodo 0. A continuación se comprobará que está fórmula verifica las propiedades del enunciado.

- Identidad. Si los periodos base y actual coinciden $t \equiv 0$, entonces: $I_{0,0} = \frac{x_0}{x_0} = 1$, por lo que queda comprobado que el índice resultante es unitario.
- Inversión. Si se intercambian los periodos base y actual, entonces: $I_{0,t} = \frac{x_0}{x_t}$. Para comprobar que se verifica la propiedad de inversión y, como consecuencia, el nuevo índice es el inverso del índice inicial $I_{t,0}$, basta dividir numerador y denominador por el valor en el periodo base $x_0$: $I_{0,t} = \frac{x_0}{x_t} = \frac{\frac{x_0}{x_0}}{\frac{x_t}{x_0}} = \frac{1}{I_{t,0}}$.
- Circular. Si se considera $x_{t'}$, valor de la magnitud $X$ en un periodo $t'$ $(0 < t' < t)$, pueden calcularse los índices simples $I_{t,t'} = \frac{x_t}{x_{t'}}$ e $I_{t',0} = \frac{x_{t'}}{x_0}$. Entonces, partiendo de la fórmula del índice simple $I_{t,0}$ y sin más que multiplicar numerador y denominador por $x_{t'}$: $I_{t,0} = \frac{x_t}{x_0} = \frac{x_t}{x_0}\frac{x_{t'}}{x_{t'}} = \frac{x_t}{x_{t'}}\frac{x_{t'}}{x_0} = I_{t,t'}I_{t',0}$, por lo que queda comprobado que el índice simple del periodo $t$ con base 0 puede obtenerse como producto de los índices calculados a través del periodo intermedio $t'$.
- Proporcionalidad. Si en el periodo actual la magnitud $X$ experimenta una variación proporcional, con factor de proporcionalidad $k$, es decir, $x'_t = kx_t$, el nuevo índice simple $I'_{t,0}$ variará en la misma proporción ya que: $I'_{t,0} = \frac{x'_t}{x_0} = \frac{kx_t}{x_0} = kI_{t,0}$
- Independencia de la escala de medida. Si la magnitud $X$ experimenta un cambio de unidades de medida que no depende del periodo temporal considerado, $Y = kX$, entonces el índice calculado en la nueva escala coincide con el inicial ya que: $I^Y_{t,0} = \frac{y_t}{y_0} = \frac{kx_t}{kx_0} = I^X_{t,0}$. Sin embargo, en casos como los cambios de divisa, que tienen la particularidad de que son variables a lo largo del tiempo, es decir, en cada periodo $t$, $y_t = k_t x_t$, se tiene que: $I^Y_{t,0} = \frac{y_t}{y_0} = \frac{k_t x_t}{k_0 x_0} = I^{TC}_{t,0} I^X_{t,0}$, donde $I^{TC}_{t,0}$ es un índice del tipo de cambio y no será posible garantizar entonces el cumplimiento de la propiedad de independencia de la escala de medida.

□

El planteamiento e interpretación de las propiedades de los números índices se ha efectuado desde la óptica temporal, pero sería extensible al caso espacial. Así, por ejemplo, el cumplimiento del requisito de inversión por parte del índice simple $I_{hz}$ supone que la posición relativa de dos áreas es única llegándose a resultados inversos según cuál de ellas sea adoptada como referencia; y la verificación de la propiedad de circularidad garantiza que la comparación de dos regiones conduce a idéntica conclusión tanto si se efectúa de modo directo como si se adopta un tercer territorio como referencia intermedia.

# 9 Números índices: Fórmulas habituales, variación y repercusión

## 9.1. Fórmulas habituales de precios y cantidades

Entre los principales indicadores de la economía española destacados en la web del INE figuran dos índices de precios, el Índice de Precios de Consumo (IPC) y el Índice de Precios Industriales (IPRI), y un índice cuántico, el Índice de Producción Industrial (IPI). Ello pone de manifiesto el papel relevante del estudio de las variaciones de precios y producción en una economía y justifica el interés de dedicar un apartado específico a las fórmulas más utilizadas en este contexto.

Consideremos un conjunto de $n$ bienes para los que se dispone de información sobre sus precios ($p$) y cantidades ($q$) en los periodos base y actual, y para los que se han calculado los correspondientes índices simples de precios y cantidades, aplicando la fórmula introducida en el capítulo anterior. La tabla siguiente resume la notación utilizada para el bien i-ésimo ($i = 1, \ldots, n$).

| | PRECIOS | CANTIDADES |
|---|---|---|
| Periodo base 0 | $p_{i0}$ | $q_{i0}$ |
| Periodo actual $t$ | $p_{it}$ | $q_{it}$ |
| Índices simples | $I^P_{t,0}(i) = \dfrac{p_{it}}{p_{i0}}$ | $I^Q_{t,0}(i) = \dfrac{q_{it}}{q_{i0}}$ |

El método habitual para calcular índices que muestren la evolución global de precios o cantidades consiste en considerar índices sintéticos tipo media ponderada, tanto de base fija como de base móvil. En cuanto al sistema de ponderaciones, suele considerarse el valor de cada bien, calculado como producto de precio por cantidad $w_i = p_i q_i$ $(i = 1, ..., n)$, como variable más adecuada para reflejar la importancia de la variación en el precio o la cantidad de los distintos bienes. Respecto al periodo temporal de referencia de las ponderaciones, las opciones más utilizadas son las correspondientes a las fórmulas de Laspeyres y Paasche: la primera es de base fija y utiliza como ponderaciones los valores del periodo base, mientras que la segunda es de base móvil y utiliza ponderaciones vinculadas al periodo actual.

Las fórmulas de Laspeyres y Paasche deben su nombre a los economistas alemanes Etienne Laspeyres (1834-1913) y Hermann Paasche (1851-1925).

## Fórmulas habituales de precios

**Definición 9.1.** Dado un conjunto de $n$ bienes con precios y cantidades en los periodos 0 y $t$ $\{p_{i0}, p_{it}, q_{i0};\ i = 1, \ldots, n\}$, se define el *índice de precios de Laspeyres* $(L^P)$ del periodo $t$ con base el periodo 0 como un índice media ponderada de los índices simples de precios $I^P_{t,0}(i)$ con ponderaciones $w_{i0} = p_{i0}q_{i0}$.

La expresión habitual de cálculo de un índice de Laspeyres de precios viene dada por:

$$L^P_{t,0} = \frac{\sum_{i=1}^{n} p_{it}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}} \tag{9.1.1}$$

que se obtiene de forma directa a partir de la expresión general de los índices sintéticos media ponderada 8.2.1, sin más que considerar los índices simples de precios y como ponderaciones, los valores del periodo base:

$$L^P_{t,0} = \frac{\sum_{i=1}^{n} I^P_{t,0}(i)w_{i0}}{\sum_{i=1}^{n} w_{i0}} = \frac{\sum_{i=1}^{n} \frac{p_{it}}{p_{i0}}p_{i0}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}} = \frac{\sum_{i=1}^{n} p_{it}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}}$$

**Definición 9.2.** Dado un conjunto de $n$ bienes con precios y cantidades en los periodos 0 y $t$ $\{p_{i0}, p_{it}, q_{it};\ i = 1, \ldots, n\}$, se define el *índice de precios de Paasche* $(P^P)$ del periodo $t$ con base el periodo 0 como un índice media ponderada de los índices simples de precios $I^P_{t,0}(i)$ con ponderaciones $w_{it} = p_{i0}q_{it}$.

La expresión habitual de cálculo de un índice de Paasche de precios viene dada por:

$$P^P_{t,0} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}} \tag{9.1.2}$$

que se obtiene también a partir de la expresión general de los índices sintéticos media ponderada tomando como ponderaciones las cantidades consumidas en el periodo actual valoradas a precios del periodo base:

$$P^P_{t,0} = \frac{\sum_{i=1}^{n} I^P_{t,0}(i)w_{it}}{\sum_{i=1}^{n} w_{it}} = \frac{\sum_{i=1}^{n} \frac{p_{it}}{p_{i0}}p_{i0}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}}$$

Analizando las fórmulas resultantes para los índices de precios de Laspeyres y Paasche se observa que ambas se expresan como índices agregativos que cuantifican exclusivamente variaciones en los precios de los bienes incluidos en el índice: en el índice

de Laspeyres se comparan valores agregados de los periodos actual y base, para una distribución constante en el tiempo de las cantidades $\{q_{i0}\}$, mientras que la comparación en el caso de la fórmula de Paasche se efectúa para unas cantidades variables a lo largo del tiempo $\{q_{it}\}$.

**Ejemplo 9.1.** Para estudiar la evolución global de los precios de los carburantes en una estación de servicio en la que se comercializan dos tipos de combustible: sin plomo 95 y gasóleo, pueden utilizarse las fórmulas de Laspeyres y Paasche. A partir de las cifras de precios (en céntimos de €) y cantidades (en millones de litros) de la tabla adjunta, se observa que los índices de Laspeyres y Paasche ofrecen resultados prácticamente coincidentes.

| Año | Sin plomo 95 | | Gasóleo | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $p$ | $q$ | $p$ | $q$ | $\sum_{i=1}^{2} p_{it}q_{i07}$ | $\sum_{i=1}^{2} p_{it}q_{it}$ | $\sum_{i=1}^{2} p_{i07}q_{it}$ | $L^P_{t,07}$ | $P^P_{t,07}$ |
| 2007 | 109 | 0,5 | 96 | 1,5 | 198,5 | 198,5 | 198,5 | 1 | 1 |
| 2008 | 124 | 0,48 | 129 | 1,45 | 255,5 | 246,57 | 191,52 | 1,287 | 1,287 |
| 2009 | 105 | 0,46 | 92 | 1,43 | 190,5 | 179,86 | 187,42 | 0,960 | 0,910 |
| 2010 | 119 | 0,5 | 110 | 1,42 | 224,5 | 215,7 | 190,82 | 1,131 | 1,130 |

**Propiedad 9.1.** *El índice de precios de Laspeyres verifica las propiedades de identidad y proporcionalidad y no cumple las propiedades de inversión, circularidad y homogeneidad.*

*Demostración.* La comprobación de la propiedad de identidad es inmediata sin más que tener en cuenta que si el periodo base y actual coinciden $p_{it} = p_{i0}$ y, por consiguiente, el valor del índice calculado a través de la expresión 9.1.1 es unitario.

En lo que respecta a la propiedad de inversión, si se invierten los periodos base y actual en el índice de Laspeyres de precios se obtiene el inverso del índice de Paasche de precios ya que:

$$L^P_{0,t} = \frac{\sum_{i=1}^{n} p_{i0}q_{it}}{\sum_{i=1}^{n} p_{it}q_{it}} = \frac{1}{\frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}}} = \frac{1}{P^P_{t,0}}$$

Este resultado permite comprobar que la fórmula de Laspeyres no cumple la propiedad de inversión (y tampoco la de Paasche) y, como consecuencia, tampoco verifica la circularidad ni la propiedad de cambio de base.

Para comprobar la propiedad de proporcionalidad, supondremos que los precios de los $n$ bienes experimentan una variación proporcional en el periodo actual, $p'_{it} = kp_{it}$, en cuyo caso el nuevo índice de precios de Laspeyres vendrá dado por:

$$L_{t,0}^{P'} = \frac{\sum_{i=1}^{n} p'_{it} q_{i0}}{\sum_{i=1}^{n} p_{i0} q_{i0}} = \frac{\sum_{i=1}^{n} k p_{it} q_{i0}}{\sum_{i=1}^{n} p_{i0} q_{i0}} = k L_{t,0}^{P}$$

y, por tanto, el índice de Laspeyres varía en la misma proporción.

Por último, no puede asegurarse que el índice de precios de Laspeyres verifique la propiedad de independencia de la escala de medida. Como ilustración, bastaría tener en cuenta las consideraciones ya efectuadas en el tema anterior respecto a la variabilidad temporal de los cambios de divisa en los precios. □

Respecto al índice de precios de Paasche podría establecerse un resultado análogo al de la propiedad 9.1, con una consideración adicional para el caso de la propiedad de proporcionalidad. En efecto, por tratarse de un índice de base móvil sus ponderaciones dependen de las cantidades del periodo actual $\{q_{it}\}$ y cabe esperar que el consumidor reaccione ante el encarecimiento de un bien sustituyendo su consumo, en la medida de lo posible, por el de otros bienes sustitutivos, lo que conllevaría un cambio en las ponderaciones. Por tanto, desde una óptica económica el índice de precios de Paasche no cumpliría el requisito de proporcionalidad.

**Propiedad 9.2.** *Generalmente, el índice de precios de Laspeyres toma valores superiores al índice de Paasche:* $L_{t,0}^{P} \geqq P_{t,0}^{P}$.

La relación entre los índices de precios de Laspeyres y Paasche se basa en la correlación entre los índices simples de precios y de cantidades: cuando ésta es negativa, el índice de Laspeyres supera al de Paasche y si es positiva el sentido de la desigualdad se invierte. En la práctica, la situación más frecuente es la primera ya que ante aumentos de los precios, los consumidores, como consecuencia del efecto sustitución, tienden a consumir otros productos de menor precio. Por este motivo, no puede asegurarse que el sentido de la desigualdad en la relación entre ambos índices se cumpla siempre, pero sí que es el más habitual.

La demostración de esta propiedad que relaciona los índices de Laspeyres y Paasche se lleva a cabo mediante la fórmula de Bortkiewicz (1923), en la que aparece la covarianza entre precios y cantidades, que habitualmente presenta signo negativo[1].

Ladislaus von Borkiewicz (1868-1932) fue un estadístico y economista ruso que realizó importantes contribuciones al análisis de números índices económicos y también propuso la denominada distribución de Poisson asociada a sucesos de baja frecuencia o "raros".

Teniendo en cuenta que las fórmulas de Laspeyres y Paasche no verifican algunos de los requisitos teóricos deseables de los números índices y que la primera generalmente sobreestima las verdaderas variaciones de los precios, mientras que la segunda las subestima, Fisher propuso el denominado "índice ideal" como promedio de ambos indicadores.

[1] La demostración de la propiedad que relaciona los índices de Laspeyres y Paasche excede los objetivos de este texto y puede consultarse en Calot [2]

**Definición 9.3.** Se define el *índice de precios de Fisher* ($F^P$) como la media geométrica de los índices de precios de Laspeyres y Paasche, es decir,

$$F^P_{t,0} = \sqrt{L^P_{t,0} P^P_{t,0}} \tag{9.1.3}$$

**Propiedad 9.3.** *El índice de precios de Fisher cumple las propiedades de identidad, inversión y circularidad.*

*Demostración.* Teniendo en cuenta la demostración de la propiedad 9.1, se comprueba que $F$ verifica la propiedad de inversión, ya que:

$$F^P_{0,t} = \sqrt{L^P_{0,t} P^P_{0,t}} = \sqrt{\frac{1}{P^P_{t,0}} \frac{1}{L^P_{t,0}}} = \frac{1}{\sqrt{L^P_{t,0} P^P_{t,0}}} = \frac{1}{F^P_{t,0}}$$

Asimismo, considerando un periodo intermedio $t'$ $(0 < t' < t)$ y aplicando las expresiones correspondientes a las fórmulas de Laspeyres y Paasche se comprueba el cumplimiento de la propiedad de circularidad $F^P_{t,t'} F^P_{t',0} = F^P_{t,0}$. □

Las fórmulas introducidas en este capítulo se encuentran entre las más utilizadas para cuantificar las variaciones de precios y se plantea ahora la pregunta relativa a cuál de ellas es la más adecuada. Aun cuando el índice de Fisher goza de ciertas propiedades teóricas interesantes, su utilización en la práctica resulta compleja, en parte relacionado con el hecho de que para su cálculo se precisa información sobre un doble sistema de ponderación, correspondiente a los índices de Laspeyres y de Paasche. La elección entre los índices de Laspeyres y Paasche se basa en la disponibilidad de información para la actualización de las ponderaciones. Pensemos, por ejemplo, en un Índice de Precios de Consumo: el cálculo de un índice de Paasche requeriría conocer las cantidades consumidas de los distintos bienes que integran el índice en cada periodo, lo cual no suele resultar factible por la gran cantidad de información estadística que es necesario recabar. La alternativa será utilizar la fórmula de Laspeyres, que asume una estructura de consumo fija, y renovar la base del índice con cierta frecuencia con el fin de que la estructura de ponderaciones pueda ir recogiendo los cambios en las pautas de consumo de los hogares.

En la actualidad, la fórmula de cálculo del Índice de Precios de Consumo español se basa en un índice de Laspeyres con base en el año 2006, en el que las ponderaciones de los artículos incluidos en el índice se determinan según el gasto de los hogares en el año base. Asimismo, el Índice de Precios Industriales se calcula como un índice de Laspeyres base 2005, en el que las ponderaciones de los productos se determinan según el valor de producción en el año base.

### Fórmulas habituales de cantidades

Se introducen a continuación las fórmulas habituales para medir la variación global de las cantidades (en unidades físicas) producidas o consumidas de un conjunto de $n$ bienes.

**Definición 9.4.** Dado un conjunto de $n$ bienes con precios y cantidades en los periodos 0 y $t$ $\{p_{i0}, q_{i0}, q_{it};\, i = 1, \ldots, n\}$, se define el *índice cuántico de Laspeyres* ($L^Q$) del periodo $t$ con base el periodo 0 como un índice media ponderada de los índices simples de cantidades $I^Q_{t,0}(i)$ con ponderaciones $w_{i0} = p_{i0}q_{i0}$.

La expresión habitual de cálculo de un índice cuántico de Laspeyres viene dada por:

$$L^Q_{t,0} = \frac{\sum_{i=1}^{n} p_{i0}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{i0}} \tag{9.1.4}$$

que se obtiene de forma directa como media ponderada de los índices simples de cantidades tomando como ponderaciones los valores del periodo base:

$$L^Q_{t,0} = \frac{\sum_{i=1}^{n} I^Q_{t,0}(i)w_{i0}}{\sum_{i=1}^{n} w_{i0}} = \frac{\sum_{i=1}^{n} \frac{q_{it}}{q_{i0}} p_{i0}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}} = \frac{\sum_{i=1}^{n} p_{i0}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{i0}}$$

**Definición 9.5.** Dado un conjunto de $n$ bienes con precios y cantidades en los periodos 0 y $t$ $\{p_{it}, q_{i0}, q_{it};\, i = 1, \ldots, n\}$, se define el *índice cuántico de Paasche* ($P^Q$) del periodo $t$ con base el periodo 0 como un índice media ponderada de los índices simples de cantidades $I^Q_{t,0}(i)$ con ponderaciones $w_{it} = p_{it}q_{i0}$.

La expresión habitual de cálculo de un índice cuántico de Paasche viene dada por:

$$P^Q_{t,0} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{it}q_{i0}} \tag{9.1.5}$$

que se obtiene también a partir de la expresión general de los índices sintéticos media ponderada tomando como ponderaciones las cantidades consumidas en el periodo base valoradas a precios del periodo actual:

$$P^Q_{t,0} = \frac{\sum_{i=1}^{n} I^Q_{t,0}(i)w_{it}}{\sum_{i=1}^{n} w_{it}} = \frac{\sum_{i=1}^{n} \frac{q_{it}}{q_{i0}} p_{it}q_{i0}}{\sum_{i=1}^{n} p_{it}q_{i0}} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{it}q_{i0}}$$

De forma análoga a los índices de precios, los índices cuánticos de Laspeyres y Paasche se expresan como índices agregativos que cuantifican exclusivamente variaciones en las cantidades de los bienes incluidos en el índice: en el índice de Laspeyres se comparan valores agregados de los periodos actual y base, para un sistema de precios constante en el tiempo $\{p_{i0}\}$, mientras que la comparación en el caso de la fórmula de Paasche se efectúa para unos precios variables a lo largo del tiempo $\{p_{it}\}$.

**Propiedad 9.4.** *El índice cuántico de Laspeyres verifica las propiedades de identidad, proporcionalidad y homogeneidad y no cumple las propiedades de inversión y circularidad.*

*Demostración.* La demostración es análoga a la desarrollada en la propiedad 9.1 para la fórmula de precios. La única salvedad se refiere al cumplimiento en este caso de la propiedad de homogeneidad, ya que al tratarse de un índice cuántico no se presentan problemas con los cambios de escala que varían en el tiempo. □

También es posible definir un *índice cuántico de Fisher* como media geométrica de los índices cuánticos de Laspeyres y Paasche. Asimismo el tipo de consideraciones sobre la adecuación teórica y los problemas de aplicación práctica de los distintos indicadores de precios se mantienen para el caso de los índices cuánticos, siendo el de Laspeyres el más utilizado. En particular, el Índice de Producción Industrial elaborado por el INE es un índice cuántico de Laspeyres con base en el año 2005 y con ponderaciones de los productos según el valor de producción en el año base.

## 9.2. Índices de valor

Las fórmulas de índices de precios introducidas en el apartado anterior permiten estudiar variaciones exclusivamente en precios, mientras que las fórmulas de índices cuánticos permiten analizar las variaciones de cantidades. Sin embargo, en muchos casos interesa estudiar la variación de los valores $v_i = p_i q_i$, (i=1,...,n), de un conjunto de bienes.

**Definición 9.6.** Dado un conjunto de $n$ bienes con precios y cantidades en los periodos 0 y $t$ $\{p_{i0}, p_{it}, q_{i0}, q_{it};\ i = 1, \ldots, n\}$, se define el *índice de valor* del periodo $t$ con base el periodo 0 como el índice agregativo:

$$IV_{t,0} = \frac{V_t}{V_0} = \frac{\sum_{i=1}^{n} p_{it} q_{it}}{\sum_{i=1}^{n} p_{i0} q_{i0}} \tag{9.2.1}$$

A nivel individual se verifica que la variación en el valor de un bien puede obtenerse como producto de la variación en precio por la variación en cantidad. Por ejemplo, si este año se triplica el precio de un bien y se duplica la cantidad consumida, su valor se verá multiplicado por seis. La extensión de esta idea a un conjunto de $n$ bienes da lugar a la siguiente definición.

**Definición 9.7.** Dada una fórmula de cálculo (I) de índices de precios y cantidades, se dice que cumple el *criterio de reversión de factores* o *compatibilidad* si verifica: $IV = I^P I^Q$.

**Propiedad 9.5.** *Las fórmulas de Laspeyres y Paasche verifican el criterio de reversión de factores de forma cruzada, es decir:* $IV_{t,0} = L^P_{t,0} P^Q_{t,0} = L^Q_{t,0} P^P_{t,0}$.

*Demostración.* Teniendo en cuenta las fórmulas de cálculo de Laspeyres y Paasche podemos comprobar, por ejemplo, la primera parte de la identidad anterior:

$$L_{t,0}^{P} P_{t,0}^{Q} = \frac{\sum_{i=1}^{n} p_{it} q_{i0}}{\sum_{i=1}^{n} p_{i0} q_{i0}} \frac{\sum_{i=1}^{n} p_{it} q_{it}}{\sum_{i=1}^{n} p_{it} q_{i0}} = IV_{t,0}$$

Por consiguiente las fórmulas de Laspeyres y Paasche no cumplen la propiedad de compatibilidad en un sentido estricto, sino de forma cruzada. □

## 9.3. Deflactación

Las magnitudes económicas valoradas en unidades monetarias de un periodo de tiempo $t$ se dice que están expresadas en términos monetarios nominales o a precios corrientes del periodo $t$. La comparación de valores de una magnitud a precios corrientes de diferentes periodos de tiempo da como resultado una variación aparente. Dado que los precios varían con el paso del tiempo, una solución para comparar valores correspondientes a diferentes periodos sin que se vean afectados por los cambios en los precios consiste en expresar dichos valores a precios constantes de un periodo de referencia 0. La comparación de valores de una magnitud a precios constantes del periodo 0 da como resultado una variación real.

Si se considera un valor agregado del tipo $V_t = \sum_{i=1}^{n} p_{it} q_{it}$, podría expresarse a precios constantes de un periodo 0 de forma directa como $\sum_{i=1}^{n} p_{i0} q_{it}$. Pero no todas las magnitudes monetarias son susceptibles de ser expresadas como agregados de precios por cantidades o, en ocasiones, no se dispone de información de los precios del periodo base, por lo que es preciso establecer un procedimiento general que permita transformar series a precios corrientes en series a precios constantes, descontando el efecto de la variación de los precios entre los dos periodos considerados.

**Definición 9.8.** La *deflactación* es la operación que permite transformar un valor expresado a precios corrientes de un periodo $t$ ($x_t$) en otro a precios constantes de un periodo de referencia 0, a través de la expresión:

$$\frac{x_t}{I_{t,0}^{P}}$$

donde $I_{t,0}^{P}$ es un índice de precios que refleja la variación de los precios entre los periodos 0 y $t$ y que recibe el nombre de *deflactor*.

La elección del deflactor dependerá del problema al que nos enfrentemos en cada caso concreto. No obstante, como criterio general interesará utilizar como deflactor un índice de precios referido al periodo considerado y que incluya un conjunto de

bienes y/o servicios lo más parecidos posible a los que integran la magnitud que se desea deflactar. En el caso de valores agregados el deflactor adecuado sería el índice de precios de Paasche con la misma cobertura de bienes, ya que:

$$\frac{V_t}{P^P_{t,0}} = \frac{\sum_{i=1}^{n} p_{it}q_{it}}{\frac{\sum_{i=1}^{n} p_{it}q_{it}}{\sum_{i=1}^{n} p_{i0}q_{it}}} = \sum_{i=1}^{n} p_{i0}q_{it}$$

En la práctica es habitual acudir a índices de precios publicados por organismos oficiales. Así, para transformar la renta disponible de una familia a unidades monetarias constantes de un determinado año, el deflactor adecuado será el IPC, mientras que para deflactar el valor de un conjunto de productos industriales resulta más adecuado utilizar como deflactor el Índice de Precios Industriales (IPRI).

## 9.4. Índices encadenados

Los índices introducidos hasta el momento comparan directamente dos periodos de tiempo, 0 y $t$. La idea que se plantea en este apartado consiste en fragmentar el paso del periodo 0 al $t$ mediante el encadenamiento de las variaciones parciales en los periodos intermedios.

**Definición 9.9.** Se define el *índice encadenado o tipo cadena* entre los periodos 0 y $t$, asociado a la expresión de cálculo $I$, como el valor de la expresión:

$$CI_{t,0} = I_{t,t-1}I_{t-1,t-2}\cdots I_{1,0} = \prod_{k=1}^{t} I_{k,k-1} \tag{9.4.1}$$

Esta expresión puede ser aplicada a cualquiera de las fórmulas de cálculo introducidas en el apartado 9.1 (Laspeyres, Paasche, etc.). Así, por ejemplo, la expresión de un índice encadenado de precios de Laspeyres será:

$$CL^P_{t,0} = L^P_{t,t-1}L^P_{t-1,t-2}\cdots L^P_{1,0} = \prod_{k=1}^{t} \frac{\sum_{i=1}^{n} p_{i,k}q_{i,k-1}}{\sum_{i=1}^{n} p_{i,k-1}q_{i,k-1}}$$

Los índices encadenados han cobrado especial relevancia desde el año 2001 en que se renovó la metodología del IPC elaborado en España, pasando a ser calculado en la actualidad mediante un índice encadenado de Laspeyres.

### Ventajas e inconvenientes de los índices encadenados

- Los índices cadena verifican la propiedad de circularidad ya que si se considera un periodo intermedio $t'$ $(0 < t' < t)$, entonces: $CI_{t,0} = (I_{t,t-1} \cdots I_{t'+1,t'})(I_{t',t'-1} \cdots I_{1,0}) = CI_{t,t'}CI_{t',0}$. El cumplimiento de esta propiedad no depende de que la fórmula de cálculo $I$ verifique el requisito de circularidad; no obstante, en el caso de que sí lo verifique, la cifra de variación proporcionada por el índice encadenado coincidirá exactamente con la obtenida por comparación directa entre los periodos 0 y $t$ mediante la fórmula $I$.

- Desde un punto de vista práctico tienen mayor flexibilidad: la consideración de índices en periodos intermedios permite compensar un posible envejecimiento de la base del índice así como ir introduciendo cambios en su composición (por ejemplo, la incorporación de productos de nueva aparición en el mercado).

- Una limitación práctica de este tipo de índices es la necesidad de gran cantidad de información estadística. Pensemos por ejemplo en un índice de precios de Laspeyres: para la fórmula directa es necesario conocer los precios en los periodos 0 y $t$ y las cantidades tan solo en el periodo 0, mientras que para el cálculo del índice encadenado se precisan adicionalmente los precios y las cantidades en todos los periodos intermedios.

- El resultado de la variación global proporcionada por un índice encadenado depende generalmente del número de eslabones o periodos intermedios considerados. De hecho, el sesgo asociado a la utilización de una expresión de cálculo determinada aumentará de dimensión cuanto más larga sea la cadena.

## 9.5. Variación de un índice y repercusión

En el ejemplo 9.1 se obtuvieron las variaciones de los precios de los carburantes en el periodo 2007-2010, tomando como referencia el año 2007. La cuestión que surge ahora es la siguiente: ¿sería posible conocer a partir de dichos resultados la variación de los precios entre los años 2009 y 2010? Para ello se introduce el concepto de variación relativa.

**Definición 9.10.** Dada una serie de índices calculados tomando como referencia el periodo 0 $\{I_{t,0}\}$, se define la *variación relativa* (en %) del índice entre los periodos $t$ y $t$' como el valor de la expresión:

$$V_{t \to t'} = \frac{I_{t',0} - I_{t,0}}{I_{t,0}} 100 \tag{9.5.1}$$

La aplicación de esta expresión a las cifras del índice de precios de Laspeyres del ejemplo nos lleva a concluir que los carburantes aumentaron su precio en un 17,9 % entre los años 2009 y 2010.

Las cifras de variación entre dos periodos obtenidas a partir de índices sintéticos son valores promedio de la variación de los distintos componentes del índice. Sería interesante analizar separadamente los distintos efectos, pues una misma cifra de variación global puede responder a situaciones muy diferentes. Por ejemplo, una variación nula entre dos periodos puede ser consecuencia de que todos los componentes se han mantenido constantes en el tiempo o, alternativamente, que unos han experimentado cambios positivos y otros cambios negativos, que se han compensado entre sí.

**Definición 9.11.** La *repercusión* (en %) de un componente en la variación relativa del índice general entre dos periodos se define como la parte de la variación o efecto individual que corresponde a dicho componente.

Teniendo en cuenta esta definición, la repercusión de un componente puede interpretarse como la variación del índice si sólo hubiera experimentado cambios dicho componente. La expresión de cálculo de la repercusión depende del tipo de fórmula utilizada. A continuación se efectúa la deducción para el caso de un índice tipo media ponderada de base fija.

**Propiedad 9.6.** *La repercusión de un componente $i$ en la variación de un índice tipo media ponderada de base fija con ponderaciones $w_{i0}$ $(i = 1, ...n)$ al pasar de un periodo $t$ a un periodo $t'$ viene dada por el valor de la expresión:*

$$R_{t\to t'}(i) = \frac{1}{I_{t0}} \left[ (I_{t'0}(i) - I_{t0}(i)) \frac{w_{i0}}{\sum_{i=1}^{n} w_{i0}} \right] 100 \tag{9.5.2}$$

*Demostración.* Sustituyendo la expresión de cálculo de un índice media ponderada en el numerador de la fórmula de la variación relativa 9.5.1, se tiene que:

$$V_{t\to t'} = \frac{I_{t',0} - I_{t,0}}{I_{t,0}} 100 = \frac{1}{I_{t,0}} \left[ \frac{\sum_{i=1}^{n} I_{t'0}(i) w_{i0}}{\sum_{i=1}^{n} w_{i0}} - \frac{\sum_{i=1}^{n} I_{t'0}(i) w_{i0}}{\sum_{i=1}^{n} w_{i0}} \right] 100$$

$$= \frac{1}{I_{t,0}} \sum_{i=1}^{n} \left\{ [I_{t'0}(i) - I_{t0}(i)] \frac{w_{i0}}{\sum_{i=1}^{n} w_{i0}} \right\} 100$$

La expresión de la repercusión propuesta en el enunciado 9.5.2 se corresponde con el sumando i-ésimo de la relación anterior ya que si entre los periodos $t$ y $t'$ sólo experimenta variaciones el componente i-ésimo, todas las diferencias de los índices simples entre corchetes serán nulas a excepción de la correspondiente a dicho componente. □

La expresión 9.5.2 puede aplicarse a los índices de Laspeyres; en particular, si se considera la fórmula de precios con ponderaciones $w_{i0} = p_{i0}q_{i0}$, se deduce fácilmente la expresión de la repercusión para $L^P$:

$$R_{t\to t'}(i) = \frac{\left[I^P_{t'0}(i) - I^P_{t0}(i)\right] \dfrac{p_{i0}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}}}{L^P_{t0}} 100 = \frac{\left[\dfrac{p_{it'}}{p_{i0}} - \dfrac{p_{it}}{p_{i0}}\right] \dfrac{p_{i0}q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}}}{L^P_{t0}} 100$$

$$= \frac{\left[p_{it'} - p_{it}\right] \dfrac{q_{i0}}{\sum_{i=1}^{n} p_{i0}q_{i0}}}{L^P_{t0}} 100$$

Tanto la variación relativa de un índice como las repercusiones pueden ser negativas, positivas o nulas. Además, de la propia definición de la repercusión se desprende que la suma de todas las repercusiones da como resultado la variación del índice, es decir: $\sum_{i=1}^{n} R_{t\to t'}(i) = V_{t\to t'}$.

En el ejemplo 9.1, el tipo de carburante con mayor efecto individual en la variación global de precios entre los años 2009 y 2010 es el gasóleo, con una repercusión del 12,6 %. O equivalentemente, si el precio de la gasolina sin plomo no hubiera cambiado entre los años 2009 y 2010, el índice global de precios de los carburantes habría aumentado en un 14,2 %. De forma complementaria se obtiene que la repercusión de la gasolina sin plomo fue del 3,7 %.

# 10 El Índice de Precios de Consumo y sus aplicaciones

## 10.1. El Índice de Precios de Consumo (IPC)

El *Índice de Precios de Consumo* (IPC) es un indicador mensual elaborado por el INE, cuyo objetivo es medir la evolución del nivel de precios de los bienes y servicios de consumo adquiridos por los hogares residentes en España.

EL IPC es uno de los principales indicadores de la coyuntura económica del país, utilizado principalmente como medida de la inflación. Pero tiene además otras muchas aplicaciones y de gran importancia en los ámbitos económico, jurídico y social: revalorización de las pensiones, actualización del salario mínimo interprofesional y de los sueldos de los funcionarios de las Administraciones Públicas, revisiones salariales pactadas en convenios colectivos, revisión de los contratos de arrendamiento de inmuebles, etc.

La metodología para la elaboración del IPC es compleja y su diseño responde a la necesidad de satisfacer dos requisitos básicos[1]:

- Representatividad. Dado que cada familia tiene sus propias pautas de consumo, ¿cómo será posible obtener un indicador que represente adecuadamente los cambios en los precios de los productos que consumen todas las familias? Para ello, los artículos incluidos en el índice deben ser los más consumidos por la mayoría de la población, los establecimientos donde se observan los precios deben ser los más visitados y las ponderaciones o pesos relativos de cada artículo en el índice deben responder a las tendencias de consumo de los hogares.

- Medir las variaciones de precios "puras", sin verse afectadas por cambios en la calidad de los productos, ni en las pautas de consumo de los hogares, ni en la metodología de elaboración del índice,.... Es decir, se trata de que al comparar las cifras de IPC de dos periodos de tiempo, la variación obtenida sea debida exclusivamente a cambios en los precios.

En este marco, un Sistema de Índices de Precios de Consumo consta de una serie de elementos cuya determinación viene guiada por los requisitos anteriores. Los más destacables son los siguientes:

- El *estrato de referencia*, que es el grupo de población cuya estructura de consumo sirve de base para el cálculo del Índice de Precios de Consumo.

[1] Este apartado es un resumen de la metodología detallada que está disponible en la web del INE www.ine.es

- La *cesta de la compra* o conjunto de bienes y servicios que consumen habitualmente los integrantes del estrato de referencia y que, por tanto, serán objeto de observación para el cálculo del índice. Es conveniente tener presente que se consideran únicamente bienes destinados al consumo, por lo que están excluídos, entre otros, los gastos en bienes de inversión tales como la adquisición de una vivienda.

- Las *ponderaciones*, que representan la importancia relativa que tiene cada artículo que compone la cesta de la compra frente a los demás. Tanto la cesta de la compra como las ponderaciones se determinan a partir de la información proporcionada por las encuestas de presupuestos familiares.

- La *muestra de municipios y establecimientos*. Para elaborar el IPC se precisa también disponer de información permanentemente actualizada de los precios de los artículos que integran la cesta de la compra. Estos datos se obtienen mediante una encuesta mensual realizada a una muestra de municipios y establecimientos representativos en los que se observan los precios de dichos artículos.

- La *fórmula de cálculo*. Dado que el objetivo del IPC es aislar los cambios en los precios, su cálculo se basa en un índice de precios, generalmente mediante fórmulas de base fija, tipo Laspeyres.

### El IPC base 2011

Los elementos que componen el Sistema de IPC deben permanecer estables a lo largo del tiempo con el fin conseguir la comparabilidad temporal de las variaciones en los precios. No obstante, las pautas de consumo de los hogares van experimentando cambios con el paso del tiempo: se reduce el consumo de algunos productos y, por otra parte, aparecen nuevos productos en el mercado; pensemos por ejemplo en los cambios tan rápidos vinculados a las nuevas tecnologías de la información. Por ello, es necesario revisar y actualizar los componentes del IPC cada cierto tiempo, de modo que se garantice su representatividad. En este sentido, el IPC español se renueva cada cinco años, siendo la última renovación realizada la correspondiente al año 2011. A continuación se analizan los aspectos concretos del sistema de IPC vigente:

- El estrato de referencia del IPC base 2011 abarca a toda la población residente en viviendas familiares en España. Por tanto, únicamente queda excluido el gasto de los residentes en hogares colectivos (conventos, residencias de ancianos, prisiones, ...), así como el de los residentes en el extranjero.

- La composición de la cesta de la compra y su estructura de ponderaciones se basa principalmente en la información sobre gastos de consumo de los hogares proporcionada por la Encuesta de Presupuestos Familiares. La cesta de la compra del IPC base 2011 está integrada por artículos representativos de las diferentes parcelas de consumo que superan el 0,3 por mil del gasto total de los hogares, un total de 489 artículos.

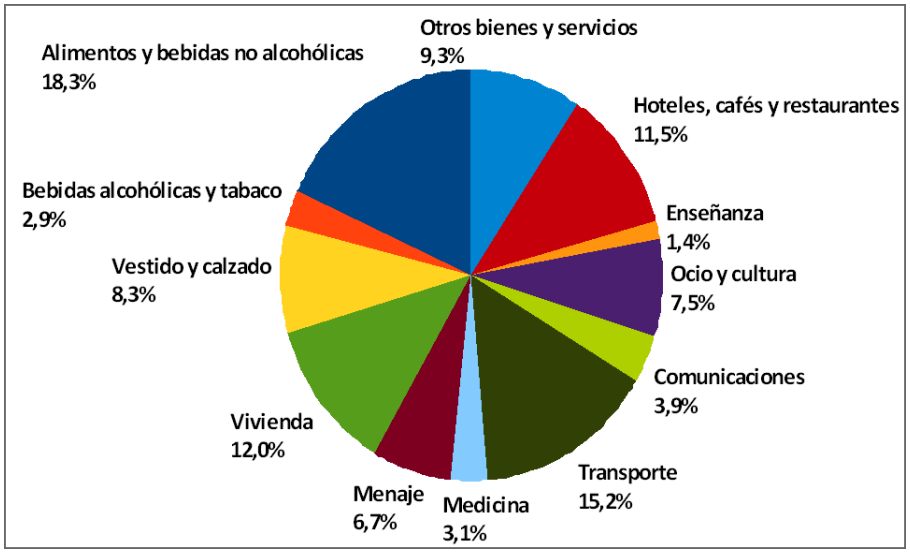



Figura 10.1: Ponderaciones IPC 2011 (actualización 2012)

- La ponderación asociada a cada uno de los artículos que forman parte de la cesta de la compra representa la relación entre el gasto realizado en dicho artículo y el gasto total:

$$w_i = \frac{\text{gasto realizado en las parcelas representadas por el artículo i}}{\text{gasto total}}$$

Los artículos de la cesta de la compra se clasifican en 12 grandes grupos de gasto, de los que las mayores ponderaciones corresponden al grupo Alimentos y bebidas no alcohólicas, seguido de Transporte. Las ponderaciones se actualizan cada año según la última información anual disponible de la Encuesta de Presupuestos Familiares.

Según los datos recogidos en la figura 10.1, podemos interpretar que, por término medio, una familia del estrato de referencia destinará de cada 1.000€ de su presupuesto para consumo, 183€, a gastos de alimentación y bebidas no alcohólicas, 152€, a transporte, 115€, a gastos relacionados con Hoteles, cafés y restaurantes, etc.

En el caso español, el Instituto Nacional de Estadística ha elaborado a lo largo del tiempo distintos sistemas de Índices de Precios de Consumo. El primero de ellos, bajo el nombre de Índices de Coste de la Vida, se inició en 1939, tomando como base el mes de julio de 1936. Las características de este índice eran muy diferentes a las del IPC actual: el estrato de referencia estaba constituido por las familias de

4 o 5 miembros con ingreso mensual de 600 pesetas (unos 787€ de 2011) y la cesta de la compra estaba integrada por una media de 115 artículos en cada capital de provincia, clasificados en 5 grupos, de los que el grupo de alimentación y bebidas tenía un peso del 60 %, en torno al triple que en el IPC base 2011.

- Otro de los elementos que intervienen en el diseño del IPC es la muestra de municipios y establecimientos. En el sistema base 2011, se han seleccionado 177 municipios (las 52 capitales de provincia y 125 municipios no capitales), garantizando una cobertura de población del 50 % de cada comunidad autónoma. En dichos municipios se ha seleccionado una muestra de unos 29.000 establecimientos comerciales de modo que estén representados todo tipo de establecimientos (hipermercados, mercados, tiendas especializadas, ...) y de zonas comerciales y que, además sean los de mayor afluencia de público y ventas de la localidad. Entre los días 1 y 22 de cada mes, agentes entrevistadores del INE acuden a los establecimientos de la muestra y toman los precios de los artículos de la cesta de la compra con las características especificadas previamente. Es importante este punto puesto que, de no mantenerse las especificaciones de los productos todos los meses, se estarían midiendo no sólo los cambios en los precios sino también en la calidad. De este modo, se recogen aproximadamente 220.000 precios cada mes.

- Por último, la fórmula de cálculo del IPC base 2011 es el índice de precios de Laspeyres encadenado. Esto significa que para comparar el periodo actual ($t$) con el periodo base (0) se consideran otros periodos intermedios ($k$), que en el sistema base 2011 corresponden a los meses de diciembre de todos los años. Así, el índice del mes $m$ del año $t$ se obtendrá como: $IPC_{mt,06} = IPC_{mt,dic(t-1)} \times IPC_{dic(t-1),11}$. Una ventaja de la utilización de fórmulas encadenadas es la posibilidad de incorporar anualmente las actualizaciones de las ponderaciones, lo que permite ir adaptando el IPC a los cambios del mercado y de los hábitos de consumo en plazos muy breves de tiempo.

Los resultados del IPC de cada mes son dados a conocer en la primera quincena del mes siguiente, de acuerdo con el calendario de disponibilidad establecido por el INE. En la fecha prevista pueden consultarse en la web del INE los datos del IPC general y con distintos niveles de desagregación:

- Desagregación geográfica: además del IPC general para el conjunto del territorio nacional, se publican índices por comunidades autónomas y por provincias.

- Desagregación funcional: se publican datos para los doce grandes grupos de gasto, para grupos especiales y, ya más específicos, para pequeños grupos o rúbricas de gasto.

Además de los índices se publican distintas tasas de variación, tanto globales como por grupos, que para el mes $m$ del año $t$ se calculan a través de las siguientes expresiones:

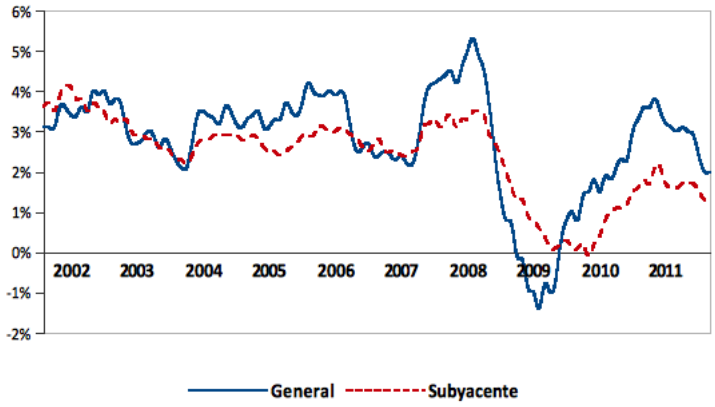



Figura 10.2: Tasas interanuales de inflación general e inflación subyacente

- Tasa intermensual o variación sobre el mes anterior:

$$V_{(m-1)t\to mt} = \left(\frac{IPC_{mt,11}}{IPC_{(m-1)t,11}} - 1\right) 100$$

- Tasa acumulada o variación en lo que va de año (toma como referencia el mes de diciembre del año anterior):

$$V_{dic(t-1)\to mt} = \left(\frac{IPC_{mt,11}}{IPC_{dic(t-1),11}} - 1\right) 100$$

- Tasa interanual o variación en un año:

$$V_{m(t-1)\to mt} = \left(\frac{IPC_{mt,11}}{IPC_{m(t-1),11}} - 1\right) 100$$

Como consecuencia de su definición, la tasa acumulada (que cuantifica la variación del IPC en lo que va de año) coincidirá en el mes de enero de cada año con la tasa intermensual y en el mes de diciembre con la tasa interanual.

Entre los grupos especiales cabe destacar, por su importancia en los análisis económicos, el resultado correspondiente a la *inflación subyacente*, que se obtiene a partir del Índice General excluyendo los alimentos no elaborados y los productos energéticos. La inflación subyacente suele tener un comportamiento más estable como se puede observar en el gráfico 10.2, dado que no incluye productos cuyos precios pueden ser más erráticos por motivos climáticos o estacionales (caso de algunos alimentos no elaborados) o por tratarse de productos importados (caso del petróleo).

### Enlace de series

Cada vez que se lleva a cabo un cambio en la base del IPC, se produce una ruptura en la continuidad de las series como consecuencia de los cambios introducidos en el sistema: actualización de ponderaciones, nueva cesta de la compra y cambios metodológicos en general. En la práctica los usuarios precisan series continuadas y dado que los cambios de base son inevitables, el INE ha diseñado procedimientos que permiten conocer la variación de los precios entre dos meses cualesquiera, aunque correspondan a sistemas de IPC calculados con distintas bases. Hasta que entró en vigor el sistema de IPC base 2006, el INE publicaba unos coeficientes de enlace entre las series de distintas bases; así, por ejemplo, los coeficientes de enlace correspondientes al paso de base 1992 a base 2001, se obtuvieron a partir de las cifras de IPC de diciembre de 2001 calculadas según ambas metodologías. A partir del año 2006 no se precisan coeficientes de enlace ya que por ser la fórmula del IPC un índice encadenado, el propio método de cálculo permite realizar el enlace con la serie base 2001.

El INE ha implementado en su pagina web una aplicación que permite obtener de forma inmediata la tasa de variación del IPC entre meses de dos años cualesquiera desde 1961, respondiendo así a una de las demandas más frecuentes por parte de los usuarios.

## 10.2. El IPC armonizado

El *Índice de Precios de Consumo Armonizado* (IPCA) es un indicador cuyo objetivo es proporcionar una medida común de la inflación que permita realizar comparaciones entre los países miembros de la Unión Europea.

Se trata de un indicador fundamental para el funcionamiento de la Unión Europea y muy especialmente para los países que forman parte del euro: el artículo 127 del Tratado de Funcionamiento de la Unión Europea establece que "*el objetivo principal del Eurosistema será mantener la estabilidad en precios ...*", estabilidad en precios que el Banco Central Europeo define como "*un incremento interanual del Índice de Precios de Consumo Armonizado (IPCA) para la zona euro inferior al 2 %*".

> Asimismo, el IPCA fue el indicador utilizado para examinar el cumplimiento del requisito en materia de inflación establecido por el Tratado de Maastrich, según el cual los países comunitarios que desearan formar parte de la Unión Económica y Monetaria deberían tener un índice de precios al consumo no superior a 1,5 puntos porcentuales del IPC medio de los tres países con la inflación más baja.

El IPCA se obtiene a partir de los IPC nacionales de cada Estado miembro de la Unión Europea, mediante un proceso de armonización establecido por EUROSTAT. En cada país el IPCA cubre las parcelas de consumo que superan el uno por mil del total de gasto de la cesta de la compra nacional y se realizan ajustes para conseguir la comparabilidad deseada mediante determinadas inclusiones o exclusiones de partidas de consumo. En el caso de España, desde enero de 2001 la única diferencia entre el IPCA y el IPC en cuanto a cobertura de bienes y servicios se refiere al tratamiento

de los seguros y la compra de automóviles usados. En lo que se refiere a cobertura de población el IPCA incluye los gastos de los turistas en España y excluye el de los españoles en el extranjero. En cambio, el IPC nacional sólo contempla el gasto realizado por los hogares residentes en España (independientemente del lugar en el que se haya realizado); por ello, la principal diferencia en la estructura de ponderaciones entre el IPC y el IPCA está en el grupo "Hoteles, cafés y restaurantes" con ponderaciones respectivas 12 % y 14,5 % (actualización de 2010). El IPCA se calcula mediante la fórmula del índice de precios de Laspeyres encadenado, tomando como base el año 2005.

Eurostat publica mensualmente las cifras del IPC armonizado de los 27 países miembros de la Unión Europea y de otros tres países adicionales, Islandia, Noruega y Suiza. Además publica tres índices agregados:

- El Índice de Precios de Consumo de la Unión Monetaria, calculado como una media ponderada de los IPCA de los 16 países que integran la zona euro.
- El Índice de Precios de Consumo Europeo, que es una media ponderada de los IPCA de todos los países de la UE.
- El Índice de Precios de Consumo del Área Económica Europea, que además de los países de la UE incluye Islandia y Noruega.

Además, desde el año 2009 ha empezado a publicarse el IPCA a Impuestos Constantes (IPCA-IC), que tiene como objetivo medir la evolución de los precios de consumo pero considerando que los impuestos permanecen constantes. De este modo la diferencia entre el IPCA y el IPCA-IC representa el efecto de cambios en los impuestos.

## 10.3. Aplicaciones económicas del IPC

La capacidad de los índices de precios, y muy especialmente del IPC, para medir la inflación justifica su gran popularidad y el gran número de aplicaciones en el ámbito económico, que centraremos en la deflactación y, en un sentido inverso, la indexación.

El IPC es uno de los deflactores más utilizados. Su uso será especialmente adecuado cuando se trate de analizar la evolución real de magnitudes monetarias vinculadas con su cobertura, es decir, con todos aquellos valores relacionados con el consumo de los hogares. El IPC será por tanto el índice de precios adecuado para deflactar los gastos a precios corrientes de los hogares o para conocer la evolución de su poder adquisitivo a través del estudio de la evolución real de las rentas salariales.

En un sentido inverso a la *deflactación* se plantea la indexación o actualización como un proceso mediante el cual se revisan determinados valores monetarios según la variación experimentada por un índice de precios. Este procedimiento se utiliza habitualmente para actualizar ingresos monetarios tales como salarios, pensiones de la Seguridad Social o pensiones alimenticias en sentencias de divorcio, con el fin de garantizar que se mantiene el poder adquisitivo de los perceptores. A modo de ilustración, se incluye a continuación la legislación relativa a algunas de estas actualizaciones

en las que se utiliza el IPC: en primer lugar la revalorización de pensiones de la Seguridad Social, en segundo lugar la actualización de contratos de arrendamiento y, por último, un ejemplo de incremento salarial en el convenio colectivo de un sector.

- "*Las pensiones de la Seguridad Social en su modalidad contributiva, incluido el importe de la pensión mínima, serán revalorizadas al comienzo de cada año, en función del correspondiente índice de precios al consumo previsto para dicho año*". [Ley 24/1997, de 15 de julio, de consolidación y racionalización del Sistema de Seguridad Social, Artículo 11].

- "*Durante los cinco primeros años de duración del contrato, la renta sólo podrá ser actualizada por el arrendador o el arrendatario en la fecha en que se cumpla cada año de vigencia del contrato, aplicando a la renta correspondiente a la anualidad anterior la variación porcentual experimentada por el Índice General Nacional del Sistema de Índices de Precios de Consumo en un periodo de doce meses inmediatamente anteriores a la fecha de cada actualización, tomando como mes de referencia para la primera actualización el que corresponda al último índice que estuviera publicado en la fecha de celebración del contrato, y en las sucesivas el que corresponda al último aplicado*". [Ley 29/1994 de 24 de Noviembre, de arrendamientos urbanos, Artículo 18].

- "*En lo que se refiere al incremento salarial, se establece para el año 2009 el IPC real de dicho año con un mínimo garantizado del 0,5 % sobre las tablas salariales anejas a este convenio, con efectos a partir de 1 de enero de dicho año*". [Convenio colectivo estatal para las empresas de publicidad, BOE 24/2/2010, Artículo 28].

En el último caso y teniendo en cuenta que la cifra interanual de inflación de 2009 (correspondiente al mes de diciembre) fue del 0,8 %, el convenio publicado garantiza que un trabajador del sector de publicidad con un salario mensual de 1.000€ en 2008 tendría un salario de 1.008€ en 2009.

# 11 Series temporales: planteamiento y tendencia

La comprensión de gran parte de los fenómenos de interés en economía se verá notablemente facilitada si disponemos de información histórica sobre los mismos.

En efecto, en el ámbito económico-empresarial aparece frecuentemente la necesidad de adoptar decisiones que se hallan condicionadas por el valor futuro de cierta característica. La elaboración de pronósticos para estos valores futuros exige disponer de información sobre la evolución histórica de las correspondientes magnitudes.

Este sería el caso si, por ejemplo, queremos elegir una opción entre varias alternativas posibles de inversión cuya rentabilidad futura debemos estimar o si decidimos adquirir un terreno bajo el supuesto de que la población -y en consecuencia los precios del suelo- aumentarán en determinada zona. En cualquiera de estos supuestos la decisión no se basará solamente en los valores actuales de la variable sino que se fundamentará también en el análisis de sus valores históricos así como en otras técnicas, como la opinión de expertos, que proporciona información de tipo cualitativo o el análisis de la relación con otras variables, a través de los estudios de regresión.

La historia de una serie temporal y por tanto los periodos en los que se llevaron a cabo las observaciones van a tener un papel relevante en su análisis. Según que la perspectiva adoptada sea de corto, medio o largo plazo, nuestro análisis se centrará en distintos componentes de la serie.

El análisis clásico de series temporales consiste en identificar en las mismas distintos componentes que pueden ser aislados y analizados separadamente. De estos componentes el que tendrá un mayor peso en la trayectoria de la serie será la tendencia que estudiamos en el apartado final de este tema.

Este análisis clásico es equiparable a una "*disección estadística*" que se basa en distintos instrumentos de análisis y sirve de impulso a otras ramas de investigación como el análisis de coyuntura.

## 11.1. Evolución temporal de magnitudes

**Definición 11.1.** Denominamos en términos genéricos, *serie temporal* (también designada como *serie histórica* o *cronológica*) a una sucesión de observaciones de una variable a través del tiempo.

Aunque este concepto conlleva la consideración conjunta de dos variables que podrían admitir la interpretación de *variable dependiente* en el caso de la magnitud analizada e *independiente* el tiempo, esta traslación de los conceptos de regresión no es totalmente

válida, resultando más adecuado efectuar un planteamiento no causal en el que el tiempo actúa únicamente como referencia o soporte, siendo la magnitud analizada la única variable del estudio.

En este tipo de estudios denominados *de corte longitudinal* se recoge la evolución de una o varias magnitudes a lo largo del tiempo. La diferencia fundamental respecto al análisis causal presente en regresión es que la explicación -y posteriormente la realización de predicciones- se basará ahora en la información directa que el pasado proporciona sobre el fenómeno que estamos estudiando.

La información inicial de un estudio temporal va referida a una magnitud económica, que puede pertenecer a dos modalidades: *stock* y *flujo*.

- En las variables stock o nivel cada observación se refiere a un instante determinado.
- En el estudio de variables flujo las observaciones van referidas a un periodo de tiempo.

Las categorías de variables stock y flujo aparecen claramente recogidas en la distinción -apuntada ya en la obra de Adam Smith (1723-1790)- entre los conceptos de riqueza y renta. El primero de ellos cuantifica los fondos o sotcks de una nación o sociedad mientras el segundo se define habitualmente como corriente de los bienes y servicios generados en la sociedad durante un periodo de tiempo y pertenece, por tanto, a la categoría de flujo.

La notación utilizada para el análisis de series temporales depende de cómo estemos considerando el periodo de observación:

- Periodos correlativos $1, \cdots, T$, entonces se denota por $Y_t$ el valor de la magnitud en el periodo $t$. En este caso la serie se representa:

| Periodo | $Y_t$ |
|---|---|
| 1 | $Y_1$ |
| 2 | $Y_2$ |
| $\vdots$ | $\vdots$ |
| $T$ | $Y_T$ |

- Si consideramos subperiodos, tendremos un doble subíndice, $i\,(i = 1, \ldots, n)$ que indica el periodo (generalmente el año) y $j\,(j = 1, \ldots, m)$ que señala el correspondiente subperiodo (habitualmente mes, trimestre, semestre,...). Así, para series mensuales denotamos por $Y_{ij}$ el valor de la magnitud en el mes $j$ del año $i$, y la representación de la serie completa sería:

| Años \ Meses | 1 | $\cdots$ | $m$ |
|---|---|---|---|
| 1 | $Y_{11}$ | $\cdots$ | $Y_{1m}$ |
| 2 | $Y_{21}$ | $\cdots$ | $Y_{2m}$ |
| $\vdots$ | $\ddots$ | $\ddots$ | $\vdots$ |
| n | $Y_{n1}$ | $\cdots$ | $Y_{nm}$ |

En el ámbito económico resulta necesario efectuar algunas consideraciones referidas a la repetición de la magnitud en el tiempo, la estabilidad de las estructuras que condicionan su evolución o la permanencia de su definición, aspectos que resultan claves para garantizar la comparabilidad temporal de la variable.

- Usualmente supondremos que las observaciones se distribuyen regularmente en el tiempo, pero en muchos casos este supuesto no garantiza la comparabilidad. La presencia de fiestas móviles cada año, la existencia de meses con distinto número de días o de fines de semana, la existencia de fenómenos meteorológicos, ... hacen que en algunos casos las observaciones no sean totalmente comparables. Así, por ejemplo los índices de producción estarán muy afectados por el número de días laborables de cada mes.

- En algunos casos se pueden aplicar ciertas soluciones que ayudan a depurar y centrar mejor la serie.

  - *Efecto calendario*: podríamos ajustar la duración de cada mes a un periodo de 30 días, corrigiendo, por ejemplo, la producción de febrero mediante el factor $\frac{30}{28}$, la de marzo con el factor $\frac{30}{31}$, etc., con el inconveniente de que el total anual no coincidirá con la suma de los doce meses así ajustados.
  - Existen otros mecanismos más satisfactorios para corregir el efecto calendario, excluyendo las fiestas nacionales, regionales o locales y considerando también los fines de semana de cada mes. No obstante, estos métodos requieren un tratamiento más avanzado al de este libro.
  - También existen métodos más o menos sofisticados para valorar impactos como el *Efecto Pascua* (que es un efecto movible entre marzo y abril de cada año) o el *Puente de la Constitución* (que es un efecto permanente en el mes de diciembre). Como en el caso anterior, estos enfoques rebasan el ámbito de este libro.

- Por lo que se refiere a los cambios estructurales o coyunturales, que resultan inevitables en las series de carácter histórico, pueden revestir diferentes niveles de gravedad:

  - Cuando estos cambios se producen de forma lenta (alteración de costumbres, modas, etc.) influirán en el movimiento a largo plazo de la variable considerada, y resultarán en consecuencia incluidos en alguno de los componentes de la serie.
  - En el caso de efectos bruscos y pasajeros (huelgas, catástrofes, accidentes, cambios excepcionalmente extremos, ....) podría resultar conveniente, para que no distorsionen el análisis global del fenómeno, aminorar su impacto considerando estos valores como "impulsos" o valores atípicos dentro de la serie.

- Si, por el contrario, los cambios son bruscos y de carácter permanente pueden hacer aconsejable la consideración de distintos subperiodos diferenciados en el estudio del fenómeno, o bien tratar la serie completa incluyendo un "escalón" en la misma.

- Otro problema que se plantea a menudo en el estudio de series temporales es la ausencia de homogeneidad en las observaciones, hecho que puede ser debido a distintas razones, entre las que se encuentran la posible mejora de los métodos de observación, las variaciones en las definiciones estadísticas o las alteraciones en los productos o en la estructura social.

Aunque la permanencia de la definición de la magnitud estudiada parece un requisito teórico imprescindible para el análisis de su evolución temporal, en la práctica nos enfrentaremos a menudo con series en las que se han operado cambios metodológicos de diversa índole (alteración de unidades de medida, cambios de base en índices, actualización de ponderaciones o de las muestras de artículos considerados, etc.) que -aunque incorporan mejoras en la fiabilidad de la variable- alteran su carácter e invalidan las comparaciones.

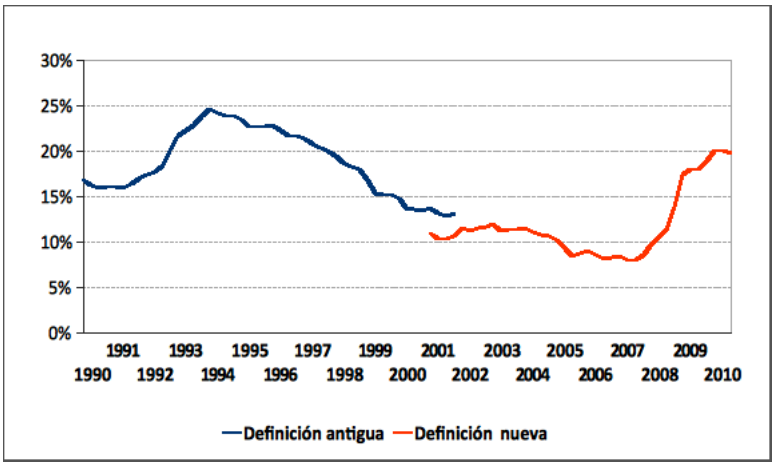



Figura 11.1: Tasa de paro en España ( %)

La figura 11.1 recoge la evolución temporal de la tasa de paro en España en los últimos años según los datos de la EPA. Llama la atención el salto en la continuidad de la serie que se produce en el año 2002, justificado por un cambio en la definición de "parado". En efecto, a partir de 2002 la clasificación de una persona como parada añade a los requisitos ya existentes la búsqueda activa de empleo, lo que supuso que por motivos metodológicos algunas personas consideradas paradas en encuestas anteriores pasaran a la categoría de inactivas.

## 11.2. Componentes de una serie temporal

La aproximación inicial de un estudio temporal se efectuará mediante el análisis de sus datos numéricos o del gráfico temporal que los representa. De este modo es

posible apreciar las características más sobresalientes del fenómeno en estudio, aunque en etapas posteriores nos interesará profundizar más en sus diferentes componentes.

Vamos a definir y posteriormente intentar separar distintos elementos que influyen en una serie temporal; pero queremos dejar claro que estos componentes no son observables, ya que la única magnitud observable y por tanto la única fuente de información real es la serie temporal. La aproximación de estos factores de la serie resulta útil para su análisis e incluso para hacer predicciones, pero nunca podremos garantizar que el componente es exactamente el estimado porque, al no ser observables sus valores reales, nunca podremos contrastarlo.

En una primera aproximación, el examen de cualquier serie cronológica suficientemente amplia nos permitirá apreciar un movimiento de carácter general, que llamamos tendencia.

**Definición 11.2.** Dada una serie temporal $Y_t$ (o $Y_{ij}$), denominamos *tendencia*, $T_t$ (o $T_{ij}$) al movimiento general a largo plazo de la serie.

Este componente recoge la evolución en plazos de 10, 20 o 30 años y se obtiene mediante filtros que proporcionan los patrones o pautas generales de comportamiento de la serie.

En la práctica, y en especial en el ámbito económico, la tendencia aparece afectada por oscilaciones de carácter no regular, que inciden en el valor de la magnitud analizada en uno u otro sentido según la fase económica en la que nos encontremos. Este componente se denomina factor cíclico ($C$) y el tipo de variaciones que contempla han sido estudiadas con éxito en distintos campos.

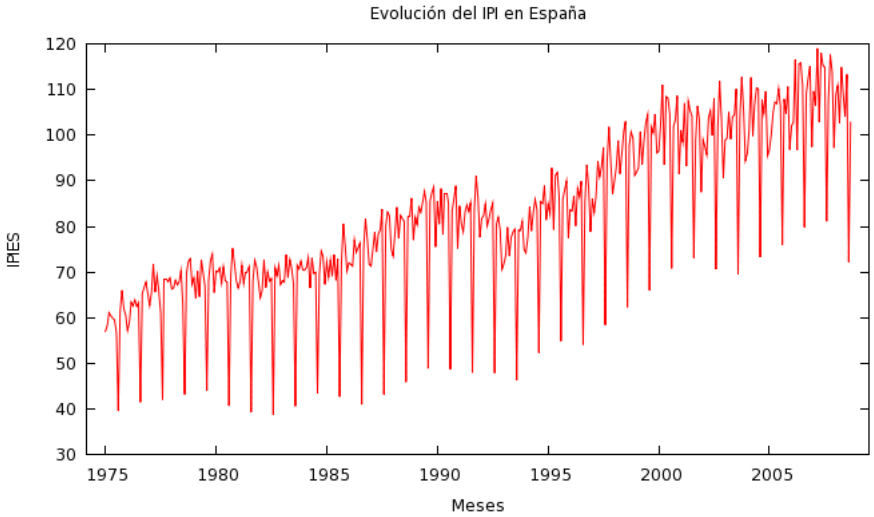
**Definición 11.3.** Dada una serie temporal $Y_t$ (o $Y_{ij}$), denominamos *ciclo o componente cíclico*, $C_t$ (o $C_{ij}$) a las oscilaciones en torno a la tendencia que se producen en un plazo medio de 3, 5, 8 o incluso más años.

Su identificación resulta especialmente compleja en el contexto económico debido a la ausencia de regularidad y a la frecuente aparición de distintos movimientos cíclicos superpuestos. Por este motivo muchas veces resulta interesante considerar conjuntamente los ciclos y la tendencia.

**Definición 11.4.** Dada una serie temporal $Y_t$ (o $Y_{ij}$), denominamos componente *extraestacional*, $E_t$ (o $E_{ij}$) a los movimientos a medio y largo plazo (superiores al año), que combinan los efectos de tendencia y de las variaciones cíclicas. Este componente se denomina también Tendencia-ciclo $TC_t$ (o $TC_{ij}$) .

Además de las oscilaciones cíclicas, en las series temporales aparecen frecuentemente una serie de movimientos a corto plazo (entendiendo como tales aquéllos cuya duración es inferior a un año).

**Definición 11.5.** Dada una serie temporal $Y_t$ (o $Y_{ij}$), denominamos componente *estacional*, $e_t$ (o $e_{ij}$), a las variaciones de carácter periódico en torno a la tendencia que se producen en el corto plazo (periodos siempre inferiores al año).




En la gráfica del Índice de Producción Industrial (IPI) observamos un comportamiento sistemático en determinados meses del año. Podemos hacer un zoom y ampliar esta gráfica para observar con mayor detalle ese comportamiento periódico.

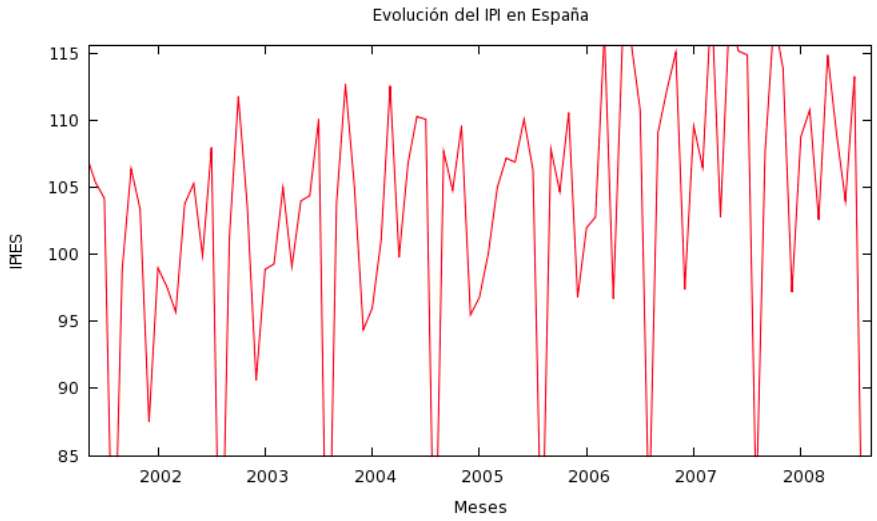



Figura 11.2: Componentes de una serie temporal

El origen de las variaciones de tipo estacional puede ser de carácter físico-natural (fluctuaciones periódicas debidas a ciclos biológicos, tiempo meteorológico, etc.) o bien institucional (calendarios laborales, horarios comerciales, etc.). En cualquier caso, su rasgo definitorio es la periodicidad, que posibilita la cuantificación de estas variaciones e incluso su eliminación de la serie original.

> Parece razonable admitir por su propio carácter que, si la estacionalidad existe, ésta sea constante para cada época considerada, hecho que facilitaría notablemente su aproximación cuantitativa. Sin embargo, en algunas ocasiones se presentan patrones de estacionalidad variable.

Por último, conviene tener presente que las series temporales no presentan un comportamiento completamente sistemático, sino que junto a las variaciones anteriores existen otras de carácter irregular, accidental o residual.

**Definición 11.6.** Dada una serie temporal $Y_t$ (o $Y_{ij}$), denominamos componente *residual o accidental*, $u_t$ (o $u_{ij}$), a las variaciones irregulares, no controladas ni modelizables que se van produciendo a lo largo del tiempo de manera no predecible.

Consideramos incluidos en este componente residual dos tipos de variaciones:

- Las aleatorias, que recogen los pequeños efectos de carácter accidental y por tanto no identificables, y
- las erráticas, que se producen como consecuencia de hechos no siempre previsibles pero que pueden ser identificados a posteriori (es el caso de huelgas, cambios institucionales, catástrofes naturales, etc.).

Existen diferentes hipótesis sobre la forma en que los diversos componentes de una serie temporal interactúan dando lugar a la observación final de la variable.

El componente residual es el único que no es modelizable. Así pues, si pudiésemos modelizar el resto de los componentes de la serie, y denotamos el efecto estimado por $f(t)$, entonces el valor residual en el periodo $t$ podría calcularse como la diferencia entre el valor registrado y el estimado en ese periodo: $u_t = Y_t - f(t)$. Por tanto el componente residual tendrá un efecto aditivo sobre la serie (los residuos no deberán estar relacionados con ningún otro componente de la serie temporal).

No obstante, será necesario efectuar supuestos sobre las influencias existentes entre la tendencia, ciclos y variaciones estacionales (movimientos a largo, medio y corto plazo, respectivamente), ya que estas interrelaciones condicionarán los métodos para su descomposición y análisis.

- Si admitimos que las desviaciones respecto a la tendencia no se ven afectadas por la magnitud de ésta (es decir, bajo el supuesto de independencia entre la tendencia y las fluctuaciones de la serie), nos encontraremos ante una hipótesis de composición que llamaremos *esquema aditivo*. En este caso tendríamos: $f(t) = T_t + C_t + e_t$, y por tanto:

$$Y_t = \underbrace{T_t + C_t}_{E_t} + e_t + u_t = E_t + e_t + u_t \tag{11.2.1}$$

donde en la última expresión hemos introducido la componente extraestacional.

- Sin embargo la hipótesis anterior de "independencia" no resulta muy creíble, especialmente en el campo económico, donde parece que las fluctuaciones deberían tener un carácter relativo y ser mayores para valores altos de la serie y menores para valores bajos. Este supuesto conduce a una hipótesis de composición que se denomina *esquema multiplicativo*, $f(t) = T_t \times Ic_t \times Ie_t = E_t \times Ie_t$.

  Respecto a la notación anterior, hemos introducido una nueva terminología, $Ic_t$ e $Ie_t$ que representan los índices de variación cíclica y estacional, y cuantifican las variaciones relativas o proporcionales respecto al valor de la tendencia del periodo. Se tiene:

$$Y_t = E_t \times Ie_t + u_t \tag{11.2.2}$$

En ambos modelos el valor final viene en las unidades de la magnitud considerada, en un caso directamente como suma y en el otro al multiplicar las unidades por unas tasas en tantos por uno.

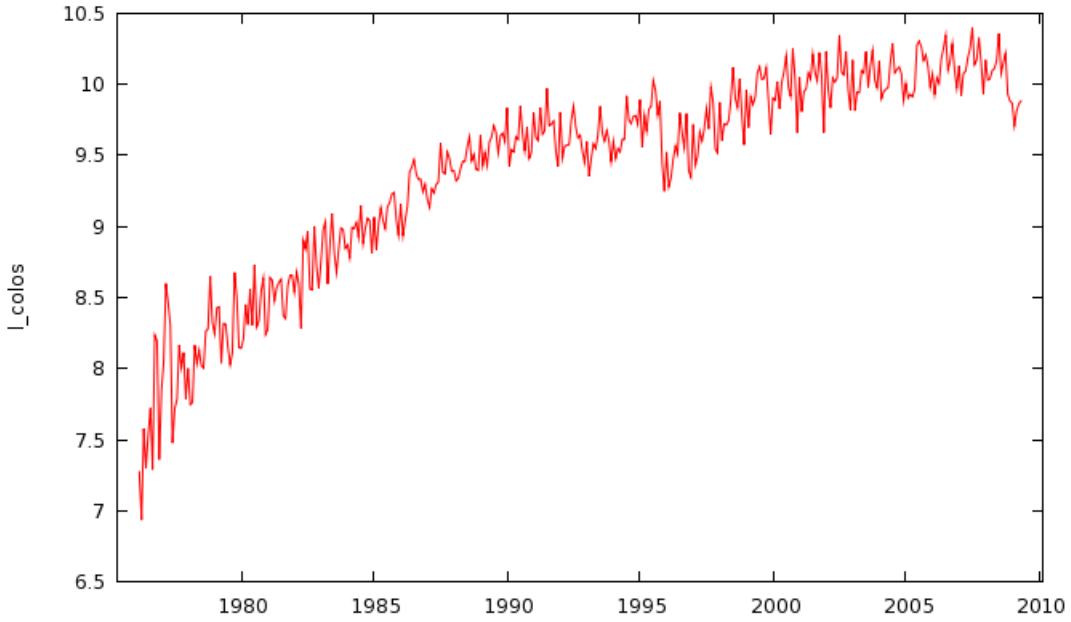
> Los esquemas de composición propuestos suponen una considerable simplificación de la realidad ya que las relaciones que aparecen suelen ser imprecisas y es probable que no se adapten a ninguna de las dos opciones. Sin embargo la experiencia nos dice que los esquemas anteriores representan bien a una parte importante de las series temporales y en particular la hipótesis multiplicativa es la más habitual en economía, ya que equivale a suponer que la relación entre dos periodos cualesquiera resulta más homogénea en términos relativos que en términos absolutos.
>
> Generalmente las series temporales expresadas en magnitudes monetarias presentarán esquema multiplicativo, mientras el esquema aditivo aparece en series de comportamiento muy estable. En el análisis de series temporales resulta habitual realizar transformaciones logarítmicas, que atenúan las oscilaciones de la serie inicial proporcionando así un esquema aditivo.

La elección del esquema de composición a considerar se efectuará en cada caso según consideraciones prácticas que resultan a menudo del examen gráfico de la serie. Así, el esquema aditivo se corresponde con variaciones de amplitud constante, mientras que en el modelo multiplicativo la amplitud de la variación es cambiante con la tendencia tal y como se recoge en la figura 11.3.

Existen diversos procedimientos para determinar el esquema de composición que resulta más adecuado. Muchos programas informáticos (incluidos los que utilizan las principales oficinas de estadística del mundo), basándose en que la inmensa mayoría de las series económicas tienen un efecto multiplicativo, toman este método por defecto, salvo que la serie contenga algún valor negativo o nulo.

Un procedimiento algo más preciso podría consistir en analizar la dispersión existente entre los valores de la serie a medida que la tendencia va creciendo y asumir una hipótesis aditiva si esta dispersión es más o menos constante, o bien una hipótesis multiplicativa si la dispersión aumenta o disminuye al aumentar la tendencia.

Dado que la tendencia de las series económicas suele ser creciente (la producción va aumentando con los años, los precios suben, al igual que los salarios o la renta disponible, el comercio aumenta, también el turismo, etc.), nos bastará con calcular




El gráfico inferior representa la evolución de las colocaciones en Asturias y muestra una dispersión creciente con la tendencia (esquema de composición multiplicativo) mientras el gráfico superior es su transformada logarítmica y presenta dispersión estable (esquema aditivo).

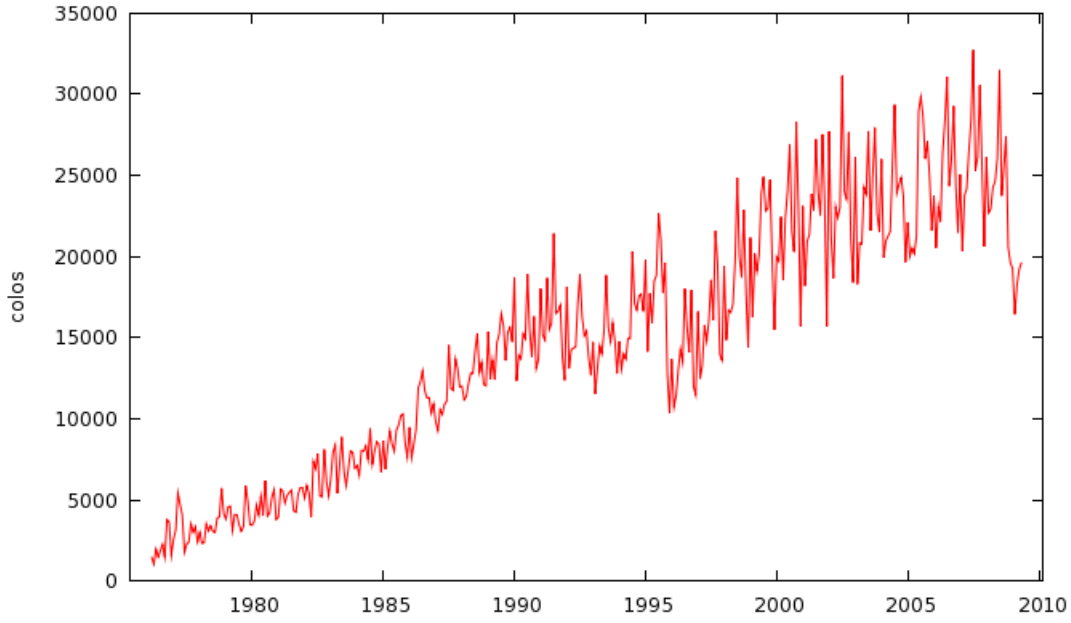



Figura 11.3: Gráficos temporales. Hipótesis aditiva e Hipótesis multiplicativa

la dispersión anual y ver si es aproximadamente constante o no. Así pues, uno de los algoritmos habituales consiste en calcular la media y la desviación típica anuales:

| Años \ Meses | 1 | $\cdots$ | $m$ | $Media_{Anual}$ | $Desv.Tipica_{Anual}$ |
|---|---|---|---|---|---|
| 1 | $Y_{11}$ | $\cdots$ | $Y_{1m}$ | $\bar{Y}_1$ | $S_1$ |
| 2 | $Y_{21}$ | $\cdots$ | $Y_{2m}$ | $\bar{Y}_2$ | $S_2$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| n | $Y_{n1}$ | $\cdots$ | $Y_{nm}$ | $\bar{Y}_n$ | $S_n$ |

Si a la vista de la tabla anterior podemos observar que la desviación típica anual es más o menos constante estaríamos en condiciones de afirmar que la hipótesis es aditiva. La situación contraria es un poco más complicada, puesto que a veces nos puede parecer que las desviaciones típicas van creciendo y sin embargo, cuando procedemos a su representación gráfica podemos no tener esa percepción. En estos casos conviene aplicar algún mecanismo más técnico; por ejemplo, efectuar un ajuste mínimo cuadrático de las desviaciones típicas sobre las medias anuales:

$$\text{Desv.Típica Anual} = b_0 + b_1 MediaAnual$$

y analizar el coeficiente $b_1$. Si la desviación típica es casi constante (hipótesis aditiva), la variación de $S$ por una unidad de variación en la media sería prácticamente nula, y por tanto $b_1 \approx 0$. Si por el contrario $b_1$ es significativamente distinto de cero, entonces la dispersión es proporcional a la tendencia (y por tanto la hipótesis multiplicativa).

Ahora el problema se centra en saber qué es significativamente distinto de cero; en el nivel de este libro todavía no estamos en condiciones de aplicar criterios que nos permitan contrastar esta situación, y consideraremos que si $|b_1| > 0,1$ la hipótesis es multiplicativa y en otro caso aditiva.

**Ejemplo 11.1.** Sobre la serie trimestral de viajeros en establecimientos hoteleros españoles es posible aplicar el método anterior para analizar el tipo de hipótesis. A partir de las medias y desviaciones típicas anuales resumidas en la tabla se obtiene en este caso:

$$\text{Desv.Típica Anual} = 206,96 + 0,25 MediaAnual$$

y al ser el coeficiente $b_1 = 0,25$ se concluye que la dispersión aumenta con la tendencia, es decir, que el tipo de esquema de composición de la serie de viajeros es multiplicativo.

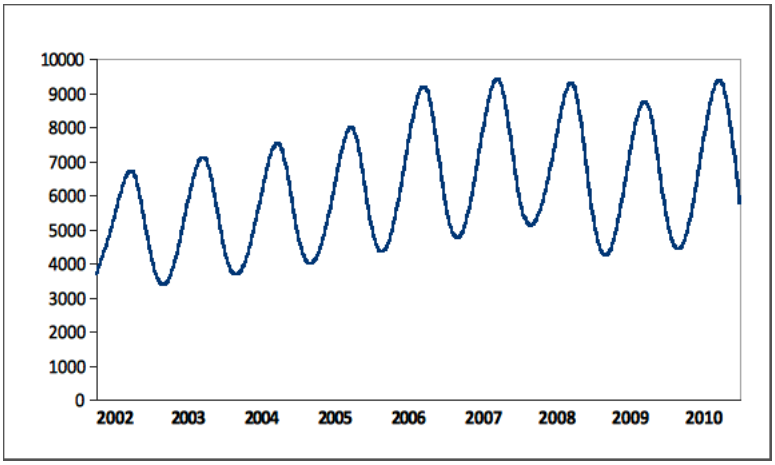



Figura 11.4: Representación gráfica de la serie de miles de viajeros

| Año \ Trimestre | I | II | III | IV | $Media_{Anual}$ | $Desv.Tipica_{Anual}$ |
|---|---|---|---|---|---|---|
| 2002 | 3.713 | 5.419 | 6.672 | 4.153 | 4.989 | 1.335 |
| 2003 | 3.619 | 5.890 | 7.009 | 4.331 | 5.211 | 1.526 |
| 2004 | 3.953 | 6.109 | 7.468 | 4.747 | 5.569 | 1.548 |
| 2005 | 4.218 | 6.409 | 7.931 | 4.986 | 5.886 | 1.638 |
| 2006 | 4.826 | 7.681 | 9.064 | 5.715 | 6.822 | 1.913 |
| 2007 | 5.055 | 7.863 | 9.275 | 5.949 | 7.036 | 1.898 |
| 2008 | 5.395 | 7.682 | 9.178 | 5.411 | 6.917 | 1.851 |
| 2009 | 4.489 | 7.184 | 8.655 | 5.386 | 6.429 | 1.860 |
| 2010 | 4.713 | 7.590 | 9.315 | 5.766 | 6.846 | 2.030 |

## 11.3. Análisis de la tendencia

La tendencia de una serie temporal es en muchos casos el componente de más peso que marca las pautas evolutivas de la variable, y en ocasiones puede enmascarar otras oscilaciones a corto o medio plazo. Por este motivo resulta interesante cuantificar y aislar la tendencia del resto de componentes de la serie temporal.

La tendencia representa la trayectoria general que sigue una serie. Así pues se trata de una línea central a lo largo de la cual se van vertebrando los distintos picos u oscilaciones de la serie.

¿Cómo determinar una tendencia? Existen distintas alternativas: gráficos (una línea más o menos intermedia), algún filtro de los datos que los haga converger hacia el centro de la serie (un alisado o suavizado, una línea de medias condicionadas -línea de regresión-, etc.).

Los métodos gráficos son utilizados con bastante generalidad como primera aproximación a la tendencia de una serie. Los más habituales consisten en trazar una línea

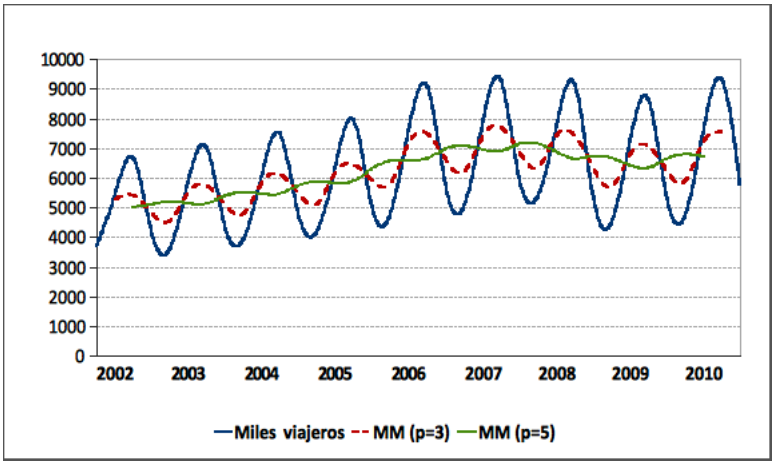



Figura 11.5: Alisados de medias móviles

que suavice el perfil inicial de la serie o la construcción de poligonales que unan puntos máximos y mínimos, acotando así una banda en la que se halla comprendida la tendencia.

### 11.3.1. Método de las medias móviles

**Definición 11.7.** Dada una serie $Y_t$, con valores en distintos periodos de observación $1, 2, ..., T$, denominamos *Media Móvil de periodo* $p$, a la sucesión de valores $MA(t, p)$ obtenidos mediante la expresión:

$$MA(t,p) = \frac{Y_{t-1} + Y_{t-2} + \cdots + Y_{t-p}}{p} \tag{11.3.1}$$

$MA(t, p)$ es un suavizado o filtro de la serie, pues calcula una media de valores y por lo tanto suaviza los valores más altos o bajos. Cuanto mayor sea $p$, más intenso es el efecto de suavizado o filtrado.

A modo de ilustración, este efecto de los alisados puede comprobarse en la figura 11.5, donde se representan la serie de viajeros con sus alisados por medias móviles de orden $p = 3$ y $p = 5$.

Otra forma de interpretar la media móvil es pensar que el valor actual viene influenciado por los $p$ últimos valores de la serie; y cuanto mayor sea $p$, mayor será la dependencia del pasado o la memoria de la serie.

Quizás uno de los puntos llamativos de este procedimiento es que las ponderaciones utilizadas son constantes, y por tanto asignamos la misma capacidad de influencia sobre el valor presente de la serie al último valor o al registrado $p$ meses antes. Estas limitaciones pueden ser superadas utilizando coeficientes de ponderación $w_i,\ i = 1, \cdots, p$ ; $(\sum_{i=1}^{p} w_i = 1)$ función del orden del mes, en cuyo caso el procedimiento se denomina *Medias Móviles Ponderadas*.

¿Qué valor de $p$ debemos tomar? La respuesta no es simple, ya que depende del tipo de serie y del nivel de alisado que queramos obtener; normalmente suelen usarse valores de $p = 5, 7, 9$ o 12, siendo recomendable utilizar un periodo anual cuando se quieren compensar las variaciones que se producen en periodos inferiores al año. En estadísticas oficiales, muchas veces se toman medias móviles de periodo 23 o 25, y así por ejemplo los algoritmos del método X12Arima, utilizado en muchas oficinas de estadística, utilizan estos parámetros.

Este método es fácil de aplicar y nos permite ir construyendo sucesivamente valores futuros de la serie o predicciones, aunque cada vez se encontrarán más y más suavizados convirtiéndose a medio plazo en una línea plana o predicciones constantes.

Sin embargo, este sistema que es muy utilizado en predicciones, puede no ser muy adecuado para analizar el perfil de una serie, donde quizás sea más realista pensar que el valor actual no solo guarda relación con sus valores pasados, sino también con sus valores futuros. Es decir cada valor de la serie se interpola entre los que le rodean y podemos plantear una media móvil que contemple los $\frac{p}{2}$ valores anteriores y los $\frac{p}{2}$ posteriores.

**Definición 11.8.** Dada una serie $Y_t$, con valores en distintos periodos de observación $1, 2, \cdots, T$, denominamos *Media Móvil Centrada* o *Media Móvil Exponencial* de periodo $p$, a la sucesión de valores $MM(t, p)$ obtenidos mediante la expresión:

$$MM(t,p) = \begin{cases} \dfrac{Y_{t-\frac{p}{2}} + \cdots + Y_t + \cdots + Y_{t+\left(\frac{p}{2}-1\right)}}{p} & \text{Si } p\text{es par} \\ \dfrac{Y_{t-\frac{p-1}{2}} + \cdots + Y_t + \cdots + Y_{t+\frac{p-1}{2}}}{p} & \text{Si } p\text{es impar} \end{cases} \tag{11.3.2}$$

Al aplicar esta fórmula el primer valor que podemos asignar es al periodo $\frac{p}{2}+1$, con lo cual no podemos obtener la media móvil centrada para los $\frac{p}{2}$ primeros meses, ni de los $\frac{p}{2}$ últimos. Por este motivo, este procedimiento nos lleva a un suavizado centrado de la serie original, que sería un buen método para el análisis y cuantificación de la tendencia, pero los valores que perdemos en la parte final no nos permiten utilizarlo con fines predictivos.

En la fórmula anterior, tenemos la casuística de si $p$ es par o impar; lo que hacen las medias móviles centradas es asignar el valor resultante a la observación central del periodo. Pero cuando el periodo es par, no existe uno sino dos meses centrales, con lo que podríamos estar sesgando el alisado de la serie (mantendríamos un cierto desfase).

En el caso de que $p$ sea par, una buena solución es aumentar el periodo en un mes y utilizar una Media Móvil Centrada y Ponderada del tipo:

$$MM(t,p+1) = \frac{Y_{t-\frac{p}{2}} + 2Y_{t-\frac{p}{2}+1} + \ldots + 2Y_t + \ldots + 2Y_{t+\frac{p}{2}-1} + Y_{t+\frac{p}{2}}}{2p} \tag{11.3.3}$$

Este procedimiento de cálculo nos garantiza que el promedio resultante se asigna al mes central sin ningún tipo de desfase.

Respecto a la amplitud del periodo son válidas las reflexiones anteriores, considerándose en muchos casos el periodo anual.

También en este caso podemos utilizar coeficientes de ponderación que asignen un peso mayor a los meses más próximos al actual y que vayan descendiendo a medida que nos alejamos.

### 11.3.2. Alisado exponencial

Hemos considerado hasta aquí una ponderación uniforme para la información suministrada por cada observación. Sin embargo, aunque nunca perdamos la memoria del pasado, es habitual considerar que los valores más recientes sean los que tienen una mayor influencia sobre los valores futuros.

**Definición 11.9.** Dada una serie temporal $Y_t$, denominamos *Alisado Exponencial Simple*, que denotamos por $ME_t$, a la serie obtenida como:

$$ME_t = \alpha Y_{t-1} + (1-\alpha)ME_{t-1} \tag{11.3.4}$$

donde $\alpha$ es un coeficiente de alisado, $0 < \alpha < 1$.

Si retrocedemos la función de alisado hacia atrás, sustituyendo en la ecuación 11.3.4, tendremos:

$$\begin{aligned} ME_t &= \alpha Y_{t-1} + (1-\alpha)ME_{t-1} \\ &= \alpha Y_{t-1} + (1-\alpha)\left(\alpha Y_{t-2} + (1-\alpha)ME_{t-2}\right) \\ &= \alpha\left(Y_{t-1} + (1-\alpha)Y_{t-2}\right) + (1-\alpha)^2 ME_{t-2} \\ &= \alpha\left(Y_{t-1} + (1-\alpha)Y_{t-2} + (1-\alpha)^2 Y_{t-3} + \cdots + (1-\alpha)^{t-2}Y_1\right) + (1-\alpha)^{t-1}ME_1 \end{aligned}$$

así pues, necesitamos un valor inicial del alisado que lo tomaremos como el primer valor de la serie: $ME_1 = Y_1$.

- Los coeficientes del alisado forman un sistema de ponderaciones, para lo que debemos comprobar que su suma es la unidad:

  $$\alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \alpha(1-\alpha)^3 + \cdots = \alpha\sum_{i=0}^{\infty}(1-\alpha)^i$$

  esta expresión es la suma de una serie geométrica $\{a_i = \alpha(1-\alpha)^i\}$ de razón $r = (1-\alpha) < 1$, con lo que el valor de la suma viene dado por $\frac{a_0}{1-r}$, en nuestro caso:

  $$\alpha\sum_{i=0}^{\infty}(1-\alpha)^i = \alpha\frac{1}{1-(1-\alpha)} = \alpha\frac{1}{\alpha} = 1$$

- La variable así obtenida se denomina alisada, ya que suaviza o alisa las oscilaciones que tiene la serie, al obtenerse como media ponderada de distintos valores.

| $\alpha$ | t-1 | t-2 | t-3 | t-4 | t-5 | t-6 | t-7 |
|---|---|---|---|---|---|---|---|
| 0,9 | 0,9 | 0,09 | 0,009 | 0,0009 | 0,00009 | 0,000009 | 0,000001 |
| 0,1 | 0,1 | 0,09 | 0,081 | 0,0729 | 0,06561 | 0,059049 | 0,053144 |

Tabla 11.1: Coeficientes de ponderación $w_1, \ldots, w_7$

- Por otra parte, el calificativo de exponencial se debe a que la ponderación o peso de las observaciones decrece exponencialmente a medida que nos alejamos del momento actual $t$, concediendo poca importancia a las observaciones que están alejadas. El alisado será más fuerte o más débil, dependiendo del parámetro $\alpha$.

  La elección del coeficiente de alisado debe acomodarse a cada serie en particular. Debemos tener presente, como se observa en la tabla 11.1, que un valor pequeño de $\alpha$ indica que estamos dando mucho peso a las observaciones pasadas, mientras que si, por el contrario, $\alpha$ es elevado se otorga mayor importancia a los valores recientes.

  Un valor en torno a 0,2 resulta apropiado en muchas series para obtener una línea de tendencia; sin embargo, para obtener un perfil ajustado a la serie, se necesitan valores de $\alpha$ superiores a 0,9.

- Por último, se emplea el calificativo de simple para distinguir este caso de otros en los que una variable se somete a una operación de doble alisado.

  Cuando en algunas series además de la tendencia, se superponen otras componentes, pueden utilizarse métodos de alisados más completos: doble alisado exponencial, alisado de Holt-Winters con tendencia y/o estacionalidad, etc.

### 11.3.3. Método de ajuste lineal

La aproximación de la tendencia puede ser realizada mediante ajuste mínimo cuadrático a la nube de puntos que muestra los valores de $Y$ en función del tiempo $t$.

Aunque la técnica empleada sea coincidente con una regresión de $Y$ sobre $t$, el planteamiento es distinto, ya que no estamos analizando la "capacidad explicativa del tiempo". El supuesto más habitual es el de tendencia lineal, si bien para ciertas magnitudes económicas su propio carácter -o su representación gráfica- aconseja ajustes de tipo parabólico, exponencial, etc.

Ajuste lineal:

$$T_t = b_0 + b_1 t$$

Teniendo en cuenta las expresiones obtenidas en la regresión lineal:

$$b_1 = \frac{S_{Y,t}}{S_t^2} \;\; ; \;\; b_0 = \bar{Y} - b_1 \bar{t}$$

Si la tendencia fuese ajustada mediante un polinomio de grado 2 o superior, se tendría:

- Ajuste parabólico: $T_t = b_0 + b_1 t + b_2 t^2$
- Polinomio de orden $k$: $T_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k$

# 12 Series temporales: estacionalidad y predicción

El movimiento general o tendencia de una variable económica sólo podrá ser apreciado con claridad si conseguimos eliminar las fluctuaciones presentes en la serie que encubren la evolución real del fenómeno. De ahí el interés del análisis de la estacionalidad, que afecta a la casi totalidad de series económicas, resultando incluso frecuente la aparición de varios movimientos estacionales de diferentes amplitudes que se presentan superpuestos y que, en ocasiones, resulta difícil identificar.

El interés de aislar la componente estacional viene justificado por argumentos de diversa índole. Así, por ejemplo, la estacionalidad no incide de igual modo sobre las sucesivas etapas del proceso productivo, originando importantes desfases entre demanda y oferta (en procesos industriales, por ejemplo, aparecen a menudo volúmenes de producción constantes frente a demandas sujetas a oscilaciones estacionales).

Aunque aislar las componentes de una serie temporal es importante para su análisis, el objetivo final del estudio de series temporales es predecir los valores futuros de las series a partir de la experiencia de los valores pasados, tal y como abordaremos en la última parte del tema.

## 12.1. Análisis de la estacionalidad

A menudo la presencia de estacionalidad puede ocultar al observador superficial el verdadero movimiento económico (por ejemplo, las cifras de paro laboral suelen decrecer en los meses de julio y agosto en los países donde el turismo es importante) con el consiguiente riesgo de llegar a conclusiones equivocadas. De hecho, la existencia de una fluctuación estacional puede llegar -en periodos de inestabilidad económica- a transformar la evolución de la magnitud (por ejemplo, dentro de un movimiento inflacionista, un alza estacional de los precios reaviva el movimiento general alcista pudiendo desencadenar una "espiral de inflación").

Las variaciones estacionales pueden no ser únicas, apareciendo varios movimientos estacionales superpuestos. Por ejemplo, si en el sector hostelero analizamos los ingresos totales, se podría apreciar un movimiento estacional de periodo un año (mayores ingresos registrados en los meses veraniegos) superpuesto a otro de carácter semanal (aumento de ventas los fines de semana) e incluso otro diario.

En lo que sigue vamos a suponer que tenemos una serie temporal $Y_{ij}$, donde $i = 1, \ldots, n$ representa el año y $j = 1, \ldots, m$ indica el mes o trimestre del año.

No existe un esquema único para la cuantificación de la componente estacional; uno de los métodos más utilizados para su determinación será el denominado de *razón a*

*la media*, en el que distinguiremos las siguientes etapas:

1. Determinación del componente extraestacional
2. Eliminación del componente extraestacional de la serie
3. Eliminación del componente residual
4. Cuantificación de la variación estacional y en su caso elaborar los índices de variación estacional.

## Cuantificación del componente extraestacional

Para la determinación de $E_{ij}$ son a su vez posibles varios procedimientos alternativos, correspondientes a los métodos ya analizados para describir el movimiento general de una serie: medias móviles y ajuste.

Dado que el componente estacional se produce en periodos inferiores al año, para recoger las variaciones sistemáticas que se producen en periodos superiores (extraestacionales), debemos tomar el año como periodo de referencia para su cuantificación.

### Medias móviles

El número de meses o trimestres que se incluyen en el año es par, y por lo tanto para aplicar el método de las medias móviles corrigiendo el periodo se utiliza la expresión 11.3.3.

Para series mensuales, se suavizaría el valor actual teniendo en cuenta los 6 meses anteriores y los 6 posteriores (en total se utilizan 13 meses en vez de 12). El sistema de ponderación para el cálculo de las medias móviles sería:

$$\frac{1}{24}, \underbrace{\frac{2}{24}, \cdots, \frac{2}{24}}_{\text{11 meses centrales}}, \frac{1}{24}$$

Para series trimestrales utilizaríamos los dos trimestres anteriores y posteriores al actual, con coeficientes de ponderación:

$$\frac{1}{8}, \underbrace{\frac{2}{8}, \frac{2}{8}, \frac{2}{8}}_{\text{3 trimestres centrales}}, \frac{1}{8}$$

### Ajuste lineal

En este caso como la unidad temporal es el año, cada mes $j$ de un año $i$ puede ser representado como: $t = (i-1) + \frac{j}{m}$

Así pues, utilizaremos la expresión del ajuste lineal de la tendencia:

$$E_t = b_0 + b_1 t$$

donde $t$ es el descrito anteriormente: $t = \frac{1}{m}, \cdots, \frac{m}{m}, 1 + \frac{1}{m}, \cdots, (i-1) + \frac{m-1}{m}, i$

## Eliminación del componente extraestacional

Una vez obtenidos -por cualquiera de los métodos expuestos- los valores $E_{ij}$, su eliminación de la serie inicial será por diferencia o por cociente, dependiendo del tipo de hipótesis de composición:

### Hipótesis aditiva

En este caso, partiendo de la expresión 11.2.1, se tiene:

$$Y_{ij} = E_{ij} + e_{ij} + u_{ij}$$

en cuyo caso, la serie:

$$Y_{ij} - E_{ij} = e_{ij} + u_{ij}$$

no tiene componente extraestacional, y conseguimos nuestro objetivo. La serie resultante contiene la variación estacional ("componente bruto de variación estacional"), cuyos valores se encuentran «contaminados» por el componente accidental o residual.

### Hipótesis multiplicativa

Partiendo de la ecuación 11.2.2, podemos expresarla como:

$$Y_{ij} = E_{ij} \times Ie_{ij} + u_{ij}$$

dividiendo por la componente extraestacional, se tiene:

$$\frac{Y_{ij}}{E_{ij}} = Ie_{ij} + \frac{u_{ij}}{E_{ij}}$$

En el segundo miembro tenemos el índice de variación estacional y un residuo aminorado, porque hemos relativizado ese efecto accidental respecto a la línea general de la serie; por lo tanto se trataría de un nuevo residuo $v_{ij}$, con las mismas propiedades, pero con menor magnitud. Así pues, como en el caso aditivo, la serie obtenida como cociente, solo incluye la componente estacional (expresada como índice) y el residuo.

## Eliminación del componente residual

La variación residual puede introducir distorsiones en la serie, pero no tiene un patrón de comportamiento con lo cual esperamos que la suma o la media de los residuos a lo largo de un periodo de tiempo tengan un impacto nulo.

Denotamos por $YC_{ij}$ la serie corregida del componente extraestacional, obtenida anteriormente:

$$\begin{cases} YC_{ij} = e_{ij} + u_{ij} \\ \\ YC_{ij} = Ie_{ij} + v_{ij} \end{cases}$$

Si calculamos la media por años, para cualquier mes $j = 1, \ldots, m$, y teniendo en cuenta el supuesto anterior sobre la media de los residuos, se tiene:

$$\begin{cases} \sum_{i=1}^{n} \frac{YC_{ij}}{n} = \sum_{i=1}^{n} \frac{e_{ij}}{n} + \overbrace{\sum_{i=1}^{n} \frac{u_{ij}}{n}}^{=0} = \bar{e}_{.j} \\ \sum_{i=1}^{n} \frac{YC_{ij}}{n} = \sum_{i=1}^{n} \frac{Ie_{ij}}{n} + \underbrace{\sum_{i=1}^{n} \frac{v_{ij}}{n}}_{=0} = \bar{Ie}_{.j} \end{cases}$$

En el caso de esquema aditivo la nueva serie solo incluye la variación estacional media del mes $j$, y en el caso multiplicativo el índice medio de variación estacional del mes $j$, habiendo eliminado ya el componente residual.

## Cuantificación de la estacionalidad

La serie obtenida anteriormente nos proporciona el componente estacional, pero posiblemente no se encuentre normalizado para que tenga una interpretación intuitiva.

Por este motivo, es necesario efectuar un último ajuste, para garantizar que la media del componente estacional sea nula en el caso aditivo, o bien la media de los índices de variación estacional sea unitaria en el caso multiplicativo. Para conseguir este objetivo es necesario llevar a cabo una última transformación consistente en comparar el componente estacional de cada mes con su valor medio del año (mediante desviaciones en el caso aditivo y por cociente en el multiplicativo):

| | **Esquema Aditivo** | | **Esquema Multiplicativo** | |
|---|---|---|---|---|
| Meses | Comp.Estacional | $Ve_j$ | Comp.Estacional | $IVE_j$ |
| 1 | $\bar{e}_{.1}$ | $\bar{e}_{.1} - \bar{e}_{..}$ | $\bar{Ie}_{.1}$ | $\frac{\bar{Ie}_{.1}}{\bar{Ie}_{..}}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| m | $\bar{e}_{.m}$ | $\bar{e}_{.m} - \bar{e}_{..}$ | $\bar{Ie}_{.m}$ | $\frac{\bar{Ie}_{.m}}{\bar{Ie}_{..}}$ |
| Media | $\bar{e}_{..}$ | 0 | $\bar{Ie}_{..}$ | 1 |

Observamos que cuando la hipótesis es aditiva, los componentes estacionales $Ve_j$ son valores expresados en las mismas unidades que la magnitud estudiada, que recogen saldos positivos si la estacionalidad genera aumentos respecto a la tendencia o negativos en caso contrario.

Por el contrario, cuando la hipótesis es multiplicativa, la definición de los índices de variación estacional permite una interpretación de los mismos en términos de números índices, cuantificando la "proporción o porcentaje en el que la estacionalidad actúa

sobre las observaciones, haciendo que éstas se desvíen (por exceso o por defecto) de su tendencia".

> El método expuesto resulta adecuado bajo el supuesto de estacionalidad estable, esto es, si admitimos que los coeficientes de estacionalidad no se ven afectados por ninguna tendencia. Podría ocurrir, sin embargo, que el efecto de la estacionalidad variase a lo largo del tiempo (por ejemplo, las diferencias estacionales de precios hosteleros podrían irse atenuando progresivamente) resultando en este caso necesario modificar en cierta medida el cálculo de índices.

**Ejemplo 12.1.** La aplicación de este procedimiento sobre la serie de viajeros en establecimientos hoteleros españoles permitiría cuantificar el componente estacional de esta serie que, al presentar esquema de composición multiplicativo, se calcularía mediante los índices de variación estacional.

Tal y como hemos descrito anteriormente, el primer paso sería la aproximación del componente extraestacional mediante el procedimiento de medias móviles o de ajuste a la tendencia. En el segundo caso, la recta de ajuste anual vendría dada por la expresión:

$$E_t = 5.083,83 + 221,09t$$

y sustituyendo $t$ por los valores; $t = \frac{1}{4}, \frac{2}{4} \cdots, 9$ se llegaría a los resultados de la tabla:

| Año \ Trimestre | I | II | III | IV |
|---|---|---|---|---|
| 2002 | 5.139 | 5.194 | 5.250 | 5.305 |
| 2003 | 5.360 | 5.415 | 5.471 | 5.526 |
| 2004 | 5.581 | 5.637 | 5.692 | 5.747 |
| 2005 | 5.802 | 5.858 | 5.913 | 5.968 |
| 2006 | 6.023 | 6.079 | 6.134 | 6.189 |
| 2007 | 6.245 | 6.300 | 6.355 | 6.410 |
| 2008 | 6.466 | 6.521 | 6.576 | 6.631 |
| 2009 | 6.687 | 6.742 | 6.797 | 6.853 |
| 2010 | 6.908 | 6.963 | 7.018 | 7.074 |

A continuación sería necesario dividir la serie inicial entre este componente extraestacional, obteniendo los resultados que siguen:

| Año \ Trimestre | I | II | III | IV |
|---|---|---|---|---|
| 2002 | 0,72 | 1,04 | 1,27 | 0,78 |
| 2003 | 0,68 | 1,09 | 1,28 | 0,78 |
| 2004 | 0,71 | 1,08 | 1,31 | 0,83 |
| 2005 | 0,73 | 1,09 | 1,34 | 0,84 |
| 2006 | 0,80 | 1,26 | 1,48 | 0,92 |
| 2007 | 0,81 | 1,25 | 1,46 | 0,93 |
| 2008 | 0,83 | 1,18 | 1,40 | 0,82 |
| 2009 | 0,67 | 1,07 | 1,27 | 0,79 |
| 2010 | 0,68 | 1,09 | 1,33 | 0,82 |
| Media | 0,74 | 1,13 | 1,35 | 0,83 |
| IVEj | 0,73 | 1,12 | 1,33 | 0,82 |

A partir de dichos resultados podemos afirmar que el componente estacional actúa a la baja en el primer trimestre, ya que el número de viajeros en este periodo es un 73 % del valor tendencial (por tanto se reduce un 27 %), mientras en el segundo trimestre la estacionalidad se manifiesta al alza aumentando las entradas de viajeros en un 12 %. Este efecto de estacionalidad al alza es todavía más acentuado en el tercer trimestre, ya que en este periodo los viajeros aumentan un 33 % respecto a la tendencia. En cambio, en el último trimestre de cada año se observa una reducción estacional de los viajeros, que se sitúan en el 82 % de su nivel tendencial (lo cual supone una disminución del 18 %).

## 12.2. Desestacionalización

Inicialmente disponíamos de una serie temporal $Y_{ij}$, y a lo largo de los epígrafes anteriores, hemos separado los componentes de tendencia o extraestacional y el componente estacional.

En muchos casos, las variaciones estacionales actúan como agentes distorsionadores que dificultan el análisis de la evolución de la serie. Así, por ejemplo, si hablamos de la evolución de los ocupados, existen determinados meses en los que la ocupación aumenta o disminuye por un efecto estacional, y como consecuencia la señal de aumento o descenso en los ocupados, puede aparecer distorsionada. Lo mismo sucedería si analizamos series sobre movimiento de pasajeros, ingresos por turismo, consumo, producción agraria, etc.

En estos casos nos va a interesar disponer de una serie temporal limpia de variaciones estacionales y que nos permita una interpretación directa de las variaciones de la serie. Lógicamente, para eliminar la componente estacional debemos tener en cuenta el tipo de composición de la serie:

- Aditiva:
$$Y_{ij} - Ve_j = E_{ij} + u_{ij}$$
- Multiplicativa:
$$\frac{Y_{ij}}{IVE_j} = E_{ij} + v_{ij}$$

Podemos observar que el procedimiento descrito consiste en eliminar del valor inicial de la serie correspondiente al subperiodo $j$ del año $i$, la estacionalidad correspondiente al subperiodo $j$. Esta eliminación se lleva a cabo por diferencia o cociente, según que el esquema de la serie sea aditivo o multiplicativo.

Los valores resultantes de una serie desestacionalizada aparecerán corregidos al alza o la baja según la estacionalidad tenga en ese mes una influencia negativa o positiva respectivamente.

> Como hemos comentado, podría suceder que los datos estacionales se viesen afectados de cierta tendencia, situación que conduciría a índices de variación estacional diferentes

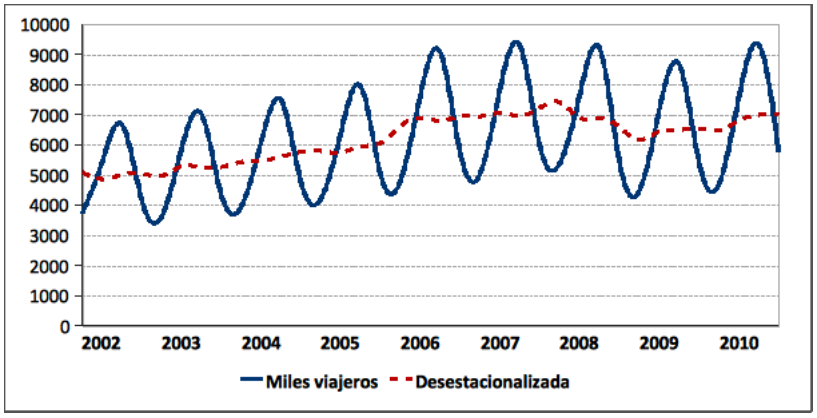



Figura 12.1: Serie desestacionalizada

en cada periodo. Como consecuencia, en la desestacionalización le correspondería a cada dato de la serie original un índice de variación estacional distinto.

Para resolver este problema han sido desarrollados varios procedimientos informáticos, siendo el más conocido el Census Method 12, versión X-l2, desarrollado por Julius Shiskin para el US Bureau of the Census y ampliamente utilizado en todo el mundo.

**Ejemplo 12.2.** La desestacionalización de la serie de viajeros estudiada podría ser llevada a cabo mediante este procedimiento. Así, por ejemplo, para el primer trimestre del año 2002 dividiríamos el valor observado de viajeros en ese periodo (3.713) entre el correspondiente índice de variación estacional (0,73), obteniendo el resultado 5.098, que como podemos observar es superior al inicial ya que en este caso hemos eliminado la estacionalidad que actuaba a la baja en el primer trimestre. De modo análogo, en los restantes trimestres del mismo año la serie desestacionalizada proporcionaría los resultados:

$$\frac{Y_{2002,II}}{IVE_{II}} = \frac{5419}{1,12} = 4859; \quad \frac{Y_{2002,III}}{IVE_{III}} = \frac{6672}{1,33} = 5005; \quad \frac{Y_{2002,IV}}{IVE_{IV}} = \frac{4153}{0,82} = 5044$$

Como consecuencia de este proceso de desestacionalización se obtienen series con menos oscilaciones que la inicial, tal y como podemos observar en la figura 12.1.

## 12.3. Predicción

El principal objetivo del estudio de las series temporales es realizar predicciones para anticipar valores futuros de variables que han sido observadas de forma temporal en el pasado.

Las predicciones con métodos descriptivos (que son los estudiados en este libro), se vienen utilizando desde hace aproximadamente un siglo, ya que se vivió un periodo de esplendor de estas técnicas durante el periodo 1902-1950. Estos métodos se pusieron en duda cuando no fueron capaces de predecir la crisis de 1929, y como consecuencia se

realizó una crítica importante a las técnicas de series temporales, porque no estaban basadas en la teoría económica; es decir, no había hipótesis teóricas que justificasen su comportamiento.

A partir de la década de 1950 se fueron incorporando nuevas técnicas de predicción más o menos revolucionarias que se suponían capaces de adelantar los periodos de crisis y expansión, pero no lograron sus objetivos. Posteriormente, en el periodo 1950-1970 la investigación se centró en la "Econometría Aplicada", que introducida por el premio Nobel de Economía Lawrence R. Klein (1920-), se sigue utilizando en la actualidad para explicar y predecir mediante análisis causal el comportamiento de distintas economías.

Estas técnicas no fueron capaces de explicar la crisis de la década de 1970 y aparece una nueva metodología para el tratamiento de series temporales, denominada Box-Jenkins en honor de sus autores, Georges E.P. Box (1919-) y Gwilym Meirion Jenkins (1933–1982). Esta técnica se aplicó durante mucho tiempo y sigue considerándose como un método de predicción muy eficiente a corto plazo, pero tampoco esta metodología fue capaz de adelantar la crisis de la década de 1990.

Más recientemente se desarrollaron otras teorías, basadas en métodos no lineales, distribuciones apriori, números difusos, redes neuronales, etc. Lamentablemente ninguno de éstos métodos fue capaz de predecir la crisis de 2007 y sin duda, en los próximos años se desarrollarán nuevas técnicas de predicción que deberán ser contrastadas en las próximas crisis.

La predicción basada en el análisis clásico, que es el estudiado en este tema, presupone que una serie temporal se comportará en el futuro de modo análogo al pasado y aprovechará la experiencia de los datos para predecir los valores futuros, ceteris paribus el resto de causas que pueden influir en la series. Así pues, aun asumiendo la crítica de que el análisis clásico de las series temporales no está basado en ninguna teoría económica o hipótesis de comportamiento, esta técnica puede resultar de gran utilidad.

Resulta conveniente tener presente el riesgo inherente en toda predicción temporal, que aparece relacionado con el periodo de tiempo analizado. Cuanto más amplia sea la experiencia (o sea, mayor el número de datos disponibles), más fácil será determinar las regularidades estadísticas de la serie y por tanto mejor sería nuestra predicción. Así, aunque los algoritmos que usamos en el análisis clásico de series temporales no exigen un número mínimo de datos, conviene disponer de un tamaño suficientemente elevado para que los filtros tengan sentido y repartan las variaciones irregulares en los distintos años.

En términos generales, para obtener una tendencia se recomendaría disponer de al menos 7 o 10 años; al aplicar medias móviles debemos tener en cuenta que perdemos un año, y si queremos distribuir un efecto accidental parece razonable disponer de al menos 6 años. Así pues orientativamente resulta conveniente disponer en series mensuales de al menos 70 datos para el análisis o la predicción con métodos descriptivos de las series temporales.

Otra consideración interesante es la referida al objetivo de la predicción, ya que podemos estar interesados en anticipar el valor exacto de una serie en determinado mes de cierto año, o bien en aproximar el componente tendencial e incluso extraestacional de dicha serie.

**Ejemplo 12.3.** Supongamos que nuestro objetivo es predecir una serie temporal de

carácter socioeconómico como el número de ocupados o parados en España. Dentro de este ámbito podemos tener diversos objetivos, y así podríamos estar interesados en anticipar el número de parados que se registrarán en el mes de mayo del próximo año (es decir, el valor $Y_{T+1,5}$) o bien el valor tendencial de esta variable, que vendría aproximado por el componente de tendencia, el extraestacional o incluso la serie desestacionalizada.

Los ejemplos anteriores muestran distintos objetivos y como consecuencia podríamos contemplar diferentes métodos de predicción.

**Predicciones sobre la tendencia:** Si analizamos una serie $Y_t$, y queremos obtener una predicción tendencial para el periodo $T + 2$, existen distintas alternativas.

- Si la serie muestra una trayectoria lineal clara, el método del ajuste lineal de la tendencia puede resultar el método más adecuado:

$$T_{T+2} = b_0 + b_1(T + 2)$$

- Si el perfil de la serie es más irregular, los métodos más adecuados para hacer predicciones son el alisado exponencial y el de medias móviles, donde se predice el valor futuro en función de sus valores pasados. El alisado exponencial que lleva ponderaciones variables, ponderando más los valores más próximos y menos los más alejados, sería nuestra primera propuesta. Además este método tiene otras dos ventajas: que se encuentra implementado en la mayor parte de las hojas de cálculo, y que existen variantes de este algoritmo que permiten perfeccionar el método incluyendo tendencia o componente estacional:

$$ME_{T+1} = \alpha Y_T + (1 - \alpha)ME_T$$

  El procedimiento garantiza una buena predicción para el periodo $T + 1$; para valores superiores seguimos manteniendo el registro del periodo $T$ y actualizamos solo la parte de la predicción:

$$ME_{T+2} = \alpha Y_T + (1 - \alpha)ME_{T+1}$$

  Sabemos que cuanto más bajo sea $\alpha$ más suave es la línea y por lo tanto tendremos una predicción más centrada y menos ajustada al valor puntual de un mes concreto. Valores $0,2 \leq \alpha \leq 0,3$ pueden conseguir una buena aproximación a la tendencia-ciclo. La predicción con medias móviles es menos habitual, aunque es muy fácil de aplicar o implementar en una hoja de cálculo:

$$MA_{T+1} = \frac{Y_T + Y_{T-1} + \ldots + Y_{T-p+1}}{p}$$

$$MA_{T+2} = \frac{MA_{T+1} + Y_T + \ldots + Y_{T-p}}{p}$$

  valores altos de $p$ (superiores al periodo anual), consiguen predicciones tendenciales aceptables.

**Predicciones sobre la serie desestacionalizada:** La serie desestacionalizada, como hemos visto antes, se obtiene al eliminar de la serie original el componente estacional. Esta serie tiene cierto paralelismo con la serie de tendencia o tendencia-ciclo, y existe abundante literatura sobra la conveniencia de usar una u otra serie en determinados análisis como la extracción de señales.

Sin embargo, existe una cierta diferencia conceptual entre ambas series: la tendencia-ciclo realiza un filtrado de los componentes estacional y residual, de forma que en la serie resultante sus efectos se habrán diluido sobre la tendencia; sin embargo, en la serie desestacionalizada se ha eliminado el componente estacional y se ha suavizado algo el residuo, pero probablemente la serie contiene una buena parte de este último componente. Así pues, cuando hacemos predicciones sobre la serie desestacionalizada, obtendremos valores más ajustados a los observados que los que se obtienen mediante la serie de tendencia.

**Predicciones sobre valores futuros de la serie original:** A la vista de las consideraciones anteriores, podemos contemplar básicamente dos alternativas para predecir los valores futuros de la serie:

- Aprovechar los métodos de predicción de tendencia, con valores altos del parámetro de alisado $\alpha$ ($0,9 \leq \alpha \leq 0,999$), en algunos casos los programas permiten estimar el $\alpha$ óptimo, que ofrecen una predicción aceptable de los valores observados; o bien con valores bajos del periodo de la media móvil (aunque suele conducir a mejores predicciones el método del alisado).

- Una segunda alternativa es hacer predicciones sobre la serie desestacionalizada y posteriormente actualizar las predicciones incorporándoles el componente estacional, para el que asumimos estabilidad en el futuro. Esta alternativa tiene la ventaja de que podemos utilizar el método del ajuste lineal en la predicción de la serie desestacionalizada (porque suele tener un perfil suave). Sin embargo, tiene el inconveniente de que las predicciones pueden tener mayor incertidumbre puesto que repetimos varias veces el proceso de filtrado.

Cabe además señalar que, tal y como hemos comentado al comienzo del tema anterior, en muchas series temporales la existencia de valores extraños o atípicos, el efecto calendario, etc. condicionan en alguna medida los datos y por tanto sus predicciones.

Así pues el primer paso para mejorar las predicciones será depurar bien los datos; para ello consideraremos la serie tipificada $Z_t$ y llamaremos valores atípicos a aquellos valores para los que $|Z_t| > k$, con $k = 3$ o $k = 4$.

Una vez localizados los valores atípicos, podemos proceder a eliminarlos; para ello comenzaremos por el valor atípico de mayor magnitud (que asumimos corresponde al periodo $t^*$) y aplicaremos algún tipo de filtrado, por ejemplo asignando a ese valor la media de la serie, o la media del año, o la de los dos valores consecutivos, o realizando una interpolación. También es posible llevar a cabo una regresión, en la que se

introduce una variable dicotómica:

$$D_t = \begin{cases} 1 & \text{si } t = t^* \\ 0 & \text{en el resto} \end{cases}$$

Entonces realizamos la regresión de $Y_t$ sobre $D_t$, y obtenemos así el impacto del valor atípico:

$$Y_t = b_0 + b_1 D_t$$

Si ahora eliminamos de la serie original esta serie estimada obtendremos una nueva serie que ya no tiene efectos de la influencia del valor atípico.

Este proceso de eliminación de valores atípicos podría repetirse tantas veces como se considere necesario. Es decir, sobre la serie anterior volvemos a llevar a cabo la tipificación y comprobamos si existen valores atípicos. En caso afirmativo podemos empezar por el que tiene mayor peso y repetir la operación. El objetivo final es conseguir una serie limpia de valores atípicos y que por tanto resulta más adecuada para realizar predicciones.

### Evaluación de las predicciones

Dado que no existe un método que sea el ideal para utilizar con cualquier serie, sería conveniente evaluar la capacidad predictiva de los distintos procedimientos. Para ello vamos a comparar las predicciones con los datos registrados y comprobar el margen de error que comete nuestro método.

Denotaremos el error de predicción por $e_t$ y como la suma de los errores tiende a compensarse debido a su distinto signo, consideraremos medidas basadas en sus valores absolutos o sus cuadrados:

- Error absoluto medio

$$EAM = \frac{\sum_{t=1}^{T} |e_t|}{T}$$

- Error absoluto porcentual medio ( %)

$$EAPM = \frac{1}{T} \sum_{t=1}^{T} \frac{|e_t|}{Y_t} \cdot 100$$

- Raíz del error cuadrático medio

$$RECM = \sqrt{\frac{1}{T} \sum_{t=1}^{T} e_t^2}$$

- Raíz del error cuadrático porcentual medio

$$RECPM = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \frac{e_t^2}{Y_t^2}}$$

Así pues, cuando realicemos predicciones por varios métodos, podemos calcular las medidas anteriores, y adoptar como más adecuado aquel método que proporcione unos resultados más reducidos de estas medidas, que indican una mayor adecuación de las predicciones.

Nuestro objetivo es conseguir predicciones de la mejor calidad posible, para lo cual como hemos visto resulta conveniente eliminar los valores atípicos de la serie original, y seleccionar el método de predicción que en cada caso se considere más adecuado. Así en algunas ocasiones los métodos lineales resultan excesivamente simplistas, resultando más adecuado aplicar modelos de regresión más ajustados a los datos: regresión parabólica, logarítmica, etc.

También podríamos considerar algoritmos más complejos que tuvieran en cuenta no solo una ecuación de alisado como la que hemos visto, sino con dos o tres ecuaciones que sean capaces de recoger el efecto de distintos componentes.

Evidentemente, estos y otros métodos más complejos exceden el nivel de este libro, pero su aplicación será necesaria cuando deseemos ir mejorando nuestra capacidad predictiva.

# Bibliografía


[1] F. Arnaldos, M.T. Díaz, U. Faura, L. Molera, I. Parra. *Estadística Descriptiva para Economía y Administración de Empresas.* AC, 2003.

[2] G. Calot. *Curso de Estadística Descriptiva.* Paraninfo, Madrid, 1974.

[3] The Economist. *Guide to Economic Indicators: Making Sense of Economics.* 2007.

[4] I. Fisher. *The Making of Index Numbers.* Houghton Mifflin Company, 1922.

[5] FMI. *Manual del Índice de Precios al Consumidor: Teoría y Práctica.* 2006.

[6] A. García Barbancho. *Estadística Elemental Moderna.* Ariel, Barcelona, 1973.

[7] M.P. Martín-Guzmán,, F.J. Martín Pliego. *Curso Básico de Estadística Económica.* AC, Madrid, 1985.

[8] P. Martín-Guzmán, I. Toledo, F.J. López Ortega, N. Bellido. *Manual de Estadística Descriptiva.* Thomson Civitas, 2006.

[9] F.J. Martín-Pliego. *Introducción a la Estadística Económica y Empresarial. Teoría y Práctica.* Thomson, 2004.

[10] J.M. Montero. *Estadística para Relaciones Laborales.* AC, Madrid, 2000.

[11] U. Nieto de Alba. *Introducción a la Estadística.* Aguilar, Madrid, 1975.

[12] A. Novales. *Estadística y Econometría.* McGraw-Hill, 1996.

[13] R. Pérez. *Nociones Básicas de Estadística.* [en línea] <https://sites.google.com/a/uniovi.es/libros/nociones-basicas-estadistica>, 2010.

[14] R. Pérez, A.J. López, M.J. Río M.J., N. Muñoz, C. Caso, M. Alvargonzález, J.B. García. *Análisis de datos económicos I. Métodos descriptivos.* Pirámide, Madrid, 1997.

[15] A. Pulido. *Estadística y Técnicas de Investigación Social.* Pirámide, Madrid, 1976.

[16] J.L. Sánchez-Crespo, E. García España. *Estadística Descriptiva.* INE, Madrid, 1961.

# Índice alfabético