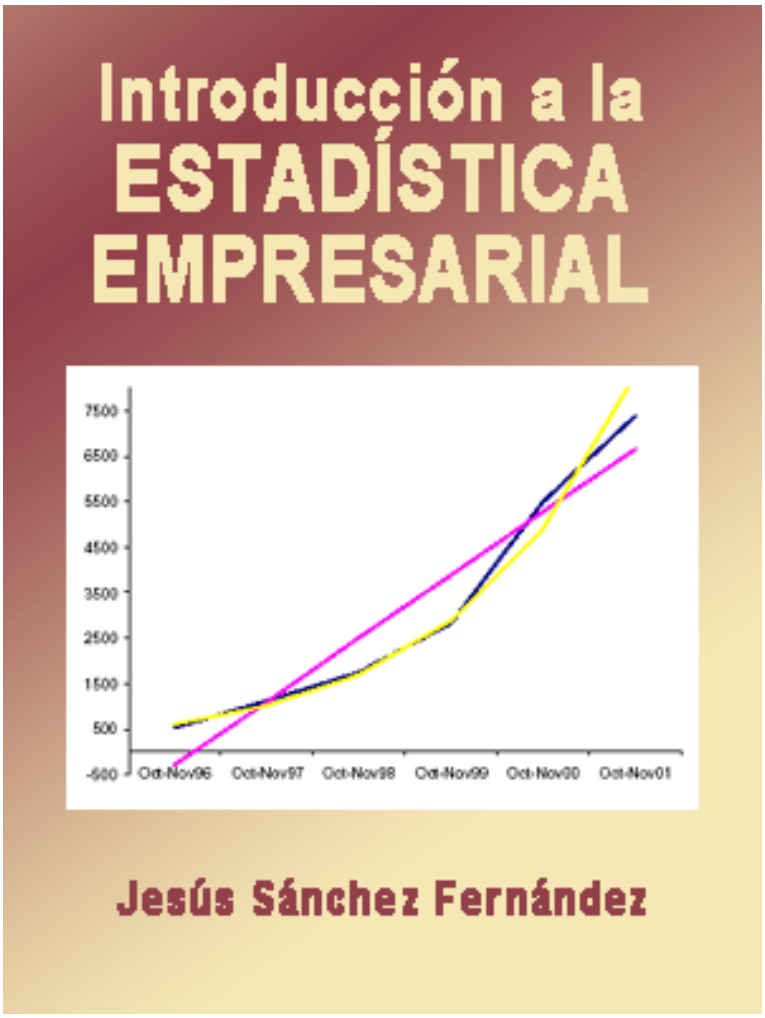


Puede enviar sus comentarios al libro directamente al autor:
j_sanchez10@terra.es

Para citar este libro puede utilizar el siguiente formato:

Sánchez Fernández, J. (2004) Introducción a la Estadística Empresarial
Edición electrónica en http://www.eumed.net/cursecon/libreria/index.htm

editado por eumed.net

CAPITULO 1.- ESTADÍSTICA: CIENCIA Y DATOS

1.1 Introducción.

En este manual se exponen un conjunto de instrumentos de análisis estadístico cuya finalidad es ayudar a manejar, de una forma cómoda y útil, la cada vez mayor cantidad de información de tipo cuantitativo, e incluso cualitativo, que cualquier agente económico tiene en sus manos. Pero esos instrumentos son de utilidad no solo para los agentes económicos (consumidores y productores). En realidad las técnicas estadísticas también son de tanta utilidad o más en otros ámbitos que nada tienen que ver con la economía.

La gran aportación de la estadística es, precisamente, ese arsenal de instrumentos y técnicas que permiten tratar y sintetizar esa gran cantidad de información, en un intento de buscar las posibles regularidades que la misma esconde detrás de la enorme variabilidad con la que se presenta. El objetivo último de ese tratamiento estadístico de la información es reducir, en la medida que ello sea posible, la incertidumbre inherente a la variabilidad de la información, para que la toma de decisiones, de cualquier agente (económico o de otra naturaleza), se lleve a cabo con el menor grado de incertidumbre posible.

Una vez que se han señalado estas ideas, con las que se ha pretendido poner de manifiesto no ya la utilidad de los métodos estadísticos, que eso quedará a juicio de cada usuario, sino la cotidianidad con la que nos vemos obligados la gran mayoría de la población a trabajar y tomar decisiones basadas en información de tipo estadístico, cabría preguntarse si es necesario ampliar el contenido de este capitulo introductorio si, como se ha señalado en otro lugar, la mayoría de los lectores se lo saltan.

Sin embargo, y aún a riesgo de que se cumpla esa afirmación, es conveniente dejar claras dos cuestiones relevantes. La primera de ellas es que no debe sacarse la falsa idea de entender la estadística como una mera colección de métodos o técnicas útiles para el tratamiento de la información o, incluso lo que es más, concluir que la estadística es lo que hacen los estadísticos. Aunque esas dos ideas no son desacertadas, tampoco permiten tener una visión completa de lo que es la estadística. La segunda es que nuestras decisiones se basan, cada vez más, en un flujo creciente de información que necesitamos sintetizar para evitar aquello de los árboles impidan

ver el bosque. Nuestras decisiones son de tipo condicionado, pues la mismas se toman en función de algún tipo de información, tanto pasada como presente.

1.2 Concepto de Estadística.

Es frecuente que la Estadística se identifique con una tabla o colección de datos. De hecho, eso es una estadística. Pero que duda cabe que la Estadística no debe entenderse como una mera colección de datos, aunque los mismos se presenten de forma ordenada y sistemática.

Esta forma de entender la Estadística tiene su origen en el significado etimológico del término. La palabra Estadística deriva de la latina "status" y se remonta a los tiempos en los que los estados-naciones recababan datos, especialmente sobre renta y población, a efectos de recaudación impuestos y mantenimiento del ejercito. Esos datos se identificaban con el estado, razón por la cual terminaron conociéndose como estadísticas. En este sentido, la Estadística es tan antigua casi como el propio ser humano. Pero esta es una forma muy estrecha de entender y definir la Estadística.

En cambio, la Estadística entendida como ciencia tiene un origen más reciente y el gran desarrollo de la misma ha tenido lugar, fundamentalmente, a lo largo del siglo XX. Como ciencia, *la Estadística está formada por el conjunto de métodos y técnicas que permiten la obtención, organización, síntesis, descripción e interpretación de los datos para la toma de decisiones en ambiente de incertidumbre*. Ese objetivo que persigue la Estadística con la organización y síntesis de los datos tiene su razón de ser en el hecho de que la misma se preocupa del estudio de los que podemos denominar como fenómenos de masas. Es decir, la Estadística no está interesada en el estudio de datos aislados, pues si la información es escasa no tiene sentido plantearse problemas de organización ni de síntesis. Así, si se estudian los gastos en publicidad de las empresas de una determinada rama de actividad y se tiene información para solo dos empresas, entonces, con esos dos datos no ha lugar plantearse si los mismos han de presentarse mediante una tabla o un gráfico o si deben resumirse mediante un promedio. Esa escasez de información no debiera ser nunca objeto de análisis estadístico, pues la descripción de la misma es irrelevante y a partir de ella poco se puede decir en relación con los gastos en publicidad de todas la empresas. La metodología estadística adquiere entidad cuando de lo que se trata es de analizar un elevado volumen de datos, pues por lo general, tras esa "masa de datos" se esconden

ciertas regularidades o leyes de comportamiento que nos permitirán, una vez descritas, tomar decisiones en ambiente de incertidumbre, siempre que esta pueda cuantificarse en términos de probabilidad, pues esas decisiones se basan en una leyes que, a diferencia de las leyes de la física, no son exactas sino que están sujetas a errores.

En el párrafo anterior han aparecido, no por casualidad, tres términos que nos van a permitir desdoblar a la Estadística en dos ramas principales. Esos términos son: describir, probabilidad y toma de decisiones con incertidumbre. El primero de ellos da lugar a lo que se conoce como Estadística Descriptiva. Esta rama es la más antigua de la Estadística y su objeto es el análisis de los datos para descubrir o describir las posibles regularidades que presenten. Paralelamente, aunque con posterioridad, se desarrolló la Teoría de la Probabilidad. La unión ambas ha dado lugar a lo que se conoce como Estadística Matemática o Inferencia Estadística. El instrumental propio de esta rama es el que nos permite tomar decisiones en ambiente de incertidumbre. Se trata de decisiones basadas en la información que suministran los datos y que permiten generalizar los resultados obtenidos.

1.3 La observación estadística.

En el apartado anterior se ha intentado dejar claro que la Estadística se dedica al estudio de los fenómenos de masas. Es decir, la Estadística centra su interés en la observación de colectivos amplios de entes o elementos, los cuales pueden ser personas o cosas. A esos colectivos se les denomina en Estadística como Población.

Ahora bien, una vez que se tiene claro que el objeto de la Estadística es la observación y estudio de las poblaciones, la siguiente cuestión que puede plantearse es como ha de realizarse esa observación. La misma puede se exhaustiva o parcial. Las dos formas tienen ventajas e inconvenientes. En el caso de la observación exhaustiva o total, y si se asume que no hay errores de medida entonces, lo que se consigue es eliminar la incertidumbre.

Frente a esa ventaja fundamental, la observación exhaustiva tiene un grave inconveniente: el coste. Se trata tanto de un coste monetario como en tiempo. Imaginemos la siguiente situación. Un partido político, antes de unas elecciones, quiere saber cual es la intención de voto del electorado. Para ello tiene dos opciones. Preguntarle a todos los electores o solo a un subconjunto de los mismos. En el primer

caso estamos frente a una observación exhaustiva y está claro que, ante esta situación, para el partido que lleva a delante la observación no habrá incertidumbre alguna en relación con el resultado final de las elecciones, siempre y cuando no haya errores de medida. En este caso esos errores vendrán dados por la falta de sinceridad en la respuesta de los electores y por las no respuestas, además de otras posibles causas. Pero, ¿porqué ese partido, o cualquier otro, normalmente no realiza ese tipo de observación?. En este caso la respuesta parece trivial. El tamaño de la población es demasiado grande, lo que conlleva un coste tanto en tiempo como en dinero que hace desaconsejable esa opción. Un ejemplo de una operación estadística de carácter exhaustivo, dentro de la estadística oficial, son los censos de población, especialmente los realizados desde 1991, pues los anteriores eran una mezcla de observación exhaustiva y parcial.

La alternativa al enfoque anterior es la observación parcial. Esta implica que no se observa a toda la población. Dentro de esta forma de proceder se pueden distinguir dos categorías distintas. Por una lado está la subpoblación y por otro la muestra. Con la primera lo que se hace es observar a un conjunto de entes o elementos de la población que guardan entre si una cierta característica y que los diferencia de los demás. Así, siguiendo con el ejemplo anterior, los electores que no han votado antes porque en las elecciones anteriores no tenían la edad mínima exigida constituyen una subpoblación. Al proceder de esta forma se consigue realizar una operación estadística en menos tiempo y a menor coste pero, en cambio, la incertidumbre a cerca de la intención de voto del electorado es enorme, pues esa subpoblación no representa en absoluto a toda la población. Su intención de voto no tiene porque coincidir con la de los demás electores.

No obstante, esta forma de observar la población puede resultar de gran interés en determinados casos. Pensemos que nuestro interés se centra es cuantificar la ocupación hotelera en una zona turística. En esta situación, en lugar de preguntar a todos los establecimientos que se dedican a esta actividad económica, podría resultar suficiente con preguntarle solo a los hoteles a partir de una cierta categoría, por ejemplo a los de tres y más estrellas pues, en este caso, esos elementos de la población son determinantes de la población total y los demás tienen poca incidencia en el volumen de ocupación.

La segunda opción de la observación parcial consiste en tomar una muestra. En este caso se observará también un subconjunto de elementos de la población. Pero ahora,

a diferencia de lo que ocurría con la subpoblación, los elementos de la muestra no guardan ninguna característica especial que los diferencie de los demás. Al contrario, con una muestra lo que se pretende es representar a toda la población. Podríamos decir que la muestra es una población de tamaño reducido.

Las ventajas de observar la población de forma parcial y, en especial, para el caso de seleccionar muestras son, en algunos casos, evidentes. En primer lugar reduce el tiempo de observación. Si el tiempo que se dedica a observar los elementos de la población es excesivo podría ocurrir que los resultados llegaran más tarde de lo que es admisible. Siguiendo con el ejemplo de las elecciones, si el periodo de observación es superior al tiempo que dista hasta que tengan lugar las elecciones, entonces cuando se disponga de resultados sobre intención de voto ya no son necesarios. En general, si lo que se pretende al observar la población es analizar una característica que no cambia mucho con el transcurso del tiempo, entonces no importará demasiado que el periodo de observación sea razonablemente largo. Por el contrario, si esa característica está sometida a fuertes variaciones en periodos de tiempo cortos o si el plazo de presentación de resultados es breve, que duda cabe que en tales circunstancias la observación parcial, mediante una muestra, es el procedimiento más indicado.

En segundo lugar está el tema de los costes monetarios, que en la observación parcial son más reducidos que en el caso de la exhaustiva.

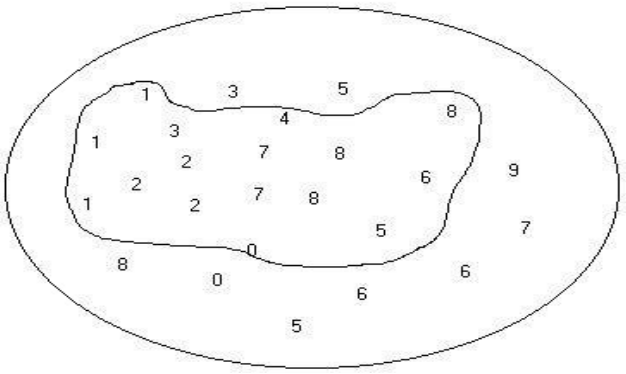
Finalmente, la observación parcial presenta también la ventaja de que reduce las pruebas destructivas. Imaginemos que nos encontramos frente a un estudio de control de la calidad de la producción de una empresa que se dedica a la fabricación de vigas de hormigón para obras civiles. Esas vigas habrán de someterse a presiones altas para conocer su resistencia a la ruptura. Pero si somete toda la producción a este tipo de pruebas destructivas entonces no hay producción. Bastaría en este caso con seleccionar una muestra y, aplicarle ese tipo de pruebas a los elementos de la misma, para tener una idea razonable de cual es la resistencia a la ruptura de las vigas producidas por la empresa.

Pero no todo son ventajas en la observación parcial. El principal inconveniente se deriva precisamente de que la observación no es exhaustiva. En estos casos las características de la población serán desconocidas, pues aunque la muestra pretenda representar lo más fielmente posible a la población, nunca dejará de ser eso, una

muestra. Con los datos de la muestra solo podremos conocer las características de esos valores muestrales. Concluir que son iguales a las de la población sería poco menos que una osadía. Entre las características observadas en la muestra y las de la población habrá siempre una diferencia que se conoce como error muestral. Es precisamente este error muestral el que lleva a que las decisiones, en relación con las características poblacionales, se tomen en situaciones de incertidumbre.

Veamos esto de una forma gráfica y sencilla. En la Figura 1 se han representado una población con todos sus elementos y una muestra de los mismos. Como puede apreciarse, la población toma valores que van del 0 al 9, mientras que en la muestra el valor 9 no está incluido. Así pues, según la muestra, los valores de la población van del 0 al 8, pero eso no es cierto, solo es aproximado. Se está cometiendo un error.

Figura 1. Población y muestra



1.3.1 Variables y atributos.

En líneas anteriores se ha señalado que el objeto de estudio de la Estadística son las poblaciones y que estas están formadas por entes o elementos. El número total de los mismo determina el tamaño de la población. Para estudiar una población, lo primero que debe hacerse es observarla de alguna de las formas que ya se ha señalado en las líneas anteriores. Pero observar una población es equivalente a observar sus elementos. Ahora bien, esos elementos poseen una serie de características que son las que realmente se observan. Por ejemplo, el conjunto de todas las empresas industriales radicadas en España constituyen una población. Los elementos de esa población son las empresas. Pero una empresa no se observa en abstracto. Lo que realmente tiene interés son las distintas características de esas empresas, como, por

ejemplo, el número de empleados, el volumen de ventas, los costes salariales, los gastos en publicidad, los beneficios de las mismas, la naturaleza de los productos que fabrican, etc.

A todas estas características de los elementos de una población se les conoce de forma genérica como caracteres. Estos últimos, según su naturaleza, pueden ser de tipo cuantitativo o cualitativo. Para el ejemplo anterior, serían caracteres cuantitativos "el número de empleados", "el volumen de ventas", "los costes salariales", "los gastos en publicidad", "los beneficios de las mismas", etc., mientras que sería cualitativo "la naturaleza de los productos que fabrican". Hay que señalar que, en general, cualquier carácter de tipo cuantitativo se puede ofrecer en términos cualitativos. Así, si el número de empleados lo agrupamos en intervalos se podría hablar de empresas pequeñas, medinas y grandes, siendo ahora el carácter "tamaño de la empresa" de naturaleza cualitativa. De manera similar se podría proceder con los demás.

Pero en estadística es más habitual hablar de variables que de caracteres cuantitativos y de atributos en lugar de caracteres cualitativos.

Las variables son susceptibles de medirse en términos cuantitativos y a cada una de esas posibles mediciones o realizaciones se les conoce como valores, datos u observaciones. A su vez, en función del número posible de valores que tome una variable, a las mismas se las puede clasificar en discretas y continuas. Serán discretas cuando el número de valores sea finito o infinito numerable, mientras que una variable será continua cuando el número de sus valores sea infinito no numerable. En los casos en los que las variables toman infinitos valores, la practica habitual es agruparlos en intervalos, como se muestra en las Tabla 1, para variable continua, y en la Tabla 2 para discreta.

Tabla 1. Retribución anual de los asalariados en Andalucía en 1996

Salarios (miles de ptas.)	Asalariados
0,000-0,454	618.604
0,454-0,909	266.378
0,909-1,363	208.329
1,363-1,818	212.353
1,818-2,272	188.908
2,272-2,727	132.436
2,727-3,181	107.217
3,181-3,636	93.230

3,636-4,090	78.737
4,090-4,544	52.578
4,544-4,999	29.889
4,999-5,453	21.318
5,453-5,908	15.217
5,908-6,362	11.367
6,362-6,817	8.433
6,817-7,271	6.504
7,271-7,725	5.322
7,725-8,180	4.311
8,180-8,634	3.425
Mayor de 8,634	13.114
Total	**2.077.670**

Fuente: IEA. Anuario Estadístico de Andalucía. 2001

Tabla 2. Tamaño de los municipios de Andalucía según su población en 1999.

Tamaño	Municipios
Menos de 501 hab.	90
De 501a 1000 hab.	95
De 1001 a 2000 hab.	138
De 2001 a 3000 hab.	94
De 3001 a 5000 hab.	106
De 5001 a 10000 hab.	115
De 10001 a 20000 hab.	69
De 20001 a 50000 hab.	40
De 50001 a 100000 hab.	11
Más de 100000 hab.	11
Total	**769**

Fuente: IEA. Anuario Estadístico de Andalucía. 2001

Los atributos no pueden medirse como ocurre con las variables. Lo único que puede hacerse con ellos es describirlos mediante palabras y clasificarlos en categorías no numéricas que sean mutuamente excluyentes. A cada una de estas categorías se le denomina modalidades. Un ejemplo es el que se recoge en la Tabla 3.

Tabla 3. Alumnos matriculados en las universidades andaluzas según rama en el curso 1999-2000

Ramas	Alumnos
Ciencias Experimentales	26155
Ciencias de la Salud	22509
Enseñanzas Técnicas	58588
Ciencias Sociales y Jurídicas	133181
Humanidades	31486
Total	**271919**

Fuente: IEA. Anuario Estadístico de Andalucía. 2001

En algunos casos, las modalidades de un atributo pueden ser objeto de ordenación, como se recoge en la tabla 4.

Tabla 4. Número de alumnos de enseñanza primaria en Andalucía según ciclo. Curso 1999-2000

Ciclo	Alumnos
De 1º ciclo	177093
De 2º ciclo	180931
De 3º ciclo	188218
Total	**546242**

Fuente: IEA. Anuario Estadístico de Andalucía. 2001

1.4 Fuentes estadísticas.

En los apartados anteriores se ha señalado que el objetivo de la Estadística es el estudio de los fenómenos de masas. Pero ello requiere el manejo de una información numérica amplia. La cuestión inmediata que surge es saber a donde se puede recurrir para encontrar esa información necesaria y sin la cual el análisis estadístico no se puede realizar. En definitiva, se trata es de conocer las fuentes que suministran información de carácter estadístico. Estas fuentes son susceptibles de clasificarse según distintos criterios. Atendiendo al agente que elabore esa información, la misma puede agruparse en endógena y exógena. La primera sería la que elabora el propio investigador. En este caso, la operación estadística conducente a recabar los datos necesarios para la realización del análisis estadístico se supone que la lleva a cabo el propio investigador. Será él quien se encargue de observar los distintos caracteres, cuantitativos o cualitativos, relevantes de los elementos de una población. El resultado será una base de datos, obtenida mediante una muestra, o cualquiera de los otros procedimientos indicados con anterioridad, que permitirá el correspondiente análisis estadístico.

Esta situación se da cuando no existe fuente alternativa exógena capaz de facilitar esa información. Pero ¿qué se entiende por fuente exógena? En general, la podemos definir como aquella cuyo objeto principal es la obtención de información estadística pero que no actúa como usuaria.

Las fuentes exógenas son múltiples y a su vez se las puede clasificar en dos categorías distintas. Por un lado están las fuentes oficiales o públicas y, por otro, las privadas. De todas ellas las que generan mayor volumen de información son las

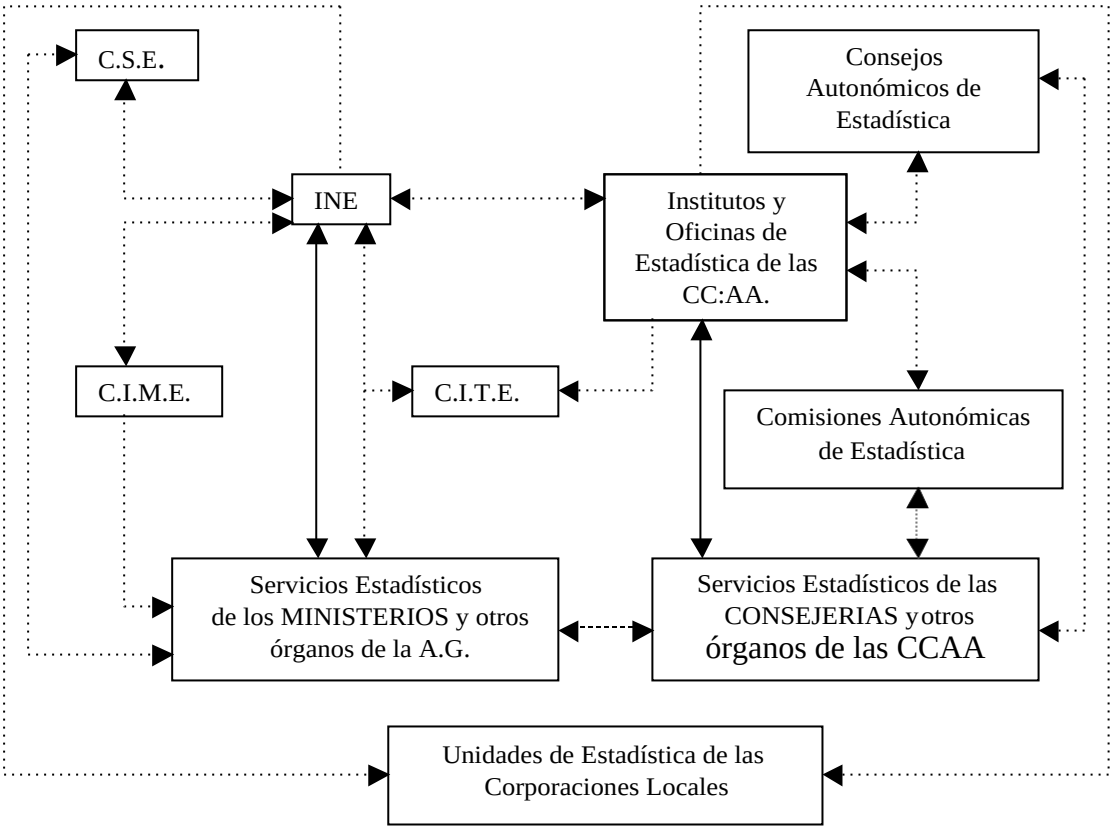
primeras, es decir, las oficiales o públicas. Estas últimas se pueden clasificar, a su vez, según el ámbito espacial en que desarrollan sus competencias en materia estadística. Así se tienen las fuentes de carácter internacional, las de ámbito estatal, las de ámbito regional o autonómico y las de carácter local.

Entre las primeras, los principales productores de información estadística son la Oficina de Estadística de la ONU (UNSTAT) y la Oficina de Estadística de la Unión Europea (EUROSTAT).

La segunda y tercera de las categorías contempladas conforman el sistema estadístico nacional, cuya organización responde al esquema de la Figura 1.

Figura 1. Esquema de la organización Estadística en España.

ADMINISTRACIÓN GENERAL DEL ESTADO ADMINISTRACIONES AUTONOMICAS (CC.AA.)



Nota: C.S.E. (Consejo Superior de Estadística), C.I.M.E. (Comisión Interministerial de Estadística), C.I.T.E. (Comité Interterritorial de Estadística), I.N.E (Instituto Nacional de Estadística).

Fuente : Página web del INE.

Dentro de la segunda categoría hay que destacar al INE por ser el órgano productor de estadísticas más importante de toda la organización estadística de España. El

mismo es el encargado de la elaboración y diseño de los planes estadísticos nacionales plurianuales que son finalmente aprobados por Real Decreto. Una vez aprobado el Plan Estadístico Nacional correspondiente a un cuatrienio concreto, el INE elabora los programas anuales, que también han de ser aprobados por Real Decreto. El objetivo final, tanto de los planes como de los programas, es definir y describir, de forma ordenada y sistemática, el conjunto de operaciones estadísticas que darán lugar a las futuras publicaciones en materia estadística, tanto del INE como del resto de los productores de ámbito estatal (Ministerios, Institutos Oficiales, Banco de España, etc). El conjunto de operaciones es muy numeroso, razón por la cual no es aconsejable describir todas y cada una de ellas. En la Figura 2 se reproduce las fichas de dos de esas operaciones, siendo la estructura de ellas idéntica a la del resto[1].

Figura 2. Ficha técnica de la operaciones estadísticas "Directorio Central de Empresas" y "Sociedades Mercantiles".

3841 Directorio Central de Empresas (DIRCE)

Fines:
Mantenimiento de un directorio central de empresas no agrarias y sus unidades locales clasificadas por actividad, tamaño y localización Marco para encuestas
Organismos que intervienen:
INE
Descripción general (principales variables):
Identificación y localización, tamaño, rama de actividad
Colectivo:
Empresas y unidades locales
Periodicidad de la recogida de la información:
Anual
Desagregación:
Municipal o inferior
Créditos presupuestarios necesarios para su financiación en el cuatrienio 2001-2004:
181,75 millones de pesetas (1.092,34 miles de euros)

3844 Sociedades Mercantiles

Fines:
Información mensual y anual de las sociedades creadas, disueltas, y de las modificaciones de sociedades
Organismos que intervienen:
INE, D.G. de los Registros y del Notariado (MJ)
Descripción general (principales variables):
Número y capital social de las sociedades constituidas, disueltas o que modifican su capital
Colectivo:
Sociedades mercantiles
Periodicidad de la recogida de la información:
Mensual
Desagregación:
Provincial
Créditos presupuestarios necesarios para su financiación en el cuatrienio 2001-2004:
97,55 millones de pesetas (586,29 miles de euros)

Fuente: Página web del INE.

[1] Para un estudio más detallado, tanto en materia de planes y programas como de operaciones estadísticas, se puede consultar la página web del INE, cuya dirección es: http://www.ine.es.

Pero previo a la formulación del Plan Nacional de Estadística está el Inventario de Operaciones Estadísticas de la Administración General del Estado (IOE), el cual contiene la descripción de la actividad estadística de los Ministerios, Banco de España e Instituto Nacional de Estadística. El Inventario es un instrumento fundamental para la coordinación y planificación estadística y el punto de partida para la formulación del Plan Estadístico Nacional. Las operaciones estadísticas[2] que recoge el inventario aparecen en fichas similares a las del Plan, como se aprecia en la Figura 3.

Paralelamente a esta producción estadística de ámbito estatal, la transferencia de competencias en materia estadística ha posibilitado la creación de Institutos y Oficinas de Estadística en las distintas Comunidades Autónomas, los cuales, a su vez, definen sus propios planes y programas estadísticos. Así, para el caso de la Comunidad Autónoma de Andalucía, su actividad estadística se enmarca dentro del Sistema Estadístico de Andalucía (SEA), que tiene como misión poner a disposición de los ciudadanos y las instituciones oficiales información estadística sobre la situación económica, demográfica, social, medioambiental y territorial de la Comunidad Autónoma.

El SEA está integrado por: el Instituto de Estadística de Andalucía (IEA), la Unidades Estadísticas de las distintas Consejerías, los Puntos de Información Estadística, el Consejo de Dirección y el Consejo Andaluz de Estadística.

En este caso, es el IEA el encargado de la elaboración de Planes y Programas para el ámbito andaluz, al igual que el INE para todo el Estado.

Este esquema de la Comunidad Autónoma de Andalucía, similar al de la Administración General del Estado, es trasladable a mayoría de las Comunidades del Estado Español.

---

[2] Como indica el propio INE: “La unidad adoptada como base del inventario es la *operación estadística*, definida como el conjunto de actividades que conducen a la obtención de resultados estadísticos sobre un determinado sector o tema a partir de datos recogidos de forma individualizada. También se incluyen en el ámbito de esta definición los trabajos de infraestructura y de normalización estadística que posibilitan la coordinación, homogeneización e integración de las estadísticas, así como la recopilación de resultados y la confección de síntesis”.

Figura 3. Ficha técnica de la operación estadística "Directorio Central de Empresas" del IOE.

30201

Directorio Central de Empresas (DIRCE)

Servicio responsable	Instituto Nacional de Estadística (INE)
Unidad ejecutora	S.D.G. de Metodología y Técnicas Estadísticas (D.G. de Procesos e Infraestructura Estadística)
Participación de otros organismos	No
Clase de operación	Marcos para censos o muestras
Sector o tema	Estadísticas de empresas y unidades de producción no referidas a sectores particulares
Subsector o subtema	Registros y directorios de unidades de producción
Nivel de desagregación	Municipal o inferior
Metodología de la recogida de datos	Obtención mediante enumeración completa de datos administrativos originales
Forma de recogida de datos	Transcripción de documento administrativo, cualquiera que sea la forma de cumplimentar éste
Objetivo general	Mantenimiento de un directorio central de empresas no agrarias y sus unidades locales clasificadas por actividad, tamaño y localización
Variables de estudio	Identificación, localización, tamaño y rama de actividad
Variables de clasificación	No aplicable
Tipo de difusión	"El Directorio Central de Empresas (DIRCE). Resultados estadísticos". Soportes papel, publicación electrónica e internet (www.ine.es)
Periodicidad de la difusión	Anual
Periodicidad de la recogida de datos	Anual
Unidades	Empresas y unidades locales
Fuente administrativa (en su caso)	Ficheros de la Administración Tributaria, Seguridad Social y comunidades autónomas de Navarra y País Vasco
Figura en el Plan Estadístico Nacional 2001-2004	Sí - 3841

Fuente: Página web del INE.

CAPITULO 2.- ANALISIS DE UNA VARIABLE .

2.1 Introducción.

En este capítulo se darán un conjunto de instrumentos que permitirán el análisis descriptivo de una variable. En primer lugar se indicará la forma de organizar y presentar la información, una vez que se ha observado la población y ha sido medido uno de los caracteres de todos y cada uno de los elementos de la misma. Esta operación nos llevará a la obtención de una distribución de frecuencias. Una vez que se tienen los datos organizados mediante esa distribución hay que iniciar el proceso de análisis de la variable. Para ello, el primer instrumento al que se puede recurrir, tanto por su sencillez como por lo fácil de su interpretación, es la representación gráfica de ese carácter. Con la representación gráfica de la variable o del atributo, según proceda en cada caso, se consigue tener una visión de conjunto del fenómeno estudiado de una forma más rápida y perceptible que con la sola inspección numérica de la tabla o distribución. Para continuar este proceso de análisis de una variable hay que definir ciertos instrumentos que nos permitan estudiar sus características más relevantes. Entre las mismas cabe destacar las siguientes: 1) forma de la distribución; 2) medidas de posición (valor central o promedios); 3) dispersión; 4) asimetría; 5) curtosis.

2.2 Distribuciones de frecuencias unidimensionales.

El adjetivo unidimensional hace referencia a que en nuestro análisis solo se tendrá en cuenta un carácter. Al mismo se le va a representar simbólicamente mediante la letra *X*, mientras que para sus posibles realizaciones (valores o modalidades, según se trate de variable o de un atributo, respectivamente) se utilizará la letra *x* minúscula.

En lo sucesivo se entenderá que el carácter observado es de tipo cuantitativo y que, en consecuencia, estamos trabajando con variables. En realidad el tratamiento que se le da a los atributos, en cuanto a distribución de frecuencias, es muy similar al de las variables discretas.

Por *distribución de frecuencias se va a entender al conjunto de valores que ha tomado una variable con sus frecuencias correspondientes*. Simbólicamente, una distribución de frecuencias vendría dada por los pares (x_i, n_i), donde x_i son los valores de la variable y n_i son sus frecuencias. Hay que señalar, en esta definición, que la frecuencia asociada a un valor de la variable es el número de veces que se repite ese valor. A la misma se le conoce como frecuencia absoluta.

2.2.1 Distribuciones de frecuencias para valores no agrupados.

Pueden considerarse básicamente dos tipos de distribuciones de frecuencias. Aquellas en las que los valores de la variable no están agrupados y las que presentan esos valores agrupados en intervalos.

Las primeras se corresponden a variables discretas. Este tipo de variables se caracteriza, como ya se indicó en su momento, por tener un número finito de valores o infinito numerable, de forma que entre dos valores consecutivos no existe otro. Pero aunque estos valores sean observables resulta que, a veces, el número de ellos es tan elevado que resulta aconsejable presentar la distribución o tabla estadística con los valores de la variable agrupados en intervalos. Esta forma de proceder podría llevarnos a pensar que estamos trabajando con variables continuas, cuando en realidad no lo son, pues en el caso de éstas, a diferencia de las primeras, dentro de cualquier intervalo de valores se pueden considerar que hay infinitos valores distintos.

La forma estándar de dar una distribución de frecuencias con valores no agrupados es la que aparece en la Tabla 1. Las frecuencias de esta tabla puede ser unitarias o mayores que uno. El primer caso tienen poco interés para la Estadística, pues como ya se indicó en la capitulo primero, el objeto de la misma era el estudio de colectivos grandes y nunca las poblaciones con un número de elementos muy reducido.

Junto a las frecuencias absolutas de los valores de una variable resulta habitual dar, también, lo que se conoce como frecuencias relativas. Para un valor concreto, la frecuencia relativa, que representaremos por f_i, es el cociente entre la frecuencia absoluta y el numero total de observaciones N. Es decir, $f_i = n_i/N$. Estas frecuencias se puede expresar en porcentajes o en tantos por uno. A su vez, las frecuencias, tanto las absolutas

como las relativas, se puede dar de forma acumulada. Las frecuencias absolutas acumuladas se representan por N_i y las relativas acumuladas por F_i.

Tabla 1. **Distribución de frecuencias para valores no agrupados**.

Valores de la variable	**Frecuencias Absolutas**	**Frecuencias Relativas**	**Frecuencias absolutas acumuladas**	**Frecuencias relativas acumuladas**
x_i	n_i	$f_i = n_i / N$.	N_i	$F_i = N_i /N$
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = N_1 + n_2$	$F_2 = F_1 + f_2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_i	n_i	f_i	$N_i = N_{i-1} + n_i$	$F_i = F_{i-1} + f_i$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_k	n_k	f_k	$N_k = N$	$F_k = 1$
	$_i\, n_i = N$	$_i\, fi = 1$		

Ejemplo 1*. En la tabla adjunta se da la distribución de las 500 hogares de un barrio según el número de sus miembros.*

Tamaño de los hogares xi	***Nº de hogares ni***	f_i	N_i	F_i
1	*40*	*0,08*	*40*	*0,08*
2	*70*	*0,14*	*110*	*0,22*
3	*110*	*0,22*	*220*	*0,44*
4	*90*	*0,18*	*310*	*0,62*
5	*48*	*0,096*	*358*	*0,716*
6	*42*	*0,084*	*400*	*0,8*
7	*40*	*0,08*	*440*	*0,88*
8	*35*	*0,07*	*475*	*0,95*

9	*20*	*0,04*	*495*	*0,99*
10	*5*	*0,01*	*500*	*1*
	500	*1*		

Esta distribución, además de dar los valores de la variable y sus frecuencias absolutas, recoge las relativas y las acumuladas. Se trata de la distribución de una variable discreta con un número finito de valores.

2.2.2 Distribuciones de frecuencias para valores agrupados.

Este tipo de distribuciones se asocia, fundamentalmente, con variables continuas, aunque, como ya se ha señalado con anterioridad, en algunos casos también es aplicable a variables discretas, especialmente en aquellas situaciones en las que la variable toma muchos valores, de forma que si éstos nos e agruparan, la tabla resultaría demasiado extensa y la función de síntesis de la misma se perdería.

La elaboración de la distribución de frecuencias de una variable continua plantea algunos problemas que no se dan en el caso de variables discretas. Se trata de decidir el número de intervalos en los que hay que agrupar los valores de la variable así como si la amplitud o recorrido de los mismos debe ser igual. Estas cuestiones no tienen una respuesta determinada de antemano. La solución dependerá de cada caso concreto, por lo que no tiene sentido entrar en la casuística de las distintas situaciones que pudieran darse. Otro problema surge cuando un valor de la variable coincide exactamente con un extremo del intervalo, con lo que hay dudas sobre su inclusión en ese intervalo o el siguiente. Como solución a este problema es habitual proceder a definir intervalos abiertos por la izquierda y cerrados por la derecha, lo que implica que intervalo definido entre *a* y *b* incluye a todos los valores menores o iguales que *b* pero mayores que *a*.

En general, una distribución de frecuencias para una variable continua será como la que se da en la Tabla 2. A partir del contenido de esta tabla hay que definir los siguientes conceptos:

a) Amplitud del intervalo. Es la diferencia entre el extremo superior y el inferior. Así para el intervalo *i-ésimo*, la amplitud vendría dada por:

$$a_i = L_i - L_{i-1} \qquad (2.1)$$

b) Marca de clase. Es el punto central de cada intervalo. Esta se suele representar por x_i. Para el intervalo *i-ésimo* viene será:

$$x_i = (L_i + L_{i-1})/2 \qquad (2.2)$$

Tabla 2. Distribución de frecuencias para valores agrupados

Variable (intervalos $L_{i-1} - L_i$)	Amplitud a_i	Marca de clase x_i	Frec. abs. n_i	Frecuencia relativa f_i	Frecuencia Absoluta Acumulada N_i	Frecuencia Relativa Acum. F_i
$L_0 - L_1$	a_1	x_1	n_1	n_1/N	$N_1 = n_1$	N_1/N
$L_1 - L_2$	a_2	x_2	n_2	n_2/N	$N_2 = n_1 + n_2$	N_2/N
$L_2 - L_3$	a_3	x_3	n_3	n_3/N	$N_3 = n_1 + n_2 + n_3$	N_3/N
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
$L_{i-1} - L_i$	a_i	x_i	n_i	n_i/N	$N_i = n_1 + n_2 + ... + n_i$	N_i / N
.	.	.	.	.	.	.
.	.	.	.	.	.	.
$L_{k-1} - L_k$	a_k	x_k	n_k	n_k/N	$N_k = n_1 + n_2 + ... + n_k = N$	$N_k/N = 1$
			$\sum_i n_i = N$	$\sum_i fi = 1$		

Esta agrupación de los valores de la variable en intervalos, aunque resulta operativamente necesaria, conlleva un problema grave que se conoce como error de agrupamiento. Este error es la consecuencia directa de la pérdida de información provocada al incluir un conjunto de observaciones en un mismo intervalo. Así, antes de agrupar las n_i observaciones en el intervalo *i-ésimo*, se sabe cuales son los valores concretos observados y sus respectivas frecuencias individuales. Ahora bien, cuando esos valores se agrupan en un intervalo se pierde esa información individualizada. En esos casos lo que se hace es sustituir a todos y cada uno de ellos por su valor medio que viene representado por la marca de clase. Pero esta solución, como se verá más adelante, implica asumir ciertos supuestos que nos inducen a error. Este error es el coste de la pérdida de información que se causa por el agrupamiento de las valores de la variable.

Ejemplo 2*. La distribución del presupuesto semanal en alimentación de un conjunto de 265 familias expresado en euros es el que figura en la tabla siguiente:*

Presupuestos L_{i-1} - L_i	a_i	x_i	**Familias** n_i	f_i	N_i	F_i
80-100	20	90	10	0,0377	10	0,0377
100-110	10	105	35	0,1321	45	0,1698
110-115	5	112,5	40	0,1509	85	0,3208
115-120	5	117,5	45	0,1698	130	0,4906
120-130	10	125	55	0,2075	185	0,6981
130-150	20	140	30	0,1132	215	0,8113
150-170	20	160	20	0,0755	235	0,8868
170-210	40	190	15	0,0566	250	0,9434
210-270	60	240	10	0,0377	260	0,9811
270-360	90	315	5	0,0189	265	1,0000
Total			**265**	**1**		

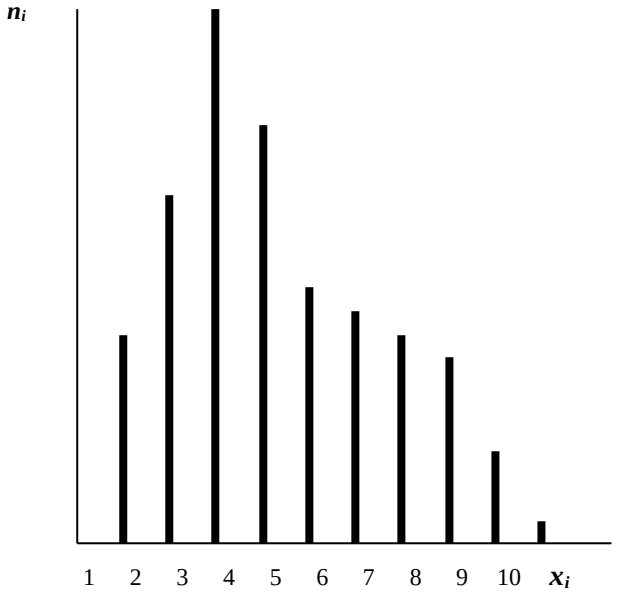
En este caso se trata de una variable continua con sus valores agrupados en intervalos cuya amplitud es variable. Este tipo de intervalos permite tratar de forma distinta a los valores de la variable, según donde se localice la mayor parte de las observaciones. En este sentido la amplitud de los intervalos es inversa a la frecuencia de los mismos. Esta forma de proceder evita que la mayor parte de las observaciones se concentre en un solo intervalo o en unos pocos.

2.3 Análisis gráfico

El tipo de representación gráfica depende en gran medida de la naturaleza del carácter de los elementos de la población con el que se esté trabajando. Así, si se trata de una variable se recurrirá al diagrama de barras en el caso de que sea discreta y sus valores no estén agrupados. Este diagrama se realiza haciendo uso de un sistema cartesiano en el que sobre el eje de abscisas se ponen los valores de la variable y sobre el de ordenadas las frecuencias, tanto absolutas (n_i) como relativas (f_i). Un ejemplo de este tipo de gráfico es el que se da en la Figura 1, donde se han representado los datos del Ejemplo 1. Hay que señalar que la anchura de las barras será, en cualquier caso, una cuestión de estética, pues la misma no hace alusión, en ningún caso, ni al valor de la variable ni al frecuencia del mismo. Dicho de otra forma, la superficie de la misma es independiente de la magnitud que se representa. En este sentido tan válido es el diagrama dado en la Figura 1 como el de la Figura 2.

Tanto en la Figura 1 como en la Figura 2 se ha representado las frecuencias absolutas. Pero estas figuras no habrían cambiado para nada si en su lugar si se hubiera trabajado con las frecuencias relativas. La única diferencia es que el eje de ordenadas tomaría como valor máximo la unidad, pero la proporción entre las barras no cambia de un gráfico a otro.

Figura 1. Diagrama de barras para la distribución del Ejemplo 1.



Otra represtación gráfica que también puede realizarse con los datos de una variable discreta es lo que se conoce como un diagrama escalonado o acumulativo. En este caso, sobre el eje de abscisas se siguen llevando los valores de la variable, mientras que sobre el de ordenadas se colocan las frecuencias acumuladas, bien absolutas (N_i) o relativas (F_i). En la figura 3 se ha representado el diagrama escalonado para la variable del Ejemplo 1.

Figura 2. Diagrama de barras para los datos del Ejemplo 1.

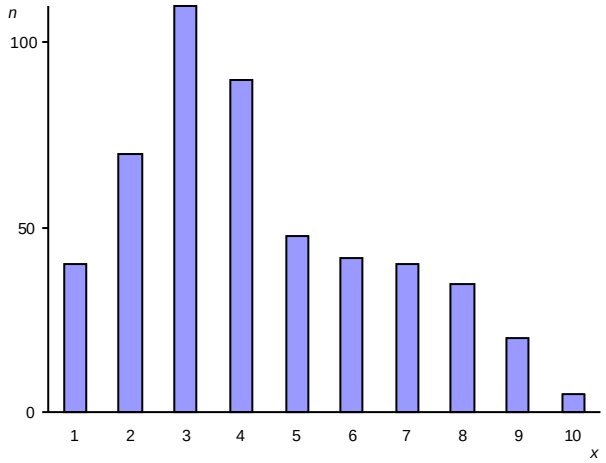
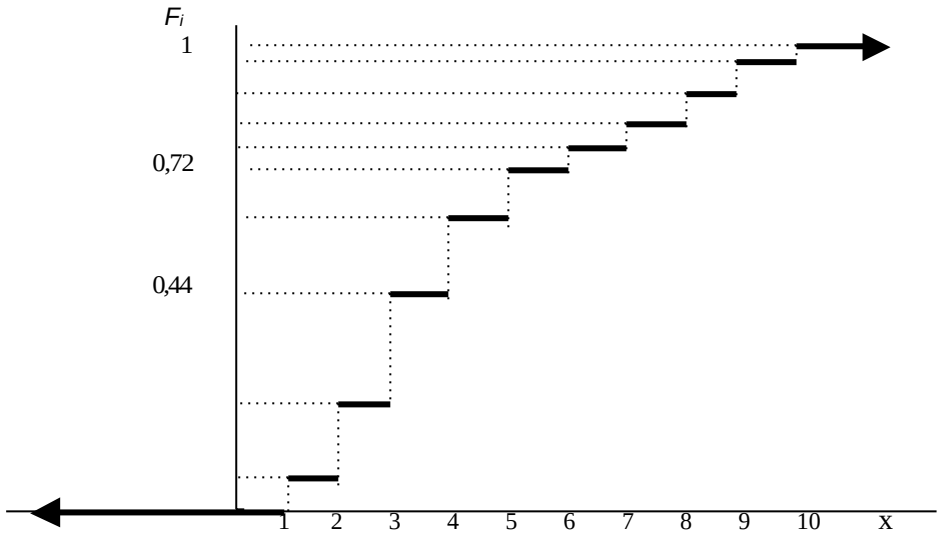


Figura 3. Diagrama escalonado correspondiente a los datos del Ejemplo 1.



Frente a este tipo de gráfico, cuando la naturaleza de la variable sea continua, entonces la representación gráfica más adecuada es el histograma o también conocido como histograma de frecuencias. Este tipo de gráficos podría utilizarse también en los casos de variables discretas con valores agrupados, aunque no resulta aconsejable hacer uso de

los histogramas para variables discretas por los problemas que conlleva asimilar una variable discreta a otra de tipo continuo.

Un histograma se realiza también haciendo uso de un sistema cartesiano, donde sobre el eje de abscisas se llevan los valores de la variable. Pero ahora ya no se trata de valores puntuales, sino de intervalos, y sobre éstos se levantan rectángulos, que tienen por base la amplitud del intervalo y por altura su frecuencia. El área de esos rectángulos deberá ser siempre proporcional a la frecuencia, de manera que cuando la amplitud de los intervalos no sea constante, entonces la altura de los rectángulos no será la frecuencia sino lo que se conoce como densidad de frecuencia definida de la forma siguiente:

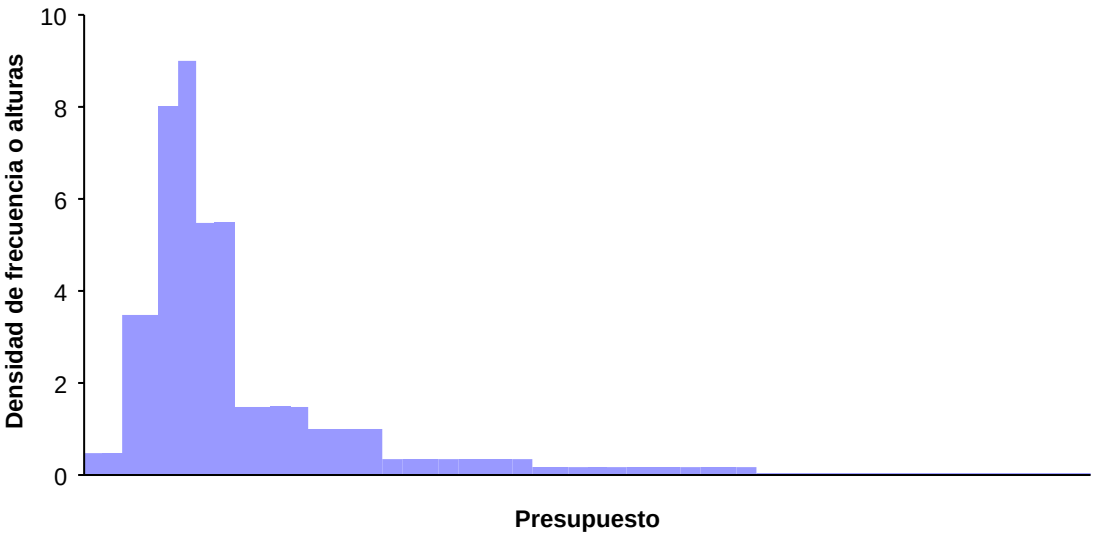
$$h_i = \frac{n_i}{a_i} \qquad i = 1,2,3,\ldots,k \qquad (2.3)$$

La Figura 4 recoge el histograma correspondiente a los datos del Ejemplo 2. En este caso se ha procedido a calcular las correspondientes densidades de frecuencias también conocidas como alturas, dado que los intervalos son de amplitud variable. Los datos numéricos que se han representado son los que aparecen en la Tabla 3.

Tabla 3. Distribución de los presupuestos familiares.

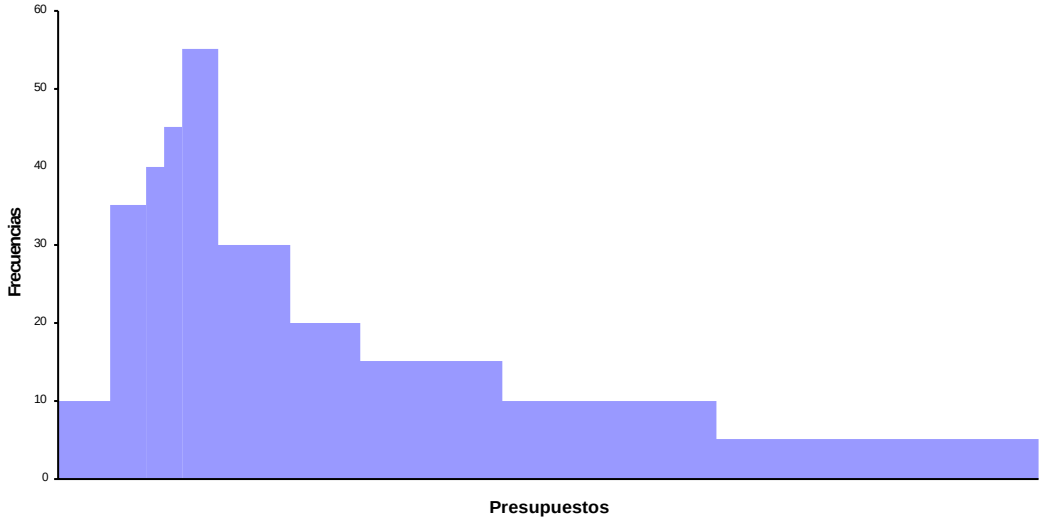
Presupuestos L_{i-1} - L_i	a_i	Familias n_i	h_i
80-100	20	10	0,50
100-110	10	35	3,50
110-115	5	40	8,00
115-120	5	45	9,00
120-130	10	55	5,50
130-150	20	30	1,50
150-170	20	20	1,00
170-210	40	15	0,38
210-270	60	10	0,17
270-360	90	5	0,06
Total		**265**	

Figura 4. Histograma para los datos del Ejemplo 2.



Si en lugar de trabajar con las alturas se hubieran llevado sobre el eje de ordenadas directamente las frecuencias, entonces el histograma correspondiente es el que aparece en la Figura 5. Como puede apreciarse, éste es muy distinto del anterior. Este último no es correcto porque el área de cada rectángulo no es proporcional a las frecuencias y, en consecuencia, muestra una realidad distorsionada.

Figura 5. Histograma para los datos del Ejemeplo 2.

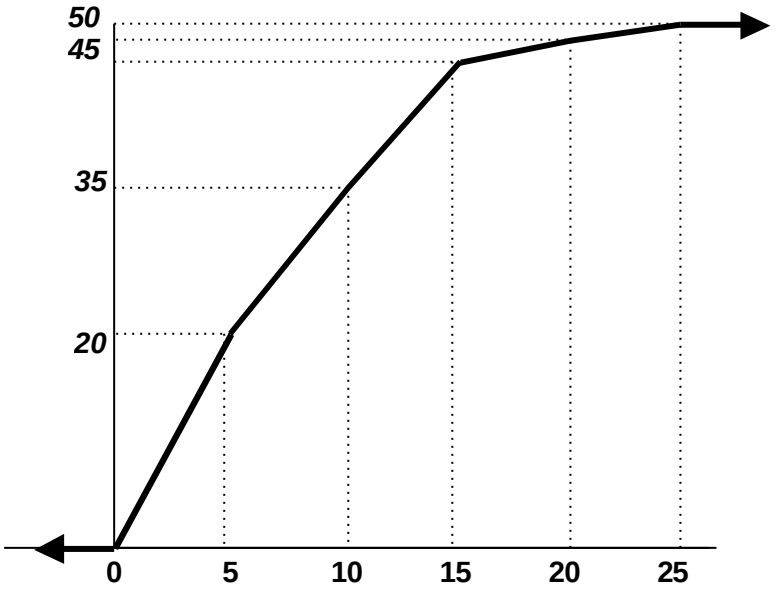


Al igual que para variables discretas se definió el diagrama escalonado para representar las frecuencias acumuladas, para el caso de continuas se puede proceder de forma similar. Pero en este caso, a la gráfica correspondiente, se le conoce como polígono de frecuencias acumuladas.

***Ejemplo 3**. A 50 clientes de una institución financiera se les ha preguntado por el tiempo que han tenido que esperar en la cola de la caja para realizar una gestión. Sus respuestas se han organizado en la siguiente tabla.*

Tiempo	*Clientes*	N_i
0-5	*20*	*20*
5-10	*15*	*35*
10-15	*10*	*45*
15-20	*3*	*48*
20-25	*2*	*50*

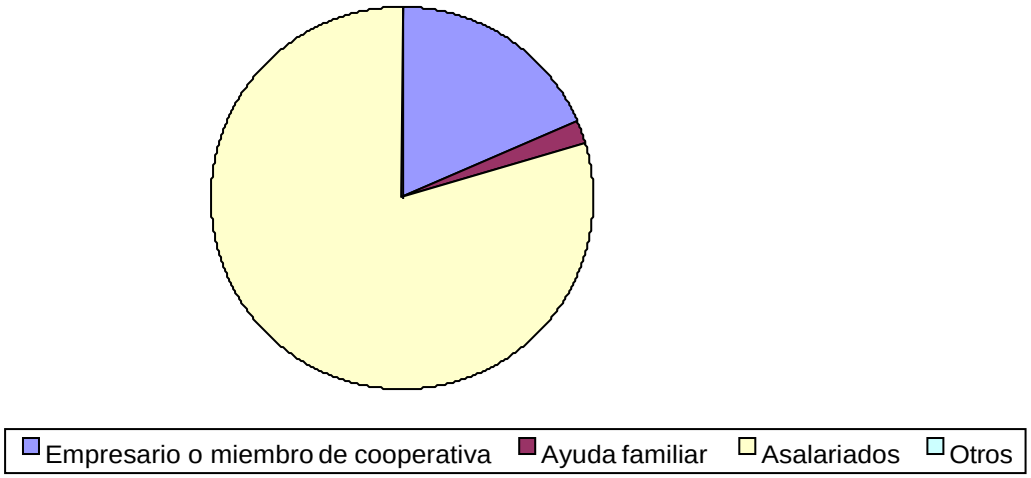
Obtenga el polígono de frecuencias acumuladas.



Una vez que se han señalado los instrumentos gráficos más habituales en el análisis de variables, quedan por introducir los correspondientes cuando de lo que se trata es de atributos. Ahora, las modalidades no tienen la dimensión numérica de los valores de las variables. Esto hace que haya que pensar en otros tipo de gráficos. Entre los más

utilizados están los diagramas de tarta. En estos casos lo que se hace es asignarle a cada modalidad del atributo un sector circular proporcional a su frecuencia. Para aclarar esta idea en la Figura 6 se ha representado la distribución de la población ocupada en España según su situación profesional, considerando cuatro modalidades distintas.

Figura 6. Población ocupada en España según situación profesional en 2000. (Miles de personas)

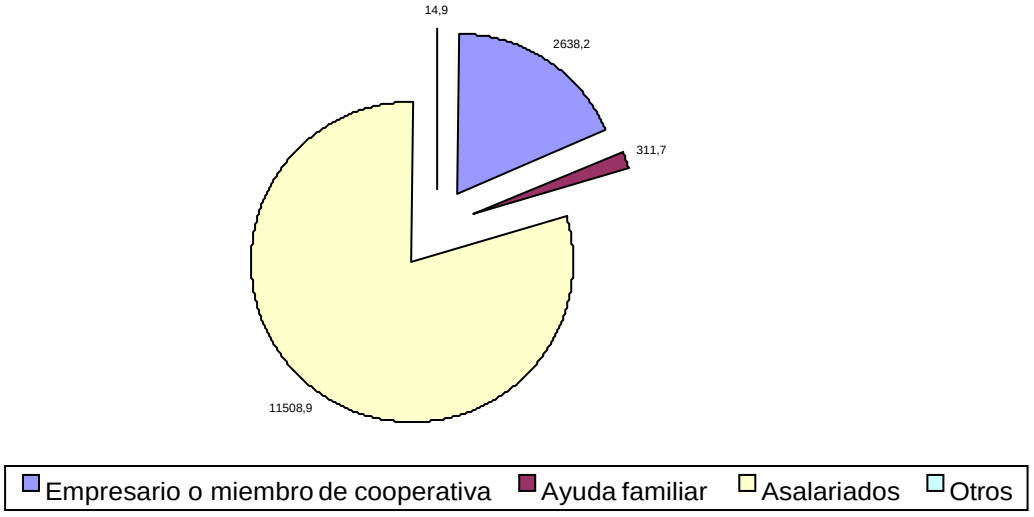


Fuente: EPA. Página web del INE

Este tipo de gráficos permite visualizar de forma bien clara cual es la estructura de un atributo concreto como es en este caso la situación profesional de la población ocupada, donde, como puede apreciarse la mayor parte de los ocupados son asalariados. Además estos gráficos se prestan a que puedan realizarse comparaciones para el mismo atributo en distintos ámbitos espaciales (por ejemplo, España y la situación en las CC.A.A. o con respecto a otros países) o temporales (por ejemplo, el año 2000 con respecto a años anteriores).

Cuando, como en el caso anterior, una de las modalidades no se aprecia porque su frecuencia es muy pequeña, entonces el gráfico se puede presentar como se hace en la Figura 7.

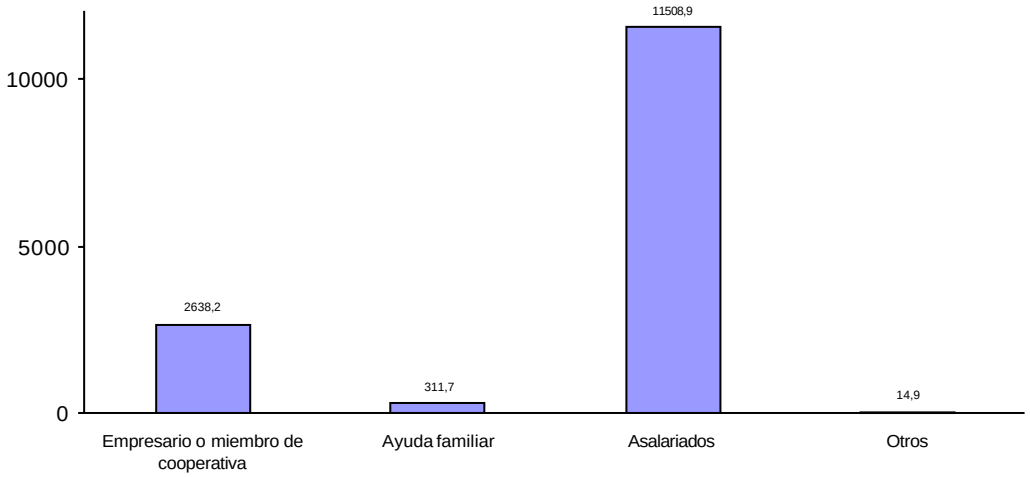
Figura 7. Población ocupada en España según situación profesional en 2000. (Miles de personas)



Fuente: EPA. Página web del INE

Este mismo atributo se ha representado en la Figura 8 haciendo uso de un diagrama de rectángulos. Este instrumento gráfico es muy similar al diagrama de barras visto para variables discretas.

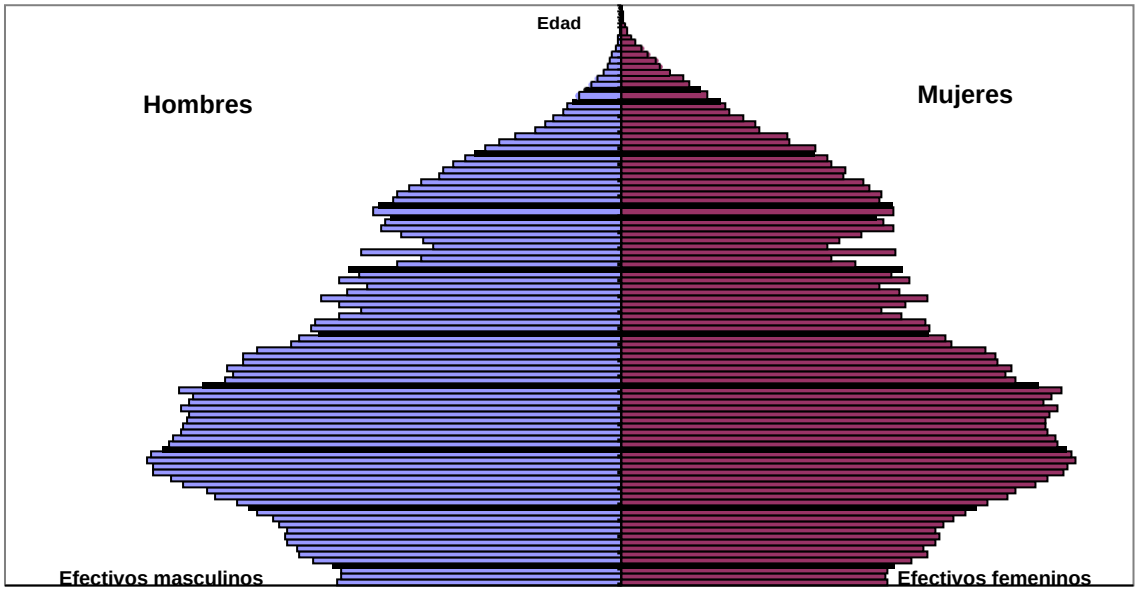
Figura 8. Población ocupada en España según situación profesional en 2000. (Miles de personas)



Fuente: EPA. Página web del INE

Pero este repertorio de gráficos no agota las posibilidades de representación. A los mismos se le puede añadir los pictogramas, cartogramas, etc. Sin embargo, los señalados son los que se utilizan con más frecuencia y, en cualquier caso, según el atributo con el que se esté trabajando habrá que seleccionar el más adecuado de entre la amplia gama de tipos de gráficos existentes.

Figura 9 . Pirámide de la población de Andalucía. 1998



Para finalizar este apartado dedicado a las representaciones gráficas vamos a dedicar unas líneas a un gráfico que tiene la particularidad de que en el mismo se hace uso de una variable continua, como es la edad, y un atributo, como es el sexo de la población. Nos estamos refiriendo a las pirámides de población, instrumento gráfico de gran utilidad en Demografía. Se trata de dos histogramas que comparten el mismo eje de abcisas, sobre el cual se lleva la edad de la población. Por otro lado, sobre el eje de ordenadas se llevan los efectivos poblacionales, tanto de hombres como de mujeres. Pero se trata de efectivos expresados no en cifras absolutas sino en porcentajes o en cualquier potencia de diez. Una muestra de este tipo de gráficos es el recogido en la Figura 9, donde se muestra la estructura por sexo y edad de la población de Andalucía para 1998.

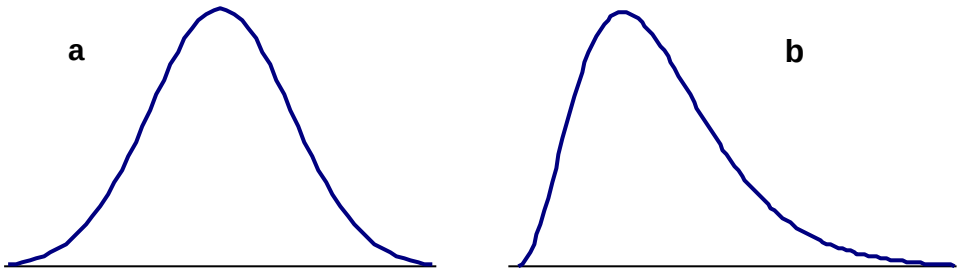
2.4 Forma de la distribución

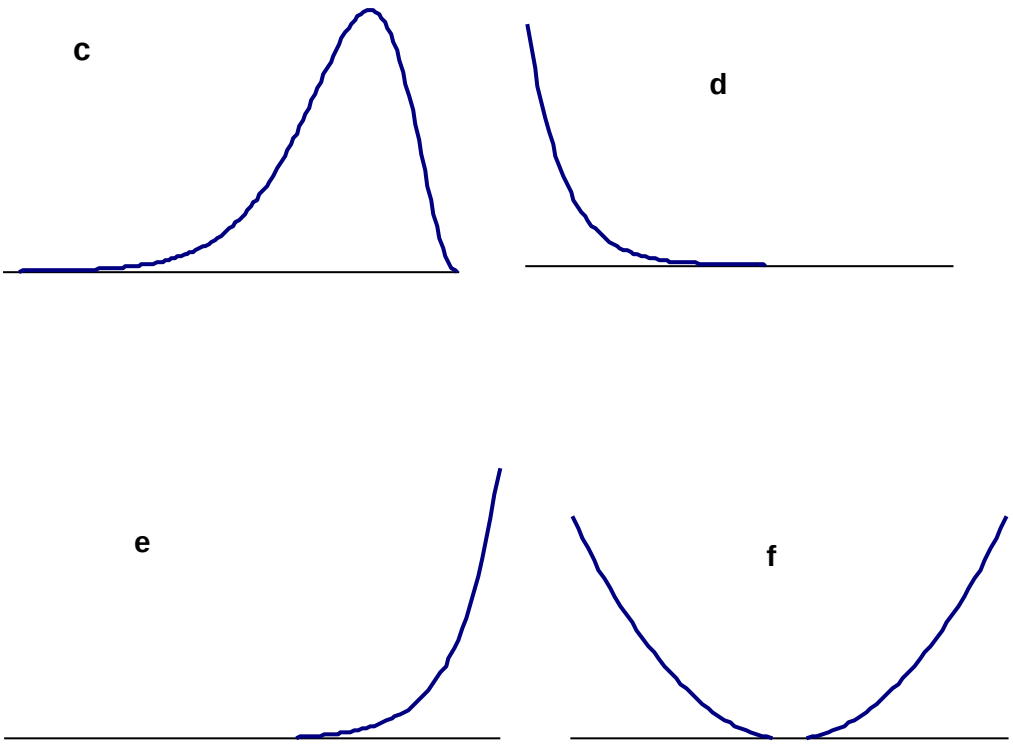
Para poder estudiar la forma de una distribución es preciso disponer de un número de observaciones de la variable suficientemente grande, como para poder determinar patrones de comportamiento o regularidades en dichas observaciones. Así, si se dispone de muy pocas observaciones, no tiene sentido hablar de la forma de la distribución.

Para estudiar la forma de la distribución la mejor herramienta de la que podemos hacer uso es de su representación gráfica, ya sea el diagrama de barras o el histograma. A partir de estas gráficas será posible, de forma fácil, concluir si las observaciones están concentradas en torno a unos pocos valores de la variable o si ocurre lo contrario, si la concentración tiene lugar en un extremo del recorrido de la variable o en el otro, etc.

Las formas más habituales de la distribución de una variable son las de tipo campaniforme (a), tanto simétricas como asimétricas (b, c). Otro formas también habituales son la que tienen forma de *jota*, invertida o normal (d, e), y las distribuciones en forma de U (f).

Figura 10. Formas de la distribución de una variable.





2.5 Promedios.

La reducción estadística que se consigue mediante la tabulación, en la mayoría de los casos, no resulta suficiente si lo que se persigue es que el "exceso" de información no nos impida ver lo que hay detrás de ella. Por tal razón, esa reducción hay que llevarla hasta el extremo de quedarnos con un solo dato que a su vez sea representativo de todo el conjunto. A ese único dato se le conoce de forma genérica como promedio. Con la obtención de promedios lo que se consigue es determinar cual es el nivel medio de la variable y, además, facilita las comparaciones entre variables. A los promedios se les conoce también como medidas de tendencia central. En principio, el único requisito que, de forma general, se le exige a cualquier promedio es que su valor esté comprendido entre los valores extremos de la variable. Con esta única condición, el número de promedios que pueden definirse es muy elevado, si bien los más habituales son la media aritmética, la media geométrica, la mediana, la moda, la media cuadrática, la media armónica, etc.

De todos ellos vamos a estudiar la media aritmética, la media geométrica, la mediana y la moda.

2.5.1 La media aritmética.

La media aritmética se define como aquel valor que resulta de dividir la suma de todas la observaciones entre el número de ellas. El resultado de este cociente no coincidirá necesariamente con valor alguno de la variable, pero si debe ser un valor del recorrido de la misma y vendrá expresado en las misma unidad de medida de esa variable. Por la forma en que se ha definido este promedio solo tiene sentido aplicarlo a variables de naturaleza cuantitativa, pues sus valores se pueden sumar, pero no las modalidades de un atributo.

Formalmente, si x_i es el valor *i-ésimo* de la variable X, siendo n_i el número de veces que se presenta ese valor, entonces la media se define como:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} \qquad (2.4)$$

Ejemplo 1: *Sea X una variable que representa el volumen de facturación de una empresa a lo largo de los 25 días laborales de un mes:*

x_i(miles €)	n_i
10,5	2
12,4	3
13,2	9
14,8	6
15,8	4
16,5	1

En este caso la media vendrá dada por:

$$\bar{x} = \frac{(10{,}5)(2) + (12{,}4)(3) + (13{,}2)(9) + (14{,}8)(6) + (15{,}8)(4) + (16{,}5)(1)}{25} = 13{,}82 \quad miles \quad €.$$

El ejemplo anterior es válido solo en el caso de distribuciones con datos no agrupados. Si los datos estuvieran agrupados, entonces los valores individuales de la variable son desconocidos, lo que impide hacer uso de la expresión cálculo anterior. Para dar solución a este problema se procede asumiendo la hipótesis de que todas las observaciones de un intervalo se distribuyen uniformemente dentro del mismo, por lo que es admisible sustituir todos los valores del intervalo por su marca de clase. Si se opta por esta solución, entonces es posible hacer uso de la expresión dada en *(2.4)* para el cálculo de la media aritmética.

Ejemplo 2: *Si la para la empresa anterior, en lugar de considerar las ventas en 25 días, consideramos las ventas en 300 días, entonces los datos serían los de la tabla siguiente:*

Ventas en pesetas (miles de €)	x_i	n_i
10,0 –12,0	*11,0*	*40*
12,0 – 13,0	*12,5*	*60*
13,0 – 14,0	*13,5*	*110*
14,0 – 15,0	*14,5*	*50*
15,0 – 16,0	*15,5*	*30*
16,0 – 18,0	*17,0*	*10*

Ahora la nueva media sería:

$$\bar{x} = \frac{(11)(40) + (12{,}5)(60) + (13{,}5)(110) + (14{,}5)(50) + (15{,}5)(30) + (17{,}0)(10)}{300} = 13{,}45 \quad miles \quad €$$

Esta forma de obtener la media aritmética implica que no se calcula de forma exacta, pues la frecuencia no tiene porque distribuirse de manera uniforme dentro del intervalo, por lo que al proceder de esa forma se comete un error que se le conoce como error de agrupamiento. Este viene condicionado por el número de intervalos que se estén considerando, así como por el tamaño de la población. Veamos este concepto de forma numérica.

Ejemplo 3. *Obtener la media de la variable del Ejemplo 1 pero con los datos agrupados según los intervalos de Ejemplo 2.*

En este caso las 25 observaciones se presentan en la tabla siguiente:

Ventas en pesetas (miles de €)	x_i	n_i
10,0 –12,0	*11,0*	*2*
12,0 – 13,0	*12,5*	*3*
13,0 – 14,0	*13,5*	*9*
14,0 – 15,0	*14,5*	*6*
15,0 – 16,0	*15,5*	*4*
16,0 – 18,0	*17,0*	*1*

Ahora, la media aritmética sería:

$$\bar{x} = \frac{(11)(2) + (12{,}5)(3) + (13{,}5)(9) + (14{,}5)(6) + (15{,}5)(4) + (17{,}0)(1)}{25} = 13{,}88 \quad miles \quad €$$

Como puede apreciarse el valor de la media ha cambiado, pasando de 13,82 a 13,88. La diferencia entre ambos es el error de agrupamiento que se ha cometido como consecuencia de trabajar con datos agrupados en intervalos.

Si lo que se persigue es obtener la media de una variable en la que los valores de la misma no tienen todos ellos la misma importancia o significación, entonces se procede a obtener la **media aritmética ponderada**, en la que cada valor de la esa variable se multiplica por sus respectivo peso o ponderación (w_i) que refleja la importancia de ese valor, pero que no es su frecuencia. Si la suma de esos productos la dividimos por la suma de las ponderaciones, lo que se obtiene es la media aritmética ponderada.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \qquad (2.5)$$

Ejemplo 4*. Un alumno ha realizado un examen que constaba de cinco preguntas. En cada una de ellas ha obtenido las siguientes puntuaciones: 5; 6,5; 7; 8 y 7,5. Obtenga la nota final del examen si las ponderaciones de esas preguntas son: 0,1; 0,25; 0,15; 0,25 y 0,25.*

$$\bar{x} = \frac{(5)(0{,}1)+(6{,}5)(0{,}25)+(7)(0{,}15)+(8)(0{,}25)+(7{,}5)(0{,}25)}{1} = 7{,}05$$

frente a una media aritmética simple de 6,8.

2.5.1.1 Propiedades de la media aritmética.

1ª La suma de las desviaciones de los valores de la variable respecto de la media vale cero. Es decir: $\sum_{i=1}^{k} (x_i - \bar{x}) n_i = 0$

La demostración de esta propiedad es como sigue:

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

2ª La media aritmética no varía si todas las frecuencias de su distribución se mul tiplican o dividen por una constante.

Llamemos *C* a constante por la que se van a multiplicar todas la frecuencias, de tal forma que $n_{Ci} = Cn_i$. En tal caso la media será:

$$\bar{x}_C = \frac{\sum_{i=1}^{k} x_i n_{Ci}}{N_C} = \frac{\sum_{i=1}^{k} x_i Cn_i}{CN} = \frac{C\sum_{i=1}^{k} x_i n_i}{CN} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \bar{x}$$

3ª La suma de las desviaciones al cuadrado de los valores de la variable respecto de una constante C cualquiera se hace mínima, $S(C) = \sum_{i=1}^{k} (x_i - C)^2$ *, cuando esa constante C coincide con la media aritmética (Teorema de Konig):*

Para demostrar esta propiedad basta con minimizar esa suma de desviaciones al cuadrado con respecto a *C*. El valor de esa constante *C* será el que anule la primera derivada y haga que la segunda sea positiva.

$$\frac{dS(C)}{dC} = \frac{d\left(\sum_{i=1}^{k} (x_i - C)^2\right)}{dC} = 2\sum_{i=1}^{k} (x_i - C)(-1) = 0$$

A partir de este resultado se obtiene que:

$$\sum_{i=1}^{k} x_i = NC$$

de forma que si se dividen ambos miembros de la igualdad por *N* queda demostrada la tercera propiedad de la media.

4ª Si a todos los valores de una variable le sumamos una constante C, la media aritmética queda aumentada en dicha constante. Es decir, la media se ve afecta por cambios de origen en la variable.

Para demostrar esta propiedad vamos a definir una nueva variable $y_i = x_i + C$. Ahora se obtendrá la media de y_i.

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum (x_i + C) n_i}{N} = \frac{\sum x_i n_i + NC}{N} = \bar{x} + C$$

5ª Si todos los valores de una variable los multiplicamos por una constante C, la media aritmética queda multiplicada por dicha constante. Es decir, la media se ve afecta por cambios de escala en la variable.

Al igual que en caso anterior, definamos una nueva variable $y_i = Cx_i$. Ahora se obtendrá la media de y_i.

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum Cx_i n_i}{N} = \frac{C\sum x_i n_i}{N} = C\bar{x}$$

6ª Si de un conjunto de valores obtenemos dos o más subconjuntos disjuntos, la media aritmética de todo el conjunto se relaciona con todas las medias aritméticas de los diferentes subconjunto disjuntos.

Supongamos que agrupamos las N observaciones en K subconjuntos disjuntos:

$$\left(x_{11}, x_{12}, \ldots, x_{1N_1}\right), \left(x_{21}, x_{22}, \ldots, x_{2N_2}\right), \ldots\ldots\ldots\ldots, \left(x_{K1}, x_{K2}, \ldots, x_{KN_K}\right)$$

Ahora la media es:

$$\bar{x} = \frac{\left(x_{11} + x_{12} + \ldots + x_{1N_1}\right) + \left(x_{21} + x_{22} + \ldots + x_{2N_2}\right) + \ldots\ldots + \left(x_{K1} + x_{K2} + \ldots + x_{KN_K}\right)}{N_1 + N_2 + \ldots\ldots\ldots\ldots + N_K} =$$

$$= \frac{\sum_{i=1}^{N_1} x_{1i} + \ldots + \sum_{i=1}^{N_K} x_{Ki}}{N_1 + N_2 + \ldots\ldots\ldots\ldots + N_K} = \frac{\bar{x}_1 N_1 + \ldots + \bar{x}_K N_K}{N_1 + N_2 + \ldots\ldots\ldots\ldots + N_K}$$

7ª La media es el centro de gravedad de la distribución.

Ventajas e inconvenientes de la media aritmética

Las principales ventajas son las siguientes:

1ª Hace uso de todos los valores para su cálculo

2ª Se puede calcular siempre.

3ª Es única.

Frente a estas ventajas el principal inconveniente es que se trata de un promedio sensible a los valores extremos de la variable, lo que en algunas ocasiones invalida su utilidad.

2.5.2 La media geométrica.

Sea una distribución de frecuencias (x_i, n_i). La media geométrica G se define como la raíz N-ésima del producto de los N valores de la distribución.

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \ldots\ldots x_n^{n_n}} = \sqrt[N]{\prod_{i=1}^{N} x_i^{n_i}} \qquad (2.6)$$

La media geométrica goza de la propiedad de que su logaritmo es igual a la media aritmética de los logaritmos de los valores de la variable. También la media geométrica es siempre menor o igual que la media aritmética.

Ejemplo 5. *Las tasas de crecimiento de la economía de un país durante diez años son las que aparecen en la tabla siguiente:*

Tasas (X) en %	***Años n_i***
1	*2*
2	*2*
3	*3*
4	*2*
5	*1*

Obtener la tasa media anual de crecimiento.

Para esta variable, y dada su naturaleza, el promedio más adecuado es la media geométrica.

$$G = \sqrt[10]{\left((0{,}01)^2 (0{,}02)^2 (0{,}03)^3 (0{,}04)^2 (0{,}05)^1\right)} = 0{,}02475$$

Se trata de un promedio que, para su cálculo, al igual que la media aritmética, hace uso de toda la información de la variable. Sin embargo es menos sensible a los valores extremos de lo que lo es la media aritmética. Frente a estas ventajas o virtudes, este nuevo promedio tiene algunas limitaciones. Entre ellas destacaremos: a) es menos intuitivo que la media aritmética; b) su cálculo no es tan inmediato; c) en ocasiones no queda determinada. Si algún valor de la variable es nulo, entonces G se anula. Si la variable toma valores negativos este promedio da problemas.

La media geométrica se utiliza especialmente para promediar porcentajes, tasas, números índices, etc., y siempre que la variable presente variaciones acumulativas.

Ejemplo 6. *La población de una determinada provincia durante los años que se indica fue la siguiente:*

Año	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*
Población	*375,1*	*385,7*	*390,6*	*410,5*	*430,3*	*450,7*

Obténgase la tasa de crecimiento medio anual.

Para obtener la tasa pedida se podría proceder de forma distintas. Supongamos en primer lugar que calculamos inicialmente las tasas de crecimiento anuales. Estas tasas[1], que se definen como: $\frac{P_t}{P_{t-1}} - 1$ *, son las que aparecen en la siguiente tabla:*

Año	*1996-95*	*1997-96*	*1998-97*	*1999-98*	*2000-99*
Tasa	*0,0282*	*0,0127*	*0,0509*	*0,0482*	*0,0474*

A partir de estas cinco tasas de crecimiento anual se podría calcular la media aritmética, pero este no sería el promedio más adecuado, dada la naturaleza de esta variable (se

[1] El concepto de tasa se expone de forma algo más amplia en el Capitulo 4 de este libro.

trata de tasas). En este caso el promedio más adecuado sería la media geométrica. Si a esa tasa media de crecimiento anual la llamamos r, entonces la misma vendría dada por:

$$r = \sqrt[5]{((0{,}0282)(0{,}0127)(0{,}0509)(0{,}0482)(0{,}0474))} = 0{,}03344$$

Otra forma de abordar el problema es la siguiente. Según se han definido las tasas de crecimiento se tiene que:

$$\frac{P_1}{P_0} = 1 + r_1$$

$$\frac{P_2}{F_1} = 1 + r_2$$

$$\frac{P_3}{P_2} = 1 + r_3$$

$$\frac{P_4}{P_3} = 1 + r_4$$

$$\frac{P_5}{P_4} = 1 + r_5$$

A partir de estas relaciones se llega a que $P_5 = P_0(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)$*. Ahora bien, si admitimos que durante todo el conjunto de años considerado la tasa de crecimiento anual ha sido la misma, entonces resulta que* $P_5 = P_0(1+r)^5$*, siendo r esa tasa de crecimiento medio anual constante. En estas condiciones, si igualamos los dos resultados tenemos que:* $(1+r)^5 = (1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)$*, por lo que finalmente se tiene que:*

$$r = \sqrt[5]{(1+r_1)(1+r_2)(1+r_3)(1+r_4)(1+r_5)} - 1$$

donde resulta que la tasa de crecimiento medio buscada se obtiene como un función de la media geométrica de las tasas de crecimiento anuales.

Este resultado se puede generalizar al caso de n periodos de tiempo, en cuyo caso se tendría:

$$r = \sqrt[n]{(1+r_1)(1+r_2)(1+r_3)(1+r_4).....(1+r_n)} - 1$$

Finalmente hay que indicar que si se procede de esta forma no es necesario calcular las tasas de crecimiento anual, pues r también se puede obtener de la forma siguiente:

$$r = \sqrt[n]{(1+r_1)(1+r_2)(1+r_3)(1+r_4).....(1+r_n)} - 1 = \sqrt[n]{\left(\frac{P_1}{P_0}\right)\left(\frac{P_2}{P_1}\right)\left(\frac{P_3}{P_2}\right)\left(\frac{P_4}{P_3}\right).....\left(\frac{P_n}{P_{n-1}}\right)} - 1 =$$

$$= \sqrt[n]{\frac{P_n}{P_0}} - 1$$

2.5.3 La mediana.

La mediana se puede definir de las siguientes formas:

1º *Es aquel valor de la distribución que ocupa el lugar central una vez los valores han sido ordenados de forma creciente, de menor a mayor*.

2º *Aquel valor de la distribución, una vez ordenada de menor a mayor, que deja a su izquierda y a su derecha el mismo número de observaciones*.

3º *El primer valor de la variable, una vez ordenados de menor a mayor, cuya frecuencia acumulada es mayor o igual que N/2*.

Con estas definiciones de la mediana *Me* lo que venimos a decir es que este promedio no se puede expresar mediante un fórmula.

Para determinar el valor de la *Me* de una distribución hay que proceder de forma distinta según se trate de distribuciones de frecuencias para variables discretas o continuas, pues en el primer caso los valores de la variable no están agrupados y sus frecuencias pueden ser unitarias o mayores que la unidad, mientras que para el segundo caso los valores se presentan siempre agrupados.

Cuando se trabaja con variables discretas con frecuencias unitarias, la mediana es el valor central una vez ordenados de menor a mayor. Si el número de observaciones fuera par, entonces la mediana sería la semisuma de los dos valores centrales.

Ejemplo 7. *Las notas de un examen de los diez alumnos de una clase son las siguientes: 3, 7, 6, 4, 8, 9, 5, 7, 4, 8. Obtener la nota mediana.*

En este caso lo primero que habría que realizar es ordenar de menor a mayor esos valores y como el número de ellos es par, entonces la mediana será la media de los valores centrales. En este caso se tendráque:

$$Me = \frac{6+7}{2} = 6{,}5$$

Si se trata de una distribución para variables discretas con frecuencias mayores que la unidad y valores no agrupados, entonces el valor de la variable que corresponde a la primera frecuencia acumulada mayor o igual que N/2 será la mediana.

Ejemplo 8. *Las notas de estadística par los cien alumnos de primero de una facultad de económicas y empresariales son los que se dan a continuación:*

Notas (x_i)	***Alumnos (n_i)***	***N_i***
0	*1*	*1*
1	*3*	*4*
2	*5*	*9*
3	*7*	*16*
4	*10*	*26*
5	*28*	*54*
6	*19*	*73*
7	*12*	*85*
8	*8*	*93*
9	*5*	*98*
10	*2*	*100*

Obtener la nota mediana.

En este caso, dado que los valores de la variable no están agrupados en intervalos, la mediana se corresponde con el primer valor de la variable cuya frecuencia acumulada es mayor o igual que N/2. La forma práctica de obtener ese valor consiste en añadir a la

tabla original una columna adicional que recoja las frecuencias acumuladas. En este caso como N/2=50, entonces la nota mediana resulta ser 5.

Cuando se trabaja con datos agrupados, entonces los procedimientos anteriores no sirven. En estos casos se busca el intervalo mediano. El mismo será el primero cuya frecuencia acumulada sea mayor o igual que *N/2*. El valor de la mediana será uno de los valores comprendidos dentro de ese intervalo. Si admitimos que la frecuencia de ese intervalo se distribuye uniformemente dentro del mismo, entonces la mediana será aquel valor que sumado al extremo inferior del intervalo acumule un frecuencia igual a *N/2*. De acuerdo con este criterio, la mediana se podrá calcular, de forma aproximada, mediante la expresión (2.7).

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i \qquad (2.7)$$

Ejemplo 9. *Obtener la mediana para los datos del Ejemplo 8 si estos se dieran agrupados de la forma siguiente:*

Notas (x_i)	***Alumnos (n_i)***	***N_i***
0-2	*9*	*9*
2-4	*14*	*23*
4-6	***50***	***73***
6-8	*20*	*93*
8-10	*7*	*100*

Para determinar el valor de la mediana en este caso habría que proceder a su cálculo aproximado mediante la expresión dada en (2.7), para lo cual hay que determinar el primer lugar el intervalo mediano, que en este ejemplo es el que se corresponde con los valores 4-6 de la variable.

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 4 + \frac{50 - 23}{50} 2 = 5{,}08$$

Como puede observarse este valor no coincide con el obtenido previamente. Ello se debe al error de agrupamiento *motivado por la falta de información que conlleva no tener los datos desagregados, lo que nos impide conocer la frecuencia de cada uno de los valores de la variable.*

2.5.3.1 Propiedades de la *Me*

1ª *La mediana hace mínima la suma de todas las desviaciones absolutas*.

$$Min \sum_i |x_i - k| n_i = \sum_i |x_i - Me|_i$$

2ª *La mediana se ve afecta por cambios de origen y cambios de escala*.

3ª *La mediana no se ve afectada si todas las frecuencias se multiplican o dividen por una misma constante*.

4ª La mediana no se influida por los valores extremos de la variable.

5ª *Para el caso de distribuciones campaniformes fuertemente asimétricas, la mediana resulta un promedio mejor que la media aritmética*.

6ª *Dado que en su cálculo no intervienen los valores extremos hace que se pueda obtener fácilmente incluso en presencia de intervalos abiertos* .

El principal inconveniente de la mediana es que, para su cálculo, no hace uso de toda la información que suministra la variable.

2 .5.4 La moda

La moda es aquel valor de la distribución que más se repite o que presenta mayor frecuencia.

De esta definición se deduce fácilmente que este promedio carec e de interés en distribuciones de frecuencias unitarias, pues en esos casos todos los valores tienen idéntica frecuencia por lo que el valor modal no existe.

Este promedio se representa por *Mo.* Al igual que la mediana, tampoco tiene formula del cálculo.

Para su determinación se procede de forma distinta según se trate de distribuciones de frecuencias no agrupadas o agrupadas.

En el primer caso se aplica simplemente la definición. Así, en términos del *Ejemplo 8* tendremos que *Mo =5*, pues ese valor de la variable es el que presenta una frecuencia mayor. En este caso la *Me* y la *Mo* coinciden.

En el segundo caso se procede de la siguiente forma. Si todos los intervalos tienen la misma amplitud, entonces se busca el de mayor frecuencia (intervalo modal) y la moda será uno de los valores contenidos en el mismo. La forma aproximada de determinar ese valor, suponiendo de nuevo equidistribución de la frecuencia dentro del intervalo, será la siguiente:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i \qquad (2.8)$$

Ejemplo 10. *Obtener la moda para los datos del Ejemplo 9:*

Notas (x_i)	***Alumnos (n_i)***
0-2	*9*
2-4	*14*
4-6	***50***
6-8	*20*
8-10	*7*

Para determinar el valor de la moda en este caso habría que proceder a su cálculo aproximado mediante la expresión dada más arriba:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i = 4 + \frac{20}{14 + 20} 2 = 5{,}176$$

Como puede observarse, y al igual que ocurría en el caso de la mediana, este valor no coincide con el obtenido previamente. Ello se debe al error de agrupamiento *motivado por la falta de información que conlleva no tener los datos desagregados, lo que nos impide conocer la frecuencia de cada uno de los valores de la variable.*

Si la amplitud de los intervalos es distinta, entonces, en lugar de buscar el intervalo de mayor frecuencia, se busca el intervalo de mayor altura y se procede de igual forma que en la situación anterior:

$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a^i \qquad (2.9)$$

Ejemplo 11. *Obtener el salario mensual más frecuente, expresado en euros, para la siguiente distribución:*

Salarios (x_i)	***Asalariados (n_i)***	***a_i***	***$h_i = n_i / a_i$***
Menos de 500	*50*	*500*	*0,1*
De 500 a 900	*70*	*400*	*0,175*
De 900 a 1200	***120***	***300***	***0,4***
De 1200 a 1800	*100*	*600*	*0,1666*
De 1800 a 2700	*50*	*900*	*0,0555*
De 2700 a 5000	*20*	*2300*	*0,0087*

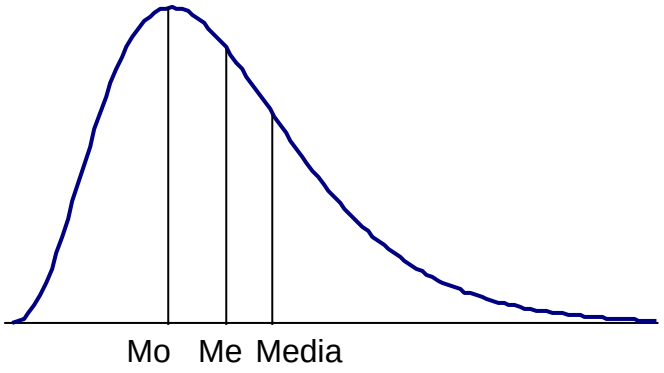
En este caso la moda será:

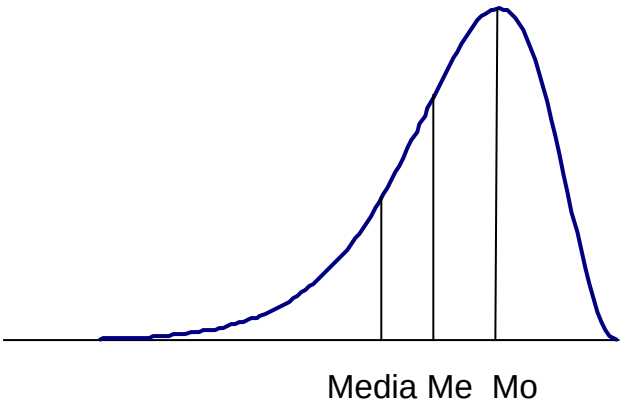
$$Mo = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} a_i = 900 + \frac{0,1666}{0,175 + 0,1666} 300 = 1046,34 €$$

Este promedio pierde interés cuando la distribución tiene más de un máximo, es decir, cuando la distribución es bimodal o multimodal, pues, como ya se ha señalado con anterioridad, la finalidad de los promedios es resumir toda la información de la distribución en un solo dato. La moda tampoco tiene gran interés cuando la distribución no es campaniforme.

En distribuciones campaniformes simétricas se cumple que la media, la mediana y la moda toman el mismo valor ($\bar{x} = Mo = Me$). Pero si no son simétricas entonces los valores son distintos, y la relación entre los mismos es la que aparece en la Figura 11.

Figura 11. Relación entre la media aritmética, la mediana y la moda en distribuciones campaniformes asimétricas.





Las propiedades de la moda son muy similares a las de la mediana, en el sentido de que se ve afectada por cambios de origen y de escala, no cambia cuando las frecuencias se multiplican o dividen por una constante y no se ve afectada por los valores extremos de la distribución.

Como inconveniente hay que señalar que no hace uso de toda la información de la tabla y que en distribuciones multimodales pierde sentido.

2.6 Las cuantilas.

Previamente se ha definido la mediana como aquel valor de la variable que divide a la distribución en dos partes iguales. Siguiendo con esta idea podríamos plantearnos buscar valores de la variable que dividan la distribución en un determinado número de partes iguales. Todos tendrían la propiedad de que entre ellos siempre queda el mismo número de observaciones. Esta forma de proceder nos lleva al concepto de cuantilas. Estas son valores de la distribución que la dividen en partes iguales, es decir, en intervalos con igual número de observaciones. A todas ellas se les conoce genéricamente como medidas de posición no centrales.

Las cuantilas de uso más frecuente son las cuartilas, las decilas y las centilas.

En consonancia con la idea de cuantilas que se ha dado previamente, las definiciones concretas de cada una de ellas serían las siguientes:

Cuartilas (Q_i): son los tres valores de la variable que dividen la distribución en cuatro partes iguales, es decir, en cuatro interval os dentro de cada cual están incluidos la cuarta parte de los valores u observaciones de la variable.

Decilas (D_i): son los nueve valores de la variable que dividen la distribución en diez partes iguales, es decir, en diez intervalos dentro de cada cual están incluidos la décima parte de los valores u observaciones de la variable.

Centilas (C_i): son los noventa y nueve valores de la variable que dividen la distribución en cien partes iguales, es decir, en cien intervalos dentro de cada cual están incluidos la centésima parte de los valores u observaciones de la variable.

Para obtener los valores de estas medidas se procederá de forma distinta según que la distribución esté agregada o no. En el caso de distribuciones discretas con valores no agregados se tendrá que:

Q_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/4*, para *i=1,2,3*.

D_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/10*, para *i=1,2,3,...9*.

C_i es el valor correspondiente a la frecuencia acumulada mayor o igual que *(iN)/100*, para *i=1,2,3,...,98,99*.

Si la distribución está agrupada en intervalos entonces la expresión general para determinar estas medidas, que es similar a la de la mediana, viene dada por:

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i} a_i \qquad (2.10)$$

en donde:

para *k=4* y *r=1,2,3*, tendremos las cuartilas
para *k=10* y *r=1,2,3,.....,9*, tendremos las decilas
para *k=100* y *r=1,2,3,.....,99* tendremos la centilas.

Ejemplo 12. *Con los datos del Ejemplo 11 obtener: a) el salario mínimo del 25% de los asalariados que más ganan; b) el salario máximo del 40% de los que menos ganan; c) el porcentaje de asalariados con salarios comprendidos entre la centila 25 y la decila 9.*

Salarios (x_i)	***Asalariados (n_i)***	***N_i***
Menos de 500	*50*	*50*
De 500 a 900	*70*	*120*
De 900 a 1200	*120*	*240*
De 1200 a 1800	*100*	*340*
De 1800 a 2700	*50*	*390*
De 2700 a 5000	*20*	*410*

a) En este caso lo que nos piden es que se obtenga la centila 75 o la tercera cuartila, pues ambos valores son iguales. Para determinar ese valor hay que determinar previamente $\frac{r}{k}N = \frac{75}{100}410 = \frac{3}{4}410 = 307,5$ *. La primera frecuencia acumulada mayor que este valor se corresponde con la del intervalo "De 1200 a 1800", por lo que el valor buscado es:*

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i}a_i = C_{75} = Q_3 = 1200 + \frac{307,5 - 240}{100}600 = 1605\ €$$

b) Ahora el valor pedido se corresponde con la centila 40 o la cuarta decila. Al igual que antes, hay que determinar previamente $\frac{r}{k}N = \frac{40}{100}410 = \frac{4}{10}410 = 164$ *. La primera*

frecuencia acumulada mayor que 164 es la correspondiente a la del intervalo "De 900 a 1200", así que el valor pedido es:

$$Q_{r/k} = L_{i-1} + \frac{\frac{r}{k}N - N_{i-1}}{n_i} a_i = C_{40} = D_4 = 900 + \frac{164 - 120}{120} 300 = 1110 \text{ €}$$

c) Según la definición de cuantilas que se ha dado, el porcentaje pedido es el 65%, que es la diferencia entre el 90% que hay por debajo de la decila 9 y el 25% que deja a su izquierda la centila 25.

2.7 Dispersión.

El proceso de reducción estadística nos ha llevado a sintetizar todos los datos de una tabla en un solo número, al que hemos llamado promedio y con el que se pretende representar a la información que hay detrás de él. Pero para que esa medida de síntesis tenga esa validez ha de ser representativa de todos los datos contenidos en ella.

La mayor o menor representatividad de esas medidas de síntesis o promedios dependerá, fundamentalmente, del grado de concentración de todos los valores de la distribución en torno a ese promedio, cuya representatividad estamos estudiando. En el caso extremo de que todos los valores fueran iguales, la media, por ejemplo, coincidiría con uno de ellos y ésta representaría perfectamente a todos. Pero esta situación extrema nunca se da, (si se diera no tendríamos variable). Lo normal es que una variable tome más de un valor. Es en estos casos donde hay que estudiar si los promedios utilizados son representativos del conjunto de valores a los que representan.

En general, el promedio más utilizado es la media aritmética. Por esta razón nos vamos a detener en definir medidas que cuantifiquen el grado de representatividad de la media. Así, diremos que la media es representativa de una distribución si los valores de la misma están muy próximos a ella. Por el contrario, si esos valores estuvieran muy dispersos o alejados, diríamos que la media no es muy representativa.

Cuando se habla de la representatividad de un promedio es más frecuente utilizar el término dispersión que el de concentración, razón por al cual en adelante usaremos esta expresión.

Para medir la dispersión de un promedio nos basaremos en el concepto de distancia o desviaciones existentes entre los valores de la distribución y el promedio que estamos utilizando. Cuanto mayores sean estas distancias o desviaciones mayor será la dispersión y menor será esta si las distancias son pequeñas.

Las medidas de dispersión se pueden clasificar en dos categorías según que sean medidas absolutas o relativas. Dentro de las primeras las más simples son las que se basan en el recorrido de la variable (recorrido y recorridos intercuartílicos). Las más elaboradas son las que se definen en términos de distancias o desviaciones de los valores de la variable respecto de algún promedio concreto (desviación estándar, variancia, desviación media). A su vez la medidas de dispersión relativas más utilizadas son el coeficiente de apertura o disparidad y el coeficiente de variación.

2.8 Medidas de dispersión absoluta.

2.8.1 Variancia y desviación estándar.

La variancia (S^2) es la medida de dispersión más conocida y utilizada de todas cuantas puedan definirse. La misma se basa en la idea de promediar las desviaciones respecto de la media aritmética. Pero el promedio que se utiliza para obtener esa medida de dispersión no es la media aritmética de las desviaciones, pues ésta sería siempre nula, por la primera propiedad de la media aritmética, que nos habla de que la suma de las desviaciones respecto de la media es siempre cero. Este inconveniente se resuelve calculando no la media aritmética de las desviaciones sino la media cuadrática, que no es otra cosa que la media de las desviaciones respecto de la media al cuadrado. A este promedio de las desviaciones se le conoce como variancia.

La variancia[2] se define como:

$$S^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2 n_i}{N} \qquad (2.11)$$

Cuanto mayor sea la variancia mayor será la dispersión de la variable y menos representativa será la media como promedio de todos los valores y viceversa. Los valores que puede tomar la variancia, dada la forma en que está definida (media cuadrática), serán siempre mayores o iguales que cero. El caso extremo se dará cuando todos los valores de la variable sean iguales, pues en tal caso todas las desviaciones serán nulas. Salvo esta situación extrema, y carente de interés para la Estadística, las desviaciones serán siempre mayores que cero y, en consecuencia, la variancia también lo será. Finalmente hay que señalar que la variancia viene expresada en las unidades de medida de la variable al cuadrado, lo que limita su interpretación.

Para salvar este último inconveniente se define la desviación estándar (S) o desviación típica, que no es más que la raíz cuadrada de la variancia tomada con signo positivo. Esta forma de definirla hace que su unidad de medida sea la misma que la de la variable.

2.8.1.1 Propiedades de la variancia.

1ª La variancia siempre será mayor o igual que cero.

Esta propiedad es evidente dada la definición de la variancia como media de sumas de desviaciones al cuadrado.

2ª La media cuadrática de las desviaciones de una variable respecto de una constante cualquiera se hace mínima cuando esa constante es la media aritmética, es decir, cuando se trabaja con la variancia. (Teorema de König).

[2] La variancia se ha definido como una media cuadrática y, en consecuencia, el denominador de la misma es el tamaño de la población. Sin embargo, resulta habitual que en los paquetes estadísticos, tales como el SPSS y otros, calculen la variancia dividiendo por (N-1). A ese cociente, para diferenciarlo del que se obtiene cuando el denominador es N, se le llama habitualmente como cuasivariancia. La explicación por la cual se utiliza como denominador N-1 en lugar de N hay que buscarla en el campo de la Inferencia Estadística, mientras que el contenido de este libro está dedicado a la Estadística Descriptiva.

La demostración de esta propiedad es inmediata a partir de la propiedad 3ª de la media aritmética.

3ª La variancia no cambia si a los valores de la variable se les suma o resta una constante (cambios de origen).

Para demostrar esta propiedad arrancaremos de la variancia de una variable *X* a la que llamaremos S_x^2 y a partir de ella definimos otra variable *Y* en la forma $Y = X + C$, donde *C* es una constante. Con esta notación vamos a obtener la variancia de *Y*.

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

4ª Si los valores de la variable los multiplicamos por una constante, la variancia queda multiplicada por el cuadrado de esa constante. (Cambio de escala).

Al igual que para la propiedad anterior, sea *X* una variable con variancia S_x^2. Definamos ahora $Y = XC$, donde *C* es una constante cualquiera. Entonces:

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

Anteriormente hemos definido la variancia mediante la expresión (2.11). Sin embargo, para calcularla se suele recurrir al siguiente desarrollo de la misma:

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x} \right)^2$$

el cual facilita[3] considerablemente su obtención.

Esta expresión es válida tanto para distribuciones no agrupadas como para las agrupadas, con la diferencia que para el primer caso da un valor exacto de la variancia de la distribución, mientras que para el segundo solo da un valor aproximado debido a los errores de agrupamiento que se comenten en este tipo de distribuciones, como ya se ha señalado en otras ocasiones. Para este segundo tipo de distribuciones x_i es la marca de clase o valor central del intervalo y se supone, como se ha hecho en otras circunstancias, que la frecuencia del intervalo se distribuye de manera uniforme entre todos los valores del mismo, pues el agrupamiento de valores en intervalos nos impide saber cual es la frecuencia verdadera de cada uno de esos valores, con la consiguiente pérdida de información.

Ejemplo 13. *Obtener la variancia de la distribución de salarios del Ejemplo 11.*

Salarios (x_i)	***Asalariados (n_i)***	***Marca de clase (x_i)***	***$x_i n_i$***	***$x_i^2 n_i$***
Menos de 500	*50*	*250*	*12500*	*3125000*
De 500 a 900	*70*	*700*	*49000*	*34300000*
De 900 a 1200	*120*	*1050*	*126000*	*132300000*
De 1200 a 1800	*100*	*1500*	*150000*	*225000000*
De 1800 a 2700	*50*	*2250*	*112500*	*253125000*
De 2700 a 5000	*20*	*3850*	*77000*	*296450000*
Tot al	***410***		***527000***	***944300000***

En primer lugar calcularemos la media como paso previo para obtener la variancia.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{N} = \frac{527000}{410} = 1285{,}4\ €$$

[3] En este caso, como en otros, se está hablando de "facilitar" el cálculo de ciertas medidas estadísticas, cuando, en realidad, hoy en día la accesibilidad y disponibilidad de paquetes estadísticos o de hojas de cálculo hace que esas preocupaciones resulten casi una cuestión del pasado.

A continuación obtendremos la variancia de esta variable:

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{944300000}{410} - (11285{,}4)^2 = 651005{,}4 \ €^2$$

Para distribuciones campaniformes y no fuertemente asimétricas se cumple que el intervalo definido por :

$$\overline{X} \pm 2S$$

contiene aproximadamente el 95% de las observaciones y el definido como:

$$\overline{X} \pm 3S$$

contiene aproximadamente el 99% de las observaciones.

Estos intervalos pueden utilizarse a su vez como medias de dispersión.

Para el ejemplo anterior, el primer intervalo contiene el 95,5% mientras que el segundo contiene el 97,25%.

2.8.2 Desviación media.

Cuando se introdujo el concepto de variancia se señaló que la media aritmética de la desviaciones no servía como indicador de dispersión porque siempre era cero. Ante tal situación se utilizó la media cuadrática para evitar que las desviaciones, positivas y negativas, se compensaran y su suma fuera nula. Otra forma de evitar esa compensación es sumar el valor absoluto de las desviaciones y dividirlo por el número de sumandos. A esta media aritmética se le conoce como desviación media. La desviación se puede definir respecto de la media, la mediana o la moda. Según el promedio que se utilice se tendría:

$$D_x = \frac{\sum_{i=1}^{N} |x_i - \bar{x}| n_i}{N} \qquad (2.12)$$

$$D_{Me} = \frac{\sum_{i=1}^{N} |x_i - Men_i| n_i}{N} \qquad (2.13)$$

$$D_{Mo} = \frac{\sum_{i=1}^{N} |x_i - Mo| n_i}{N} \qquad (2.14)$$

2.8.3 Otras medidas absolutas de dispersión.

Además de las mediadas de dispersión señaladas con anterioridad, también se puede utilizar el recorrido, definido como $R_e = x_n - x_1$, y el recorrido intercuartílico que se define como $R_I = Q_3 - Q_1$. Estas medidas, como cualquier otro tipo de recorrido, tienen menor utilidad para medir la dispersión que las vistas con anterioridad. Además, si lo que se busca son medidas de dispersión que indiquen la representatividad de los promedios, éstas no nos sirven pues en su definición no están implicados.

Ejemplo 14.- *La notas de un examen realizado por 200 alumnos son las se dan en la tabla siguiente:*

Puntuación (x_i)	Alumnos (n_i)
2	4
3	6
4	10
4.5	14
5	28
6	35
6.5	30
7	20
7.5	20
8	15
8.5	10
9	3
9.5	3
10	2
	200

Obtenga la medidas de dispersión absoluta más habituales para esta distribución.

Para la obtención de estas medidas lo primero que debe hacerse es ampliar el contenido de la tabla anterior en la forma que se indica a continuación:

Puntuación (x_i)	**Alumnos (n_i)**	$x_i n_i$	N_i	$x_i^2 n_i$	$\lvert x_i - \bar{x}\rvert n_i$	$\lvert x_i - Me\rvert n_i$	$\lvert x_i - Mo\rvert n_i$
2	4	8	4	16	16,9	18	16
3	6	18	10	54	19,3	21	18
4	10	40	20	160	22,2	25	20
4.5	14	63	34	283,5	24,1	28	21
5	28	140	62	700	34,2	42	28
6	35	210	97	1260	7,8	17,5	0
6.5	30	195	127	1267,5	8,3	0	15
7	20	140	147	980	15,6	10	20
7.5	20	150	167	1125	25,6	20	30
8	15	120	182	960	26,7	22,5	30
8.5	10	85	192	722,5	22,8	20	25
9	3	27	195	243	8,3	7,5	9
9.5	3	28,5	198	270,75	9,8	9	10,5
10	2	20	200	200	7,6	7	8
	200	**1244,5**		**8242,3**	**249,2**	**247,5**	**250,5**

Con toda esta información se tiene que:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{1244,5}{200} = 6,2$$

Mo = 6 *Me = 6,5*

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left|\bar{x}\right|^2 = \frac{8242,3}{200} - (6,2)^2 = 2,49$$

$$D_{\bar{x}} = \frac{\sum_{i=1}^{N} \left|x_i - \bar{x}\right| n_i}{N} = \frac{249,2}{200} = 1,246$$

$$D_{Me} = \frac{\sum_{i=1}^{N} \left|x_i - Me\right| n_i}{N} = \frac{247,5}{200} = 1,238$$

$$D_{Mo} = \frac{\sum_{i=1}^{N} |x_i - Mo| n_i}{N} = \frac{250,5}{200} = 1,253$$

2.9 Medidas relativas de dispersión.

Todas las medidas de dispersión definidas anteriormente tienen un inconveniente en común. Todas ellas vienen expresadas en la misma unidad de medida que la variable (o al cuadrado, caso de la variancia) y todas ellas son sensibles a los cambios de escala. Esto hace que no sea posible realizar comparaciones entre la dispersión de dos distribuciones distintas. Pero, incluso aunque las unidades fueran las mismas, bastaría con que los promedios sean muy distintos para que esa comparación tampoco fuera posible.

Una forma de eliminar esos problemas es haciendo que las mismas sean adimensionales. El resultado es un conjunto de medidas de dispersión relativas. Dentro de esta categoría se engloban el coeficiente de disparidad o apertura y el coeficiente de variación de Pearson.

El primero se define como $A = x_n/x_1$. Tiene a su favor que es muy fácil de calcular. Por el contrario, presenta el grave inconveniente de que en su definición no recoge ningún promedio y que, además, está fuertemente influido por los valores extremos de la distribución, sin tener en cuenta cual es la dispersión de los demás valores.

El segundo se define como el cociente de la desviación estándar entre la media aritmética.

$$CV = \frac{S}{\bar{x}} \qquad (2.15)$$

Este coeficiente es adimensional y será siempre mayor o igual que cero[4]. La gran limitación del mismo viene por el hecho de que cuando la media es cero entonces carece de sentido. También tiene el inconveniente de que no es invariante ante cambios de origen, pues aunque la desviación estándar lo sea, la media no lo es. En cambio si que es invariante a cambios de escala. Otra limitación de este coeficiente se deriva de que no está acotado por la derecha, lo que no permite afirmar cuando es la media representativa o deja de serlo. Como regla aproximada se puede seguir el criterio de que si el coeficiente variación es menor de 0,5, entonces se puede hablar de que la media es representativa, mientras que es si es mayor que ese valor habría que cuestionar la representatividad de ese promedio. En cualquier caso, si *CV >1*, la media es poco o nada representativa.

Ejemplo 15. *Para dos distribuciones de renta distintas, una expresada una en miles de pesetas (X) y otra en euros (Y), se tienen los siguientes resultados:*

$$\bar{x} = 2200 \quad miles \quad de \quad ptas \quad S_x = 1000 \quad miles \quad de \quad ptas.$$
$$\bar{y} = 15000 \quad € \qquad\qquad S_y = 10000 \quad €$$

Analizar la dispersión y representatividad de las medias.

Este es una caso claro donde ni las medias ni las desviaciones estándares son comparables, pues estamos trabajando con unidades de medida distintas. Todavía si las medias hubieran sido iguales y las unidades de medida también, solo en ese caso las desviaciones estándares serían comparables y se podría determinar qué media es más representativa. Pero como se ha visto esta no es la situación. Ante estas circunstancias se hace necesario recurrir al coeficiente de variación de Pearson, que elimina los inconvenientes (unidades de media y distintas medias) señalados más arriba. Para este ejemplo los coeficientes son:

[4] Cuando se afirma que el CV es mayor o igual que cero se está dando por sentado que la media de la distribución es siempre mayor que cero (cuando la media es cero este coeficiente carece de sentido e interés). Este es el caso habitual cuando se trabajan con variables de tipo económico, aunque no siempre ha de ser así. Piénsese en variables del tipo beneficios empresariales, rendimientos de acciones y otras similares. En estos casos lo que suele hacerse es tomar el valor absoluto de la media o cuestionarse si la media aritmética es el promedio más adecuado en esa situación.

$$CV_x = \frac{S_x}{\bar{x}} = \frac{1000}{2200} = 0{,}4545 \qquad CV_y = \frac{S_y}{\bar{y}} = \frac{10000}{15000} = 0{,}6666$$

En este caso la media de la distribución expresada en pesetas es más representativa, pues la dispersión relativa de esta variable es menor.

Ejemplo 16.- *La distribución de los hogares de un determinado barrio, según el tamaño de los mismos, es la que muestra la tabla siguiente:*

Tamaño de hogares x_i	***Nº de hogares*** n_i
1	*40*
2	*70*
3	*110*
4	*90*
5	*48*
6	*42*
7	*40*
8	*35*
9	*20*
10	*5*
Total	***500***

Determínese:

1º ¿Cuál es el número medio de personas por hogar?

2º ¿Cuál es el tipo de hogar más frecuente?

3º Si solo hubiera plazas de aparcamiento para el 50% de las hogares y éstas se asignaran a las de mayor tamaño, ¿a partir de qué tamaño de hogar se le asignarían plaza de garaje?

4º Si en otro barrio el coeficiente de variación e 1, ¿en cuál de los barrios la media resulta más representativa?

1º En este apartado lo que se pide es que se calcule la media aritmética. El valor de la misma es:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{2152}{500} = 4{,}3 \text{ personas}$$

2º Ahora lo que se pide es la moda, el valor de la variable que más se repite. Como se trata de una variable discreta con los valores sin agrupar, la obtención de la moda es inmediato. El valor más frecuente para este ejemplo es Mo = 3 personas.

3º El tamaño de hogar que se nos pide es el mediano, pues sabemos que la mediana es el valor de la variable que divide a la distribución en dos partes iguales, es decir, a su izquierda queda el mismo número de observaciones que a la derecha. Este promedio es el que se corresponde con la primera frecuencia acumulada mayor o igual que 250 (N/2). Ese valor es Me = 4 personas.

4º Para responder a esta cuestión se hace necesario obtener previamente la desviación estándar de esta distribución. Como sabemos $S = y\sqrt{S^2}$

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{11782}{500} - (4{,}304)^2 = 5{,}04 \text{ personas}^2$$

Por lo que S = 2,24 y el coeficiente de variación de esta distribución es

$$CV = \frac{S}{\bar{x}} = \frac{2{,}24}{4{,}3} = 0{,}52$$

que, comparado con el del otro barrio, nos lleva a que el tamaño de la familia en el barrio objeto de estudio está menos disperso, lo que hace que su media sea más representativa.

Ejemplo 17. *Responda a las cuestiones del Ejemplo 16 y compare los resultados si los datos se hubieren agregado en la forma siguiente*:

Tamaño de hogar x_i	***Nº de hogares*** n_i
De 0 a 2	*110*
De 2 a 4	*200*
De 4 a 6	*90*
De 6 a 8	*75*
De 8 a 10	*25*
Total	***500***

Ahora, al tener los valores de la variable agrupados en intervalos, se hace necesario calcular las marcas de clase con vistas a obtener tanto la media con la desviación estándar. Por ello lo primero que es conveniente hacer es ampliar la tabla con otras columnas adicionales.

Tamaño de hogar	n_i	x_i	$x_i n_i$	N_i	$x_i^2 n$
De 0 a 2	*110*	*1*	*110*	*110*	*110*
De 2 a 4	*200*	*3*	*600*	*310*	*1800*
De 4 a 6	*90*	*5*	*450*	*400*	*2250*
De 6 a 8	*75*	*7*	*525*	*475*	*3675*
De 8 a 10	*25*	*9*	*225*	*500*	*2025*
Total	***500***		***1910***		***9860***

A partir de esta tabla se tiene que:

$$\bar{x} = \frac{\sum_i x_i n_i}{N} = \frac{1910}{500} = 3{,}82 \text{ personas}$$

Como puede observarse, el nuevo valor de la media no coincide con el anterior. La diferencia es lo que se conoce como error de agrupamiento. El valor verdadero es 4,3. Sin embargo, cuando se agrupa se pierde información y el coste de esa pérdida es cometer un error. En este ejemplo es fácil ver donde está la naturaleza de este error. Basta con fijarse en la columna tercera de ésta última tabla para comprobar que los hogares de tamaño 1 no son 110, sino 40, o que los de tamaño 3 no son 200 y, así, sucesivamente.

Estos resultados muestran lo delicado que resulta agrupar los valores de las variables discretas y más cuando el número de los mismos no es muy elevado, como es este caso.

Esta agrupación nos lleva a que los demás promedios y las medidas de dispersión que se van a calcular se vean afectadas también por el error de agrupamiento, como será fácil de comprobar sin más que observar los resultados que se dan a continuación.

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i = 2 + \frac{90}{110+90} 2 = 2{,}9 \text{ personas}$$

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} a_i = 2 + \frac{250 - 110}{200} 2 = 3{,}4 \text{ personas}$$

$$S^2 = \frac{\sum_i x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = \frac{9860}{500} - (3{,}82)^2 = 5{,}13 \text{ personas}^2$$

$$CV = \frac{S}{\bar{x}} = \frac{2{,}26}{3{,}82} = 0{,}59$$

2.10 Tipificación de una variable

Relacionado con los conceptos que se han definido en apartados anteriores aparece el de tipificación (estandarización o normalización) de los valores de una variable. La tipificación no es una medida de dispersión ni un promedio. Se trata de un procedimiento que facilita la comparación entre los valores de dos distribuciones distintas.

Se dice que *una variable está tipificada, normalizada o estandarizada cuando a sus valores se les resta su media aritmética y se les divide por su desviación estándar* . El

resultado de esa operación es otra variable que tiene de media cero y de variancia la unidad. A las variables tipificadas se les representa habitualmente por la letra Z, de forma que si *X* es una variable con media $\bar{x}$ y con desviación estándar S_x, entonces, con la definición dada más arriba, se tiene que:

$$Z = \frac{X - \bar{x}}{S_x} \qquad (2.16)$$

La utilidad de la tipificación es doble. Por un lado nos lleva a una distribución muy especial, aquella en la que su media vale cero y su variancia es la unidad.

En segundo lugar, y como se ha indicado más arriba, permite realizar comparaciones entre valores de distintas distribuciones cuando éstas tienen medias y variancias diferentes.

Para dejar más clara esta utilidad se podría recurrir al siguiente ejemplo. Imaginemos que deseamos comparar las notas de un alumno en dos exámenes distintos. En uno de ellos ha obtenido un 6 y la nota media de ese examen para todos los alumnos ha sido de 7, con una desviación estándar de 2. En el otro examen ha obtenido una nota de 5, siendo la media para el conjunto de todos los que se examinaron de 4,5 con una desviación estándar de 1. Ante esta situación, si nos atenemos a las dos notas, la conclusión inmediata es que ha obtenido un mejor resultado en el primer examen. Pero no siendo falsa esta conclus ión, lo cierto es que no son comparables esos dos valores, pues proceden de dos poblaciones totalmente distintas, con distintas medias y variancias, por lo que no son comparables. Para que esas comparaciones puedan realizarse de forma correcta hay que hom ogeneizar las dos distribuciones, o lo que es igual, hay que eliminarles sus características propias y reducirlas a un único patrón. Esto se consigue tipificando las variables. En nuestro caso concreto tendríamos los siguientes resultados:

$$Z = \frac{X - \bar{x}}{S_x} = \frac{6 - 7}{2} = -0{,}5 \qquad Z = \frac{X - \bar{x}}{S_x} = \frac{5 - 4{,}5}{1} = 0{,}5$$

La conclusión a la que se llega ahora es que la nota relativa del primer examen es inferior a la del segundo, pues si bien en el primer caso obtuvo un 6, resulta que esa nota es inferior a la media del curso, mientras que en el segundo de los exámenes su nota, aunque más baja que la primera, está por encima de la media y como además la dispersión es aún más pequeña ello lleva a que muy pocos alumnos obtuvieron notas por encima de la media, lo que confirma que el resultado relativo del segundo examen es mejor que el correspondiente al primero.

Otra situación donde queda muy clara la utilidad de la tipificación es cuando se quiere comparar el poder adquisitivo de la renta de dos personas que viven en lugares distintos. Supongamos que la renta anual de un ciudadano que reside en el país A es de 20000 € anuales, siendo la renta media de ese país de 15000 € y la desviación estándar de 10000 €. En otro país B, la renta de un residente es 25000 €, mientras que la renta media de este segundo país es 30000 € y la desviación estándar es 25000 €. Con estos datos la cuestión que se plantea es saber cual de esos dos ciudadanos tiene una renta relativa más elevada. De nuevo hay que recurrir a la tipific ación para que las distribuciones de renta sean comparables u homogéneas. Ahora:

$$Z_A = \frac{X_A - \overline{x_A}}{S_A} = \frac{20000 - 15000}{10000} = 0{,}5 \qquad Z_B = \frac{X_B - \overline{x_B}}{S_B} = \frac{25000 - 30000}{25000} = -0{,}2$$

de lo que se deduce que al renta relativa del residente en A es mayor que la del residente en B.

2.11 Medidas de forma. Asimetría y curtosis.

En los epígrafes anteriores hemos estudiado distintas características de una distribución. El cuarto lo hemos dedicado a estudiar la forma de una distribución y en el mismo se describieron algunos modelos de distribuciones que se repiten con cierta frecuencia. Ello nos llevó a hablar de distribuciones campaniformes, en forma de *U* y en forma de *J*. Pues bien, este epígrafe intenta ahondar en esta cuestión, pero ahora en lugar de utilizar la representación gráfica para hablar de la forma de una distribución haremos uso de ciertas medidas sintéticas. Estas medidas servirán para determinar el grado de simetría y el de apuntamiento de la distribución (curtosis). Esta segunda característica tiene un carácter

relativo en la medida que se da en relación con otra distribución que se usa como modelo de referencia (la distribución normal).

2.11.1 Simetría de una distribución.

Diremos que una distribución es simétrica respecto de un determinado eje si existe el mismo número de valores a cada lado de ese eje, equidistantes respecto del mismo dos a dos y tales que para cada par de valores equidistantes a dicho eje tengan la misma frecuencia. En caso contrario se dice que la distribución no es simétrica. El eje de simetría que suele utilizarse como referente principal es el valor de la media aritmética.

Si tomamos la media como eje de simetría y la distribución es simétrica, entonces las desviaciones de los valores de la distribución con respecto a la media serán positivas y negativas y habrá tantas positivas como negativas, de tal manera que su suma será nula. Pero esa suma será siempre nula, incluso en aquellos casos de distribuciones no simétricas (recordemos que la suma de las desviaciones con respecto a la media es cero, incluso cuando la distribución no es simétrica). En consecuencia esta suma de desviaciones no nos permite determinar si predominan las diferencias positivas o negativas, pues si las diferencias positivas fueran mayores que las negativas ello significaría que la mayor parte de la distribución se encuentra a la derecha de la media (distribución asimétrica a la derecha o positiva), mientras que si fueran las diferencias negativas las que predominaran sobre las positivas entonces nos encontraríamos con que la mayor parte de los valores de la distribución estaría en la cola izquierda de la misma, a la izquierda de la media (distribución asimétrica a la izquierda o negativa).

Vemos pues que, una medida que se base en la suma de esas diferencias y que mantenga el signo de las que dominan será un buen indicador de la presencia o no de simetría.

Si las diferencias respecto de la media las elevamos a una potencia par evitamos que su suma sea nula, pero el signo de esas diferencias será siempre positivo, con lo cual este tipo de medias no nos sirve. En cambio si la potencia que utilizamos es impar entonces lo que conseguimos es que la suma no se anule y que además las desviaciones mantengan su signo. De todas las posibles potencias impares la más simple es la tercera. Así nuestro coeficiente de asimetría vendría dado por:

$$m_3 = (\Lambda_i(x_i - x)^3 n_i)/N \qquad (2.17)$$

de forma que si:

$m_3 = 0$ la distribución es simétrica
$m_3 > 0$ la distribución es asimétrica a la derecha o positiva
$m_3 < 0$ la distribución es asimétrica a la izquierda o negativa

Esta medida de asimetría viene expresada en las unidades de la variable al cubo y además no es invariante a los cambios de escala. Estos son dos inconvenientes que se pueden salvar fácilmente si la dividimos por el cubo de la desviación estándar. A la medida resultante se le conoce como coeficiente de asimetría de Fisher. El mismo viene dado por:

$$\gamma_1 = m_3/S^3 \qquad (2.18)$$

Como S es siempre positiva, resulta que el signo de γ_1 es el m_3 con lo que:

$\psi_1 = 0$ la distribución es simétrica
$\psi_1 > 0$ la distribución es asimétrica a la derecha o positiva
$\psi_1 < 0$ la distribución es asimétrica a la izquierda o negativa

La principal limitación de este coeficiente es que para distribuciones simétricas el mismo vale siempre cero, pero el recíproco no es siempre cierto. En estas circunstancias habría que recurrir a la representación gráfica.

Para el caso de distribuciones campaniformes unimodales y simétricas sabemos que la media y la moda coinciden. Pero sin en distribuciones campaniformes unimodales esos dos promedios no coincidieran, entonces ello sería indicativo de que la distribución no es simétrica. Basándose en este principio K. Pearson definió el siguiente coeficiente de asimetría:

$$A = \frac{\bar{x} - Mo}{S} \qquad (2.19)$$

cuya interpretación es la siguiente:

A = 0 la distribución es simétrica
A > 0 la distribución es asimétrica a la derecha o positiva
A < 0 la distribución es asimétrica a la izquierda o negativa

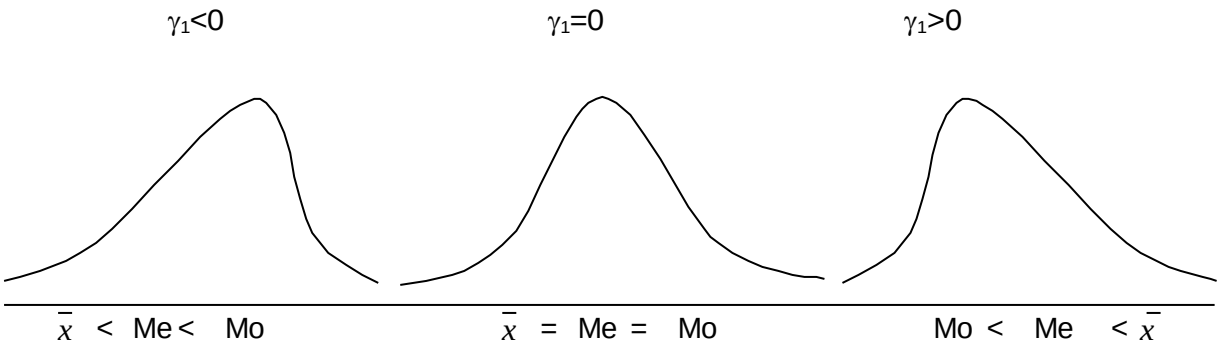
Para este tipo de distribuciones campaniformes, si además no son muy asimétricas, se cumple la siguiente relación aproximada:

$$\bar{x} - Mo \cong 3(\bar{x} - Me) \qquad (2.20)$$

por lo que también, de forma aproximada, se puede decir que:

$$A \cong \frac{3(\bar{x} - Me)}{S} \qquad (2.21)$$

Figura 12. Análisis gráfico de la simetría.



2.11.2 Apuntamiento de una distribución (Curtosis)

Esta característica de una distribución es la que tiene un menor grado de interés de todas las que se han considerado hasta el momento. La misma hace referencia al mayor o menor grado de apuntamiento de una distribución y se define en términos de la distribución normal. Para la distribución normal se verifica que $m_4/S^4=3$, donde:

$$m_4 = \frac{\sum_{i=1}^{N} \left(x_i - \underline{x}\right)_4 n_i}{N} \qquad (2.22)$$

A partir de esta relación se define el coeficiente de curtosis o apuntamiento de la forma siguiente:

$$\psi_2 = (m_4/S^4) - 3 \qquad (2.23)$$

de forma que si:

- $\psi_2 = 0$ distribución mesocúrtica (normal)
- $\psi_2 > 0$ distribución leptocúrtica
- $\psi_2 < 0$ Distribución platicúrtica

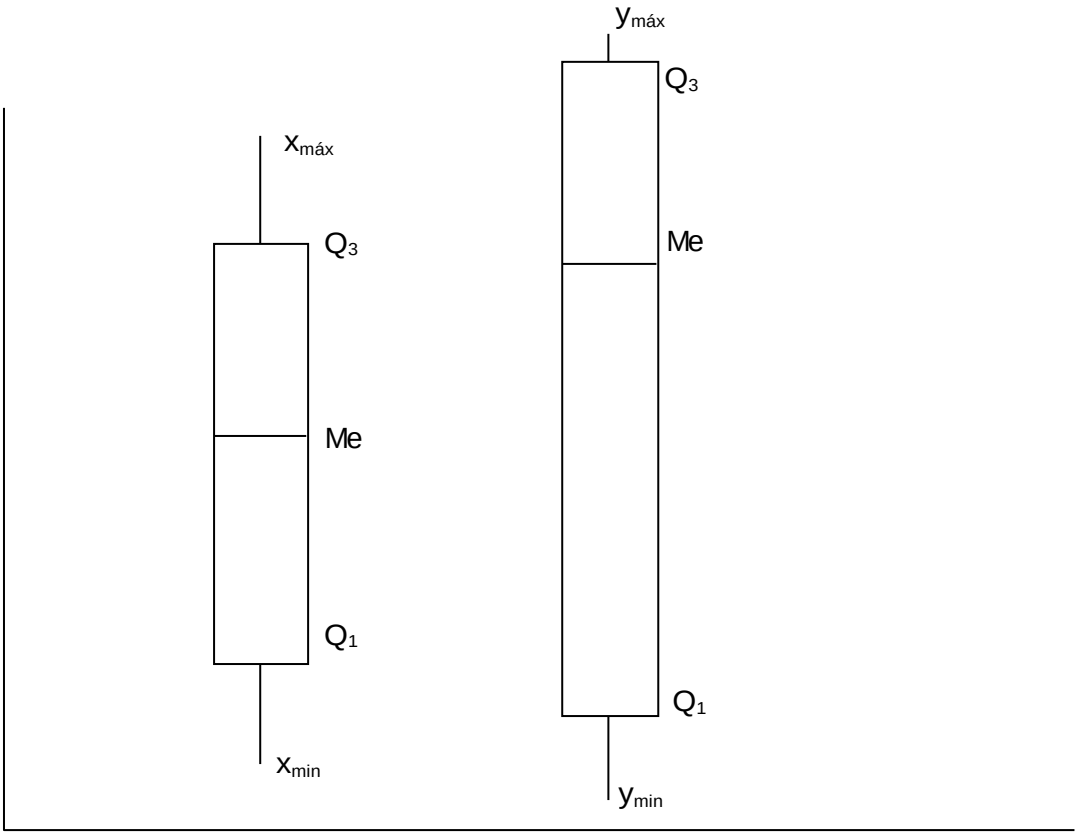
2.12 Análisis gráfico de la dispersión y la asimetría.

Estas dos características de una distribución las hemos estudiado hasta ahora de forma separada mediante medidas específicas o inspeccionando la representación gráfica de la distribución. Si embargo es posible, mediante lo que se conoce como diagramas de caja, realizar un análisis simultáneo de ambas características. Estos diagramas se basan en el recorrido de la distribución, la mediana y los valores de la primera y tercera cuartila.

De forma genérica, en la Figura 13 se han representado dos de estos diagramas, uno para la distribución de una variable *X* y otro para la de una variable *Y* cualesquiera. Para cada una de esas distribuciones se han representado los valores extremos, las medianas y la primera y tercera de las cuartilas. El resultado son dos representaciones gráficas que nos informan de lo siguiente: al variable *X* tiene menor dispersión que la *Y*, pues su

recorrido es menor y además el cincuenta por ciento de las observaciones centrales (las comprendidas entre Q_1 y Q_3) se concentran en un intervalo menor. Por otro lado, la distribución de *X* es casi simétrica, pues la mediana está en centro de los recorridos intercuartílico y total, mientras que en el caso de *Y* se encuentra más próxima a los valores grandes, lo que nos indica que se trata de una distribución asimétrica a la izquierda.

Figura 13. Diagramas de caja.



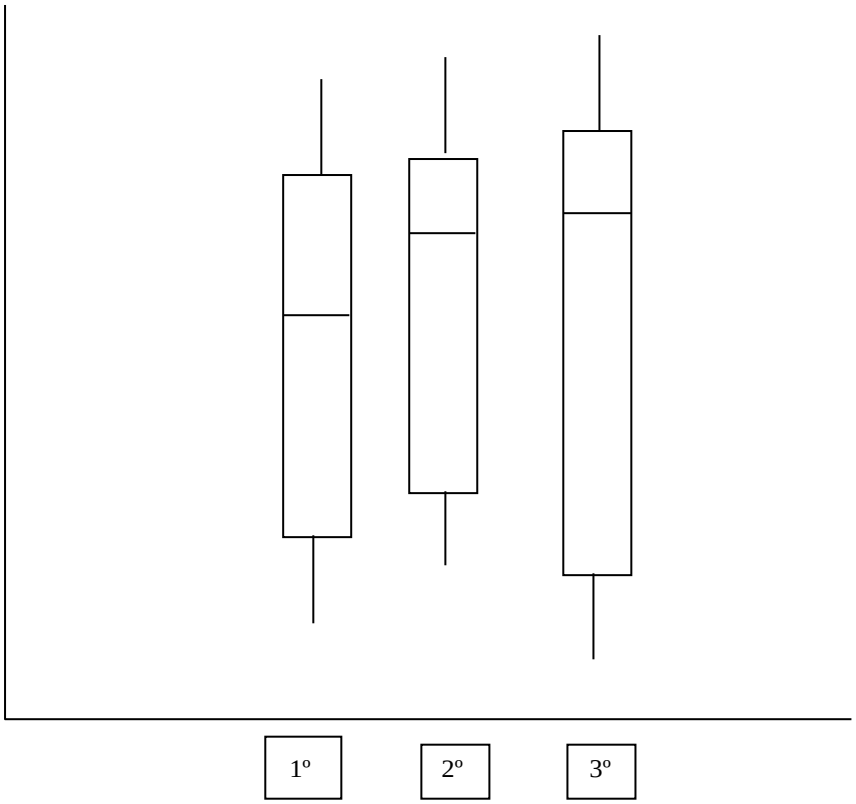
***Ejemplo 18**. En la tabla siguiente se dan los resultados de un examen en tres cursos distintos. Obtener los correspondientes diagramas de caja y estudiar la dispersión y simetría de esas distribuciones.*

Primero	***Segundo***	***Tercero***
4,7	*5,6*	*4,3*
5,2	*5,9*	*4,8*
5,2	*5,9*	*5,0*
5,7	*6,1*	*5,5*
6,3	*6,7*	*6,1*
6,4	*6,9*	*6,7*
6,9	*7,3*	*7,2*
7,1	*7,6*	*7,8*
7,2	*7,6*	*8,0*
7,2	*8,0*	*8,0*
7,8	*8,3*	*8,3*
8,1	*8,3*	*8,5*
8,1	*8,4*	*8,9*
8,6	*9,0*	*9,1*
9,1	*9,4*	*9,7*

A partir de estos datos se obtuvieron los siguientes resultados, los cuales nos permiten construir los correspondientes diagramas de caja.

	Primero	***Segundo***	***Tercero***
Mín.	*4,7*	*5,6*	*4,3*
Máx.	*9,1*	*9,4*	*9,7*
Q_1	*5,7*	*6,1*	*5,5*
Me	*7,1*	*7,6*	*7,8*
Q_3	*8,1*	*8,3*	*8,5*

Estos gráficos muestran como la tercera distribución es asimétrica a la izquierda y tiene una dispersión mayor que las otras dos. Por el contrario la primera es casi simétrica mientras que la segunda es la menos dispersa, pues el cincuenta por ciento de las observaciones centrales se concentra en un recorrido pequeño.



2.13 Momentos de una distribución.

Hasta ahora hemos venido hablando de distintas características de una distribución. Las mismas se han cuantificado mediante una serie de medidas, tales como la media aritmética, la variancia y otras más. Pues bien, algunas de ellas son casos particulares de lo que, genéricamente, se conoce como momentos. Estos son medidas que se obtienen mediante la información numérica que suministra una variable y sirven para caracterizar, de forma única, a una distribución. Tanto es así que si dos distribuciones tienen idénticos momentos, entonces se trata de la misma distribución.

Los momentos se agrupan en dos categorías. Los momentos con respecto al origen y los momentos con respecto a la media o centrales. Conceptualmente son idénticos,

diferenciándose solo en el origen que se tome como referencia. De hecho, como se verá más adelante, los unos están relacionados con los otros.

Los momentos con respecto al origen se definen de la forma siguiente:

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k n_i}{N} \qquad k = 0,1,2,3,4,..... \qquad (2.24)$$

Los momentos con respecto al origen que más suelen utilizarse son los de menor orden, especialmente: m'_1 y m'_2. El momento de orden cero vale siempre uno, no importa cual sea la variable. El de orden uno es la media aritmética y el de orden dos es el que interviene en el cálculo de la variancia.

Los momentos con respecto a la media o centrales se definen como:

$$m_k = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^k n_i}{N} \qquad k = 0,1,2,3,4,.... \qquad (2.25)$$

Al igual que ocurre con los momentos con respecto al origen, el número de momentos con respecto a la media es infinito, pero solo unos pocos, los primeros, son los que más se usan. Los dos primeros no suelen usarse casi nunca, pero tienen un valor concreto. Así el momento con respecto a la media de orden cero (m_0) vale uno, al igual que su correspondiente con respecto al origen. El de orden uno (m_1) vale cero, pues se trata de la suma de las desviaciones con respecto a la media. El de orden dos (m_2) es la variancia de la variable. El de orden tres (m_3) se utiliza para definir el coeficiente de asimetría de Fisher. El de orden cuatro se usa para el análisis de la curtosis.

A continuación vamos a dar la relación que existe entre los momentos con respecto al origen y los momentos con respecto a la media. Para ello hay que hacer uso del desarrollo del binomio de Newton:

$$\left(x_i - \bar{x}\right)^k = \sum_{j=0}^{k} (-1)^j \binom{k}{j} x_i^{k-j} \bar{x}^j$$

Sustituyendo esta relación en la definición de momento de orden k con respecto a la media se tiene:

$$m_k = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^k n_i}{N} = \frac{1}{N} \sum_{i=1}^{n} \left[\sum_{j=0}^{k} (-1)^j \binom{k}{j} x_i^{k-j} \bar{x}^j \right] n_i = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left(\frac{1}{N} \sum_{i=1}^{n} x_i^{k-j} n_i \right) =$$

$$= \sum_{j=0}^{k} (-1)^j \binom{k}{j} m'_{k-j} \bar{x}^j$$

Como caso particular de esta relación se tiene la siguiente:

$$S^2 = m_2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 n_i}{N} = \frac{\sum_{i=1}^{n} x_i^2 n_i}{N} - \left(\bar{x}\right)^2 = m'_2 - \left(m'_1\right)^2$$

donde se muestra que la variancia, (momento de orden dos con respecto a la media), es igual al momento de orden dos con respecto al origen menos la media (momento de orden uno con respecto al origen) al cuadrado.

2.14 Desigualdad

En este apartado vamos a continuar hablando de concentración pero en un sentido diferente al utilizado en los epígrafes anteriores. Ahora, cuando hablemos de concentración, estaremos haciendo referencia al mayor o menor grado de igualdad o desigualdad en el reparto o distribución de una cierta magnitud. Ahora el término concentración será sinónimo de desigualdad. Las distintas medidas de concentración que pueden utilizarse, en el sentido señalado con anterioridad, son indicadores del grado de equidistribución de la variable.

Estos indicadores tienen una gran aplicación en el ámbito de la economía y especialmente dentro de lo que podría llamarse economía social, pues lo que buscan es determinar el mayor o menor grado de equidad en el reparto de variable tales como renta,

salarios, riqueza, etc, entre los preceptores de esas magnitudes. Pero estos indicadores no solo se usan para cuantificar la desigualdad en el reparto entre personas. También pueden utilizarse para estudiar el reparto en áreas geográficas, empresas, etc.

Los indicadores más habituales que se utilizan en este ámbito son la Curva de Lorenz, el Indice de Gini y el Indice de Theil. Todos ellos se basan en el siguiente principio. Sea X una variable (por ejemplo la renta) cuyos valores ordenados son los siguientes: $x_1 \sigma, x_2 \sigma, \ldots\ldots \sigma, x_n$, donde cada x_i es lo percibido por un sujeto. Lo que se pretende saber es si el total $\Lambda_i x_i$ (la renta total en este caso) está equitativamente distribuida. Las dos situaciones extremas que se pueden dar frente a este problema son las siguientes:

1º Que todos los perceptores tengan el mismo nivel de renta, en cuyo caso se hablaría de equidistribución o concentración mínima: $x_1 = x_2 = \ldots = x_n$.

2º Que *n-1* perceptores tengan renta nula y solo uno obtenga toda la renta. En este caso la concentración sería máxima.

Entre estas dos soluciones extremas pueden existir un número infinito de soluciones con distinto grado de concentración.

2.14.1 Curva de Lorenz

Para la exposición de este instrumento de concentración haremos uso de la información contenida en la Tabla 4. Los datos de esta tabla hacen referencia a una distribución genérica (x_i, n_i) cualquiera, referida a rentas, salarios, etc.

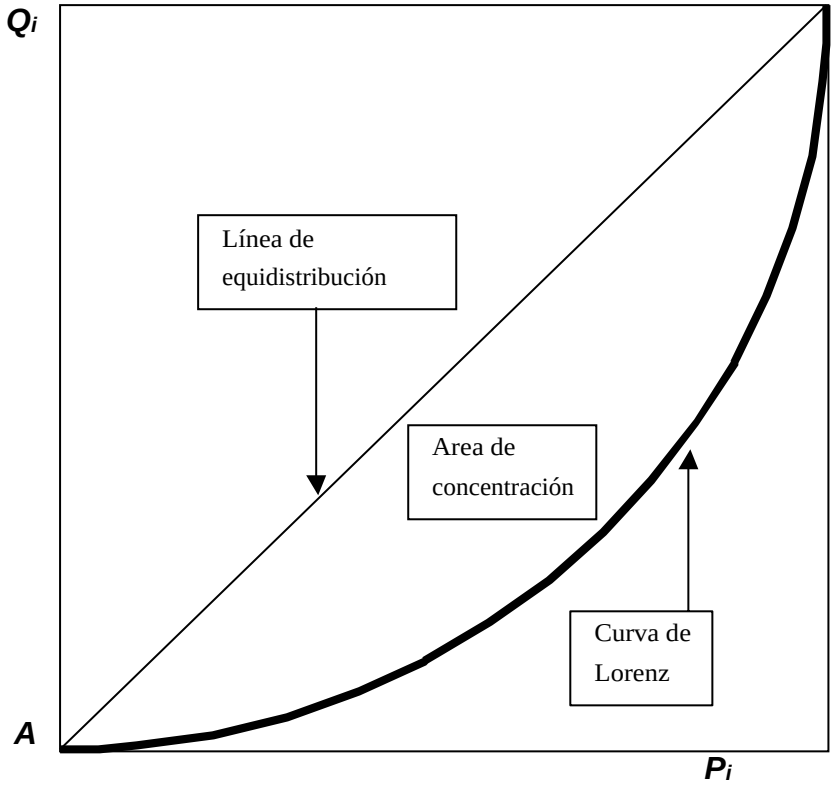
De todas las columnas de esta tabla, las que vamos a utilizar para representar la Curva de Lorenz son las encabezadas por P_i y Q_i. La representación gráfica de estos valores dará lugar a la Figura 14.

Tabla 4. Información para la Curva de Lorenz y el Índice de Gini.

x_i	n_i	$t_i = x_i n_i$	$p_i = (n_i/N)$	$q_i = (t_i/T)$	N_i	T_i	$P_i = (N_i/N)$	$Q_i = (T_i/T)$	$P_i Q_{i+1}$	$P_{i+1} Q_i$
x_1	n_1	$t_1=x_1 n_1$	n_1/N	t_1/T	N_1	$T_1=t_1$	P_1	Q_1	P_1Q_2	P_2Q_1
x_2	n_2	$t_2=x_2 n_2$	n_2/N	t_2/T	N_2	$T_2=T_1+t_2$	P_2	Q_2	P_2Q_3	P_3Q_2
x_3	n_3	$t_3=x_3 n_3$	n_3/N	t_3/T	N_3	$T_3=T_2+t_3$	P_3	Q_3	P_3Q_4	P_4Q_3
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
x_i	n_i	$t_i=x_i n_i$	n_i/N	t_i/T	N_i	$T_i=T_{n-i}+t_i$	P_i	Q_i	P_iQ_{i+1}	$P_{i+1}Q_i$
.	.	.	.	.	.	.	.	.	.	.
.	.	.		.	.	.	.	.	.	.
x_n	n_n	$t_n=x_n n_n$	n_n/N	t_n/T	$N_n=N$	$T_n=T$	$P_n=1$	$Q_n=1$	---	---
	N	$T=\Sigma x_i n_i$								

A la diagonal dibujada se le conoce como línea de equidistribución, mientras que a la línea bajo esa diagonal se le conoce como Curva de Lorenz. En realidad esa curva, cuando se obtiene como representación gráfica de un conjunto de pares de valores, como los de la Tabla 4, es una línea poligonal. Cuanto más se aleje la Curva de Lorenz de esa diagonal, mayor será el grado de concentración o desigualad en el reparto y, en consecuencia, menor la equidistribución. Esa curva siempre estará por debajo de la línea *AB,* nunca la cruzará, pues la Curva de Lorenz responde a una función monótona no decreciente. Solo coincidirá con ella cuando $P_i = Q_i$, es decir, cuando se de la situación de total igualdad en el reparto o equidistribución.

Figura 14. Curva de Lorenz



2.14.2 Índice de Gini

El índice de Gini se define como el cociente entre el área delimitada por la línea *AB* (Figura 14) y la Curva de Lorenz (área de concentración) y el área del triángulo inferior a la línea *AB*. Los valores de este índice van de cero a uno. El valor cero se alcanza cuando la Curva de Lorenz coincide con el segmento *AB*, pues en tal caso el área comprendida entre ambas es nula. Se trataría del caso de equidistribución. En este caso se habla de que no hay desigualdad en el reparto o que la concentración es nula. Todos los preceptores reciben la misma cantidad. El otro valor extremo se alcanza cuando la Curva de Lorenz está formada por los segmentos *AC* y *CB*, pues en este caso el área de concentración coincide con la del triángulo inferior. En este situación se hablaría de concentración extrema y el Indice de Gini valdría uno. Esta circunstancia se daría cuando solo un preceptor recibe toda la renta, ingreso, etc. Entre estas dos situaciones extremas y poco probables se pueden dar un conjunto infinito de valores.

La forma aproximada de calcular ese índice, que lo representaremos por *G*, se basará en la determinación del área bajo la curva. Esta área será la suma de áreas de triángulos y trapecios, pues en casos prácticos la curva de Lorenz se convierte en una línea poligonal más que en una línea continua como la dibujada en la figura anterior. Concretamente se tendría:

G = (Area de concentración)/(Area del triángulo ACB) =
=(Area de concentración)/(1/2)= 2(Area de concentración)=
= 2(1/2 – Area bajo la curva)= 1 – 2(Area bajo la curva)=

$$\sum_{i=1}^{n-1} P_i Q_{i+1} - \sum_{i=1}^{n-1} P_{i+1} Q_i \qquad (2.26)$$

Esta expresión de cálculo del Indice de Gini es solo una aproximación[5] a su verdadero valor, por cuanto no se trabaja en términos continuos, dado que, como ya se ha señalado antes, el área de concentración con la que se trabaja es la comprendida entre la diagonal principal (línea de equidistribución) y la poligonal que representa a la Curva de Lorenz. Pero esta no es la única forma de obtener el Indice de Gini. Otra, equivalente a la expresión anterior, es la siguiente:

$$G \cong 1 - \sum_{i=1} p_i \left[\sum_{j=1}^{i-1} 2q_j + q_i \right] = \frac{\sum_{j<i} (x_i - x_j) p_i q_j}{\bar{x}} \qquad (2.27)$$

Este índice es adimensional e invariante ante cambios de escala, pero no frente a cambios de origen, los cuales hacen que *G* se reduzca.

[5] Se dice que es una aproximación porque el área bajo la Curva de Lorenz habría que obtenerla mediante la correspondiente integral de la función que representa a la Curva. Ahora bien, si se admite como válido sustituir esa línea continua, que es la Curva de Lorenz, por la poligonal

Ejemplo 19. *Para la distribución de salarios mensuales (expresada en euros) dada en la siguiente tabla*

Salarios (x_i)	***Asalariados (n_i)***
Menos de 500	*50*
De 500 a 900	*70*
De 900 a 1200	*120*
De 1200 a 1800	*100*
De 1800 a 2700	*50*
De 2700 a 5000	*20*
Total	***410***

analizar la concentración tanto gráfica como analíticamente.

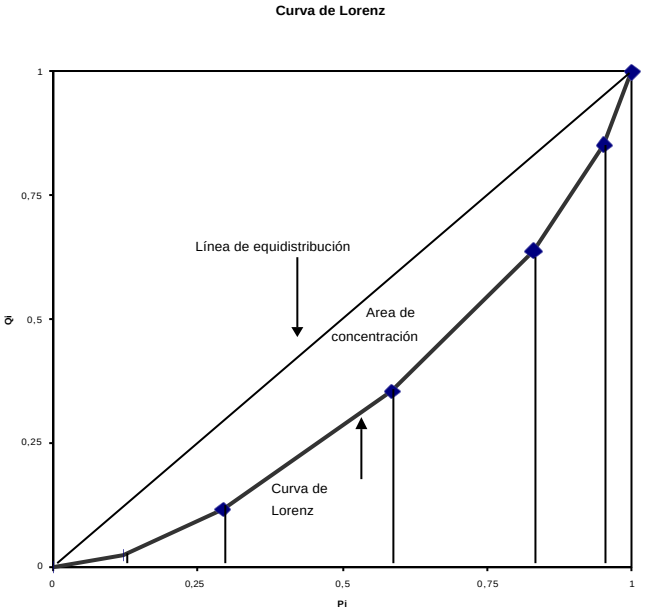
Para estudiar la concentración lo primero que se hará es completar la tabla anterior en la forma siguiente:

Marca de clase (x_i)	***t_i***	***N_i***	***T_i***	***P_i***	***Q_i***	***$P_i Q_{i+1}$***	***$P_{i+1}Q_i$***
250	*12500*	*50*	*12500*	*0,12195*	*0,02371*	*0,01423*	*0,00694*
700	*49000*	*120*	*61500*	*0,29268*	*0,11669*	*0,10413*	*0,06831*
1050	*126000*	*240*	*187500*	*0,58536*	*0,35578*	*0,37487*	*0,29504*
1500	*150000*	*340*	*337500*	*0,82926*	*0,64041*	*0,70810*	*0,60917*
2250	*112500*	*390*	*450000*	*0,95121*	*0,85388*	*0,95121*	*0,85388*
3850	*77000*	*410*	*527000*	*1*	*1*	*0*	*0*

A partir de estos datos se obtiene la Curva de Lorenz que nos permite tener una idea gráfica del grado de concentración que presenta la variable de este ejemplo. Como se aprecia, no puede decirse que concentración sea nula, pues la Curva de Lorenz se aleja de la línea de equidistribución. Además, también puede observarse por los datos de la tabla anterior que el 12% de los que menos ganan solo obtienen el 2% del total de la

correspondiente, entonces el área bajo esa poligonal obtenida mediante la expresión (2.26) es exacta, sin importar el número de valores que tome la variable.

masa salarial, mientras que el 5% de los que más ganan acaparan casi el 15% del total de los salarios. Sin embargo la concentración no es demasiado elevada.



Pero ni el gráfico anterior ni los datos elaborados de la tabla nos permiten tener una idea más precisa del grado de concentración de esta variable. Para ello hay que recurrir al cálculo del Indice de Gini. En este caso, el valor del mismo es:

$$G = \sum_{i=1}^{n-1} P_i Q_{i+1} - \sum_{i=1}^{n-1} P_{i+1} Q_i = 2{,}1526 - 1{,}8334 = 0{,}3192$$

Se trata de un valor más próximo a cero que a uno. En consecuencia, la situación no es de reparto igualitario pero tampoco puede hablarse de que la concentración sea elevada.

2.14.3 Índice de Theil.

Este indicador de desigualdad en el reparto de una magnitud entre distintas unidades preceptoras o de asignación (el reparto puede tener lugar entre personas (rentas), empresas (cuotas de mercado), unidades espaciales (provincias o regiones), etc.) fue introducido inicialmente como una medida de entropía dentro del contexto de la teoría de la información. La entropía sirve para medir el grado de desorden en un sistema (un sistema desordenado sería equivalente a otro en el que cada uno de los componentes del

mismo no están “equilibrados”) y también para comparar situaciones distintas, como veremos más adelante.

El índice de Theil se define, inicialmente, en términos de las probabilidades de los distintos valores de un distribución. Sin embargo, esas probabilidades pueden aproximarse por las frecuencias relativas observadas para esos valores o simplemente por un conjunto de proporciones, con la únicas condiciones de que sean no negativas y que su suma sea igual a la unidad. Pues bien, si nos situamos en este último contexto, podemos imaginar la situación de un conjunto de *N* empresas que se dedican a producir el mismo bien. Cada una de estas empresas tiene su propia cuota de mercado, siendo p_i la de la empresa *i-ésima*. Para este caso, el Índice de Theil viene dado por:

$$H = \sum_{i=1}^{N} p_i \log \frac{1}{p_i} \qquad (2.28)$$

Este índice está acotado y sus valores extremos son *0* y *log N.* El valor cero tiene lugar cuando la cuota de mercado de todas las empresas vale cero, salvo la de una que es la unidad (situación de monopolio). En este caso se dice que la concentración es máxima. No hay igualdad en el reparto. En general esta situación de concentración máxima se da cuando $p_i=1$ y $p_j=0$ para todo *j* distinto de *i*. En cambio si cada empresa tiene la misma cuota de mercado (competencia perfecta), es decir $p_i = p_j$, la concentración es mínima y se habla de reparto igualitario. En esta situación el valor del Índice de Theil es *log N*. Así pues cuanto mayor es el valor del índice menor desigualdad en el reparto y en la medida que se acerque a cero se puede hablar de mayor concentración.

Ahora bien, como puede observarse, el valor de este índice depende de *N* en su extremo superior, por lo que no sirve para realizar comparaciones cuando el tamaño de la población es distinto. En estos casos se podría recurrir a dividir el índice por log *N*. Si se procede de esta forma, los valores extremos serían siempre cero y la unidad.

El problema de la comparación puede resolverse como acaba de indicarse o bien realizando otro planteamiento. Supongamos que se tienen dos distribuciones distintas, *X* e *Y*, pero con un mismo tamaño poblacional. Si representamos por p_i a las participaciones

de los valores de *X* en su distribución y por q_i a las de *Y* en la suya, entonces el Índice de Theil para este caso viene dado por la expresión:

$$I(q:p) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}$$

Este indicador tomará el valor cero cuando $p_i = q_i$, es decir, cuando la igualdad o desigualad en el reparto en las dos distribuciones es idéntica. Además puede demostrarse que cuando esas proporciones no son iguales el indicador es siempre mayor que cero pudiendo hacerse infinito cuando $q_i > p_i = 0$.

Ejemplo 20. *En una determinado país existen 6 empresas que se dedican a la distribución de energía eléctrica siendo sus cuotas de mercado (proporción de energía vendida por la empresa i-ésima respecto del total de energía vendida) las que se recogen en la siguiente tabla:*

Empresa	*Cuota de mercado (p_i)*
A	*0,20*
B	*0,15*
C	*0,10*
D	*0,05*
E	*0,05*
F	*0,45*

Estudiar el grado de concentración empresarial de este sector para ese país.

En este caso, aunque también sería aplicable el índice de Gini, vamos a utilizar el de Theil, para lo cual no es necesario ordenar los valores con ningún criterio. Lo primero que debe hacerse es ampliar la tabla con aquellas columnas adicionales que recojan todos los cálculos necesarios para la obtención del índice.

Empresa	*Cuota de mercado (p_i)*	*$p_i \log(1/p_i)$*
A	*0,20*	*0,3219*
B	*0,15*	*0,2846*
C	*0,10*	*0,2303*
D	*0,05*	*0,1498*
E	*0,05*	*0,1498*
F	*0,45*	*0,3593*

Con estos datos resulta que:

$$H = \sum_{i=1}^{N} p_i \log \frac{1}{p_i} = 1{,}4956$$

que se aleja del valor cero y se aproxima bastante a log (6) = 1,7918.

Ejemplo 21. *En 1999 la distribución de la población y el VAB por provincias en Andalucía era el que se da en la tabla siguiente. Realizar un análisis de la concentración de la población y la renta en Andalucía.*

Provincias	*VAB (millones ptas)*	*Población*
Almería	*890462*	*512843*
Cádiz	*1685973*	*1119802*
Córdoba	*1132683*	*768676*
Granada	*1125698*	*813061*
Huelva	*712562*	*457507*
Jaén	*946873*	*649662*
Málaga	*2051469*	*1258084*
Sevilla	*2752314*	*1725482*

En este caso se podría plantear calcular el Índice de Theil para cada una de las distribuciones y compararlo. Pero con esto no se consigue el objetivo que se busca y que no es otro que determinar si la distribución de la población y la producción siguen el mismo patrón, es decir, si la producción per capita es parecida o si por el contrario la producción no tiene lugar donde está la población. Para conseguir este objetivo se puede utilizar el Índice de Theil dado por:

$$I(q:p) = \sum_{i=1}^{N} q_i \log\frac{q_i}{p_i} = 0{,}0018$$

Como puede apreciarse el valor del índice es muy bajo, lo que significa que la producción per capita es muy parecida entre todas las provincias, es decir que la concentración de la población es similar a la de la producción. Pero la información que contiene la tabla que se ha elaborado para calcular este índice nos advierte que hay cuatro provincias que contribuyen o concentran, proporcionalmente, más producción que población. En este caso se trata de Almería, Huelva, Málaga y Sevilla.

Provincias	VAB (millones ptas)	Población	p_i	q_i	$q_i \log(q_i/p_i)$
Almería	890462	512843	0,07020	0,07881	0,00912
Cádiz	1685973	1119802	0,15329	0,14922	-0,00400
Córdoba	1132683	768676	0,10522	0,10025	-0,00485
Granada	1125698	813061	0,11132	0,09963	-0,01109
Huelva	712562	457507	0,06262	0,06306	0,00044
Jaén	946873	649662	0,08893	0,08380	-0,00497
Málaga	2051469	1258084	0,17221	0,18157	0,00960
Sevilla	2752314	1725482	0,23620	0,24360	0,00752

1-. La distribución del presupuesto semanal en alimentación de un conjunto de 265 familias, expresado en euros, es el que figura en la tabla siguiente:

Presupuestos $L_{i-1} - L_i$	**Familias** n_i
80-100	10
100-110	35
110-115	40
115-120	45
120-130	55
130-150	30
150-170	20
170-210	15
210-270	10
270-360	5
Total	**265**

A partir de esa información:

1º Diga cual es la población, los elementos, la característica observada, el tipo de variable.
2º Represente gráficamente esta variable.
3º Obtenga la media, mediana, moda, variancia, desviación estándar, coeficiente de variación, coeficiente de asimetría, coeficiente de curtosis.
4º Determine el recorrido intercuartílico, interdecílico e intercentílico.
5º Determine el número de familias con un presupuesto inferior a la media.
6º Determine el porcentaje de familias con un presupuesto comprendido entre la media menos dos veces la desviación estándar y la media más dos veces la desviación estándar.
7º Obtenga la curva de concentración de Lorenz.
8ª Determine el Indice de Gini.
9ª Responda a todos los apartados anteriores si la unidad monetaria fuera la peseta.

2.- La distribución de los salarios semanales en una determinada industria, expresada en euros, es la que figura en la tabla siguiente:

Salarios $L_{i-1} - L_i$	**Asalariados** n_i
Menos de 90	2145
90-120	1520
120-150	840
150-180	955
180-210	1110
210-240	2342
240-300	610
300-600	328
Más de 600	150
Total	**10000**

A partir de esa información y sabiendo que la nomina total de esa industria asciende a 1765950 €

1º Diga cual es la población, los elementos, la característica observada, el tipo de variable.
2º Represente gráficamente esta variable.
3º Obtenga la media, mediana, moda, variancia, desviación estándar, coeficiente de variación, coeficiente de asimetría, coeficiente de curtosis.
4º Determine el recorrido intercuartílico, interdecílico e intercentílico.
5º Determine el salario por debajo del cual están el 35% de los asalariados.
6º Determine de forma aproximada que porcentaje de asalariados cobran más de una vez y media la el salario medio.
7º Determine el porcentaje de asalariados que cobran salarios comprendidos entre la media menos dos veces la desviación estándar y la media más dos veces la desviación estándar.
8º Obtenga la curva de concentración de Lorenz.
9ª Determine el Indice de Gini.
10ª Responda a todos los apartados anteriores si la unidad monetaria fuera la peseta.

3.- Se les ha preguntado a los 1000 establecimientos de comercio minorista de una determinada ciudad por sus ventas anuales de un cierto producto de alimentación. Los resultados obtenidos son los que refleja la tabla siguiente:

Ventas (Euros)	Establecimientos
Hasta 600	400
De 600 a 1500	225
De 1500 a 3000	175
De 3000 a 6000	120
De 6000 a 9000	75
Más de 9000	5

Además se sabe que el total de ventas para ese producto ascendió a 1953000 euros. Con toda esta información:

a) Según la forma de la distribución, indique y calcule el promedio más adecuado para representar a esta variable.
b) 0btenga la media aritmética de esta variable. Analice su representatividad en términos absolutos y relativos.
c) Determine el porcentaje de establecimientos, de entre los que menos venden, cuyas ventas acumuladas representan la cuarta parte de las ventas totales. Determine también el valor de la variable que deja a su derecha el 10% de los establecimientos que más venden.
d) Si para otra ciudad y esa misma variable se hubieran obtenido unas ventas medias de 3000 euros con una desviación estándar de 6000 € y un índice de Gini de 0,7, realice una análisis comparado de la dispersión y de la concentración de esa variable.
e) Si al siguiente año el precio de ese producto aumenta un 5%, determine cual sería el nuevo volumen medio de ventas así como su variancia.

CAPITULO 3.- ANÁLISIS CONJUNTO DE DOS VARIABLES

3.1 Presentación de los datos. Tablas de doble entrada.

En el capítulo anterior nos hemos interesado por el análisis y descripción de una sola variable. Para ello hemos definido un proceso de reducción de la información inicialmente disponible. Esta reducción ha dado como resultado la construcción de una tabla estadística donde se daba la distribución de frecuencias de la variable. Posteriormente se ha analizado la forma, se han definido medidas de tendencia central, medidas de dispersión, de simetría y curtosis. También se ha estudiado el problema de la concentración. Pero este análisis es de tipo unidimensional, pues de todos los caracteres de los elementos de una población solo nos ha preocupado observar un de ellos que, por lo regular, siempre ha sido de tipo cuantitativo. Pero que duda cabe que los elementos de una población cualquiera gozan de más de un carácter susceptible de ser observado. En este sentido, imaginemos que los elementos observados son las empresas. En ellas se pude observar de forma conjunta los beneficios y los costes de las mismas o cualquier otro par de caracteres. Así podríamos pensar en los gastos en publicidad y sus beneficios, o los costes y el número de empleados. El número de ejemplos que podríamos dar es tan amplio que no merece la pena seguir mencionándolos.

El objetivo de este capitulo será similar al del anterior, pero ahora buscando el análisis conjunto de dos variables o análisis bidimensional. Para ello se procederá a la observación de dos características de todos los elementos de una población. Inicialmente supondremos que esas características son de naturaleza cuantitativa. El resultado de esa observación conjunta será la definición de dos variables a las que llamaremos X e Y, las cuales pueden ser discretas o continuas, y nuestra primera preocupación será la de presentar de forma conjunta las frecuencias de los pares de valores de esas variables (x_i, y_j). El instrumento que se utiliza para alcanzar ese objetivo es lo que se conoce como tabla de doble entrada, tabla de correlaciones o tabla de contingencia. Esta última denominación se reserva especialmente para los casos de caracteres cualitativos. De todas las denominaciones que hemos señalado, usaremos la de tabla de doble entrada, pues la denominación de tabla de correlaciones tiene un significado que va más allá de la mera representación numérica de las la distribución conjunta de frecuencias.

Una tabla de doble entrada no es más que la representación de (x_i, y_j, n_{ij}) en la forma que se muestra en la Tabla1.

Tabla 1. Distribución conjunta de dos variables

		Y						
		y_1	y_2	.....	y_j	.....	y_k	$n_{i.}$
X	x_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1k}	$n_{1.}$
	x_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2k}	$n_{2.}$
	. . .	. . .	. . .	. ...	. . .	. ...	. . .	. . .
	x_i	n_{i1}	n_{i2}		n_{ij}		n_{ik}	$n_{i.}$
	. . .	. . .	. . .	. ...		. ...	. . .	. . .
	x_h	n_{h1}	n_{h2}	.....	n_{hi}	.....	n_{hk}	$n_{h.}$
	$n_{.j}$	$n_{.1}$	$n_{.2}$	.....	$n_{.j}$	.....	$n_{.k}$	N

La lectura del contenido de esta tabla sería el siguiente. El valor n_{ij} nos da la frecuencia conjunta con la que se presentan el valor x_i de X y el valor y_j de Y. A su vez n_{i1} da la frecuencia conjunta de x_i y de y_1. De forma similar habría que leer e interpretar el resto de las frecuencias conjuntas que son las que están dentro del cuerpo central de la tabla, es decir, las que llevan un doble subíndice alfanumérico.

Mención aparte merecen la última fila y la última columna. A esa fila y a esa columna se les conoce como distribuciones marginales de Y y de X, respectivamente. Se trata de la distribución de frecuencias de cada una de las variables tomadas por separado. Así pues la distribución marginal de X vendría dada por los pares (x_i, $n_{i.}$), mientras que la marginal de Y vendría dada por los pares (y_j, $n_{.j}$), es decir:

Tabla 2. Distribuciones marginales de *X* y de *Y*

X	$n_{i.}$
x_1	$n_{1.}$
x_2	$n_{2.}$
.	.
.	.
x_i	$n_{i.}$
.	.
.	.
x_h	$n_{h.}$

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

donde

$$n_{i.} = n_{i1} + n_{i2} + ... + n_{ij} + ...+ n_{ik} = \Lambda_j n_{ij} \qquad j = 1, 2, ..., k \qquad (3.1)$$

$$n_{.j} = n_{1j} + n_{2j} + ... + n_{ij} + ...+ n_{hj} = \Lambda_i n_{ij} \qquad i = 1, 2, ..., h \qquad (3.2)$$

Finalmente se cumple que

$$\Lambda_i n_{i.} = \Lambda_j n_{.j} = \Lambda_i \Lambda_j n_{ij} = N \qquad (3.3)$$

Además de las marginales, para una tabla de doble entrada, se pueden definir también las distribuciones condicionadas , que también son de tipo unidimensional. Estas hay que darlas en términos de una condición previa. En este sentido se tendría la distribución de los valores de la variable *X* condicionada a que la variable *Y* tome un valor concreto. De igual forma se podría hacer para la variable *Y* con respecto a los posibles valores de *X*. Si se define la condicionada de *X*, entonces los valores que puede tomar esta variable son los mismos que los de su marginal. Lo único que varía son sus frecuencias, que se representarán por $n_{i/j}$. A su vez, si de lo que se trata es de la condicionada de *Y*, los

valores de esta distribución son los de la marginal de *Y*, pero las frecuencias son distintas y se representa por $n_{j/i}$. Estas nuevas distribuciones aparecen en la Tabla 3.

Tabla 3. Distribuciones condicionales de *X* y de *Y*

X/y_j	$n_{i/j}$
x_1	n_{1j}
x_2	n_{2j}
.	.
.	.
x_i	n_{ij}
.	.
.	.
x_h	n_{hj}

Y/x_i	$n_{j/i}$
y_1	n_{i1}
y_2	n_{i2}
.	.
.	.
y_j	n_{ij}
.	.
.	.
y_k	n_{ik}

La distribución condicional no es única, al el contrario de lo que ocurre con la marginal. Habrá tantas como valores pueda tomar la variable condicionante. Así, para variables continuas el número de distribuciones condicionales será infinito.

Todas y cada una de esta nuevas distribuciones univariantes que se han definido es posible tratarlas con los instrumentos de análisis definidos en las lecciones anteriores. Además, aunque la tabla de doble entrada que se ha diseñado antes lo es para variables de tipo cuantitativo, también es posible hablar de tablas de doble entrada para variables de tipo cualitativo o mixto, en cuyo caso se les conoce como tablas de contingencia. Por otro lado, en la Tabla 1 se recogen dos variables discretas con frecuencias unitarias o mayores que la unidad. Sin embargo ese diseño de tabla de doble entrada es también válido para el caso de variables continuas. Bastaría con sustituir los valores puntuales de cada variable por intervalos.

A continuación vamos a dar un ejemplo que permita aclarar todos estos conceptos.

Ejemplo 1. *Para un conjunto de 2005 empresas de menos de 9 empleados se han observado dos caracteres de las mismas. El número de sus empleados (X) y el número de días perdidos por bajas (Y) en esas empresas. Los resultados son los que se dan en la siguiente tabla de doble entrada:*

		Y									
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$n_{i.}$
X	**1**	50	45	40	30	20	15	10	5	5	220
	2	40	50	45	40	30	20	15	10	5	255
	3	20	40	50	40	35	25	20	15	10	255
	4	15	30	30	50	40	30	25	20	15	255
	5	10	20	20	40	50	40	40	35	30	285
	6	5	10	15	30	40	50	45	40	35	270
	7	5	5	10	20	30	40	50	45	40	245
	8	5	5	5	10	20	30	45	50	50	220
	$n_{.j}$	150	205	215	260	255	250	250	220	200	2005

A partir de esos datos, obtenga

a) La marginal de X y la condicional de X/y=5.

b) La marginal de Y y la condicional de Y/x=3

a)

Marginal de X		*Condicional de X/y=5*	
x_i	$n_{i.}$	$x_i/y=5$	$n_{i/y=5}$
1	220	1	15
2	255	2	20
3	255	3	25
4	255	4	30
5	285	5	40
6	270	6	50
7	245	7	40
8	220	8	30

b)

Marginal de Y		*Condicional de Y/x=3*	
y	$n_{.j}$	$y_j/x=3$	$n_{j/x=3}$
0	150	0	20
1	205	1	40
2	215	2	50
3	260	3	40
4	255	4	35
5	250	5	25
6	250	6	20
7	220	7	15
8	200	8	10

Ejemplo 2. *En la tabla siguiente se recoge la distribución de los asalariados fijos en explotaciones agrarias según edad y tamaño de las mismas en Andalucía para el año 1997.*

	Menos de 1 Ha.	De 1 a 2 Ha.	De 2 a 5 Ha.	De 5 a 10 Ha	De 10 a 20 Ha.	De 20 a 30 Ha	De 30 a 50 Ha.	De 50 a 100 Ha.	Más de 100 Ha.	Total
< 35 años	336	195	1203	1145	671	577	234	518	2400	7279
35 a 44 años	409	468	788	452	592	448	349	852	4256	8613
45 a 54 años	231	144	657	581	751	341	418	801	5152	9076
55 a 64 años	62	601	559	212	1008	231	225	835	3260	6992
> 65 años	2	0	33	208	71	160	35	231	569	1309
Total	1041	1407	3239	2598	3094	1758	1261	3237	15637	33270

Fuente: Web INE.

A partir de esos datos, obtenga: a) la distribución marginal de la variable "edad de los asalariados fijos"; b) la distribución condicional de la variable "edad de los asalariados fijos" para explotaciones de más de 100 Ha.

a) *La distribución marginal que se nos pide viene dada por la primera y última columna de la tabla anterior. En concreto, sería la que aparece en la tabla siguiente:*

Edad	Asalariados
< 35 años	7279
35 a 44 años	8613
45 a 54 años	9076
55 a 64 años	6992
> 65 años	1309
Total	33270

b) *La distribución condicional de la variable "edad de los asalariados fijos" para explotaciones de más de 100 Ha. es la que se recoge en la tabla siguiente:*

Edad	Asalariados en explotaciones de más de 100 Ha.
< 35 años	2400
35 a 44 años	4256
45 a 54 años	5152
55 a 64 años	3260
> 65 años	569
Total	15637

3.2 La covariación.

En el apartado anterior hemos presentado una distribución frecuencias conjunta para dos variables. En ese apartado se ha señalado que tipo de distribuciones unidimensionales o univariantes se pueden definir a partir de la bivariante, y se ha indicado que las mismas podían ser tratadas con los instrumentos definidos en lecciones anteriores. Sin embargo, el interés de este capitulo no es precisamente el de realizar un análisis individualizado de todas y cada una de las distintas distribuciones univariantes que se pueda definir a partir de una distribución bivariante. Ahora, nuestro objetivo es el análisis conjunto de las dos variables que se definen en tabla de doble entrada.

Ya no se trata de estudiar solo los promedios y las medidas de dispersión de cada una de esas variables. El siguiente paso que se pretende dar con este capítulo es el análisis de la relación o dependencia que pueda existir entre dos variables. A esa relación la vamos a denominar covariación o variación conjunta**.**

La covariación es un fenómeno bastante habitual entre variables de carácter económico y de otra naturaleza. La covariación que puede darse entre dos variables *X* e *Y* cualesquiera puede ser de distinto tipo. Así puede hablarse de:

1º *Dependencia causal unilateral*. Este tipo de covariación se da cuando una variable influye en la otra y no al contrario. Es decir las variaciones de una variable pueden explicarse por las variaciones de otra, pero no a la inversa.

En este tipo de análisis, a la variable que ejerce influencia en la otra se le llama variable independiente, explicativa, variable causa o exógena. A la otra variable se le llama dependiente, explicada , variable efecto o endógena. Generalmente a la independiente se le suele representar por la letra *X* , mientras que a la dependiente se le representa por la letra *Y*.

A título de ejemplo se puede señalar los siguientes pares de variables: los impuestos y la renta, los benéficos empresariales y el volumen de ventas, los salarios y la cualificación profesional, etc.

2º *Interdependencia*. Esta situación se da cuando la influencia es recíproca entre las dos variables. En este caso se habla de una relación causal bilateral o interdependencia.

Un ejemplo muy claro en Economía de este tipo de relación se encuentra entre precio y producción de un bien. Es bien conocido que, en un sistema de mercado en régimen de competencia perfecta, estas dos variables están interrelacionadas.

3º *Dependencia indirecta*. Este tipo de covariación se da cuando existe una tercera variable que influye simultáneamente sobre *X* e *Y*. En estos casos no existe una relación de causalidad entre esas variables. Sin embargo, la presencia de una tercera que influye en ambas hace que ellas se muevan de forma sincronizada. Pensemos en la superficie quemada por incendios forestales y el número de viajeros en zonas turísticas. Estas dos variables se comportan a lo largo del año de una forma parecida. Pero no puede hablarse de una relación causa efecto entre ellas. En realidad es la variable temperatura climatológica la que condiciona su evolución paralela.

4º *Concordancia* . A veces se sabe que las variables *X* e *Y* son por naturaleza independientes. Sin embargo puede que muestren un movimiento sincronizado, lo que nos llevaría a pensar en un cierta dependencia. Tal podría ser el caso el resultado de las opiniones de un panel de expertos relativas a expectativas de crecimiento de la economía de un conjunto de países.

5º *Covariación casual o espúrea*. Ocurre cuando dos variable se mueven de forma sincronizada pero sin que exista una relación de causalidad entre ellas.

Es conveniente señalar que el tipo de relación que pueda existir entre dos variables no se puede determinar fácilmente mediante instrumentos estadísticos, por lo que ese tipo de covariación habrá que buscarla en el conocimiento previo que se tenga de esas variables. Lo que si puede hacer la Estadística, en cualquier caso, es cuantificar y formalizar matemáticamente la relación o covariación previamente señalada, con el fin de confirmar tal relación y utilizarla luego para *describir* el fenómeno, para *explicarlo* y para realizar *predicciones*.

La forma más sencilla, desde un punto de vista estadístico, de iniciar el estudio de la covariación entre dos variables es mediante un análisis gráfico. Ahora, como tenemos dos variables, recurriremos al eje de abscisas para representar los valores de la variable *X* y al de ordenadas para situar los valores de *Y*. Si en este diagrama bidimensional llevamos las parejas de valores de *X* e *Y*, el resul tado es lo que se llama un diagrama de dispersión o nube de puntos. En este tipo de diagramas se representan parejas de valores con frecuencia unitarias. Si las frecuencias fueran mayores que uno, entonces habría que recurrir a un tercer eje donde llevaríamos las frecuencias de cada una de esas parejas de valores.

Admitiendo que trabajamos con parejas de valores con frecuencias unitarias, los diagramas de dispersión mas habituales serían los siguientes:

Figura 1. Covariación directa

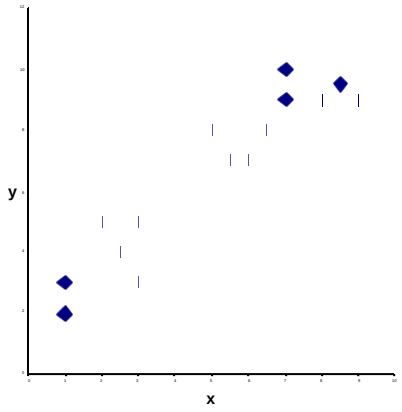


Figura 2. Covariación inversa

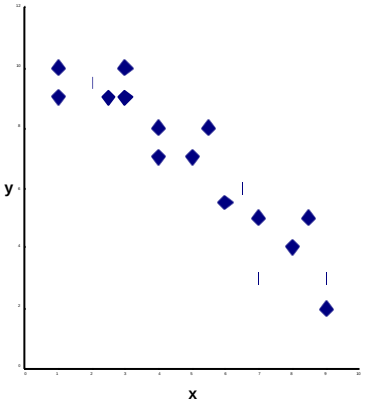


Figura 3. Ausencia de covariación

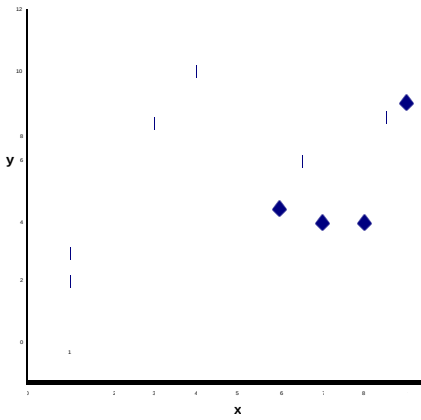
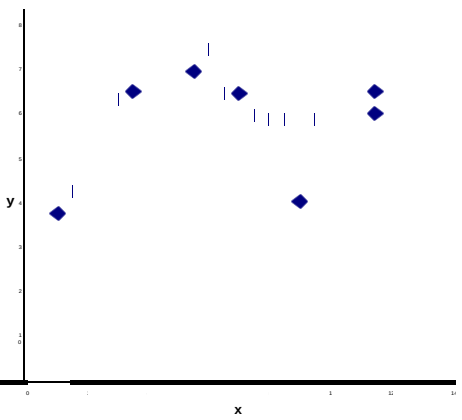


Figura 4. Covariación curvilinea



Mediante este método gráfico lo que se consigue es descubrir la posible relación que existe entre las variables. Esto representa un paso importante para un instrumento tan sencillo como es un simple gráfico.

En la Figura 1, denominada como covariación directa, se detecta una relación lineal positiva o directa. La Figura 2 nos advierte de una relación lineal negativa o inversa; la tercera nos indica que entre las variables *X* e *Y* no existe relación evidente de tipo alguno; finalmente, la última gráfica nos pone de manifiesto una relación que no es lineal.

Que duda cabe que estos cuatro modelos de diagramas de dispersión no son los únicos, pero si los más representativos.

Una vez agotada la vía gráfica para el estudio de la covariación, hay que recurrir a otros procedimientos que nos permitan cuantificar la covariación. Los dos procedimientos más utilizados son la correlación y la regresión.

Antes de finalizar este epígrafe sería conveniente resaltar que para los distintos tipos de covariación que hemos definido hay un concepto que aparece de forma recurrente. Se trata de la independencia o dependencia entre variables. Para definir este concepto en términos estadísticos haremos uso a la tabla de doble entrada que se vio en el apartado anterior. Con la terminología utilizada en esa tabla, se dice que dos variables *X* e *Y* son estadísticamente independientes si se cumple la siguiente relación:

$$n_{ij}/N = (n_{i.}/N).(n_{.j}/N) \qquad (3.4)$$

es decir, que la frecuencia relativa conjunta sea igual al producto de las frecuencias relativas marginales.

Otra forma de dar el concepto de independencia estadística es haciendo uso de las distribuciones condicionales. En este caso se dice que dos variables son estadísticamente independientes si las frecuencias relativas condicionales son iguales a sus correspondientes frecuencias relativas marginales.

$$f_{i/j} = n_{ij}/n_{.j} = (n_{i.}/N) = f_{i.} \quad (3.5)$$

$$f_{j/i} = n_{ij}/n_{i.} = (n_{.j}/N) = f_{.j} \quad (3.6)$$

Ejemplo 3. *Estudie si las variables del ejemplo 1 son o no independientes.*

En este caso, como en otros de naturaleza similar, para determinar si esas dos variables son o no independientes se procederá a aplicar alguna de las condiciones de independencia dadas con anterioridad. Para ello nos centraremos en un punto del espacio de X e Y, por ejemplo el par de valores (x=3, y=6). En este caso se tiene que

$$20/2005 = f_{36} \text{ –::} (f_{3.})(f_{.6}) = (255/2005)(250/2005)$$

$$(255/2005) = (f_{3.}) \text{ –::} (f_{3/y}) = 25/250$$

Lo anterior nos lleva a concluir que esas variables no son independientes. La selección del par (x, y) es indiferente, pues basta que para un par no se cumpla la condición de independencia para que se pueda concluir que las variables no son independientes.

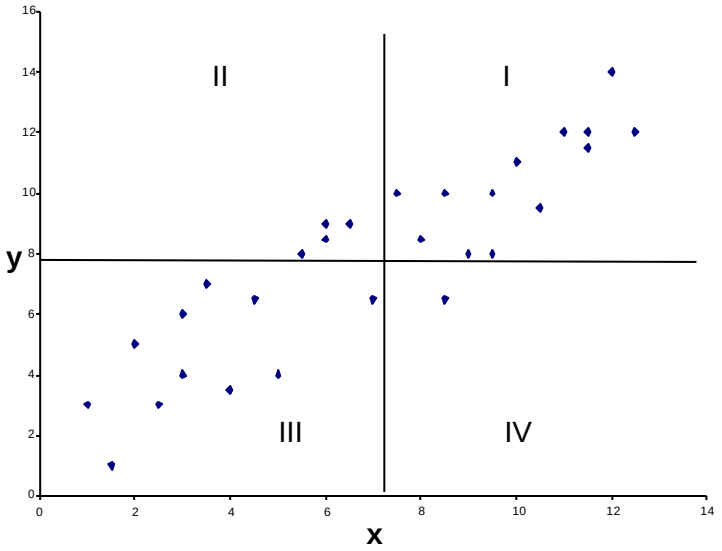
A la independencia estadística definida de esta forma se le llama determinista, frente a la estocástica.

3.3 Correlación: covariancia y coeficiente de correlación lineal.

De los distintos diagramas de dispersión que hemos mostrado en el epígrafe anterior, dos de ellos implicaban una covariación de tipo lineal, en un caso directa y en el otro indirecta o inversa. También se dijo anteriormente que una forma de cuantificar la covariación entre dos variables es mediante el análisis de la correlación. Pues bien, en lo que sigue vamos a definir un instrumento que nos va a permitir cuantificar el grado de covariación lineal entre dos variables. Se trata del coeficiente de correlación lineal.

Para deducir su expresión de cálculo y su significado haremos uso del diagrama de dispersión representado en la Figura 5.

Figua 5. Diagrama de dispersión



En este diagrama se ha realizado un cambio de origen en los ejes de forma que el nuevo origen se sitúa en el punto correspondiente a las medias para las variables *X* e *Y*. Ahora el diagrama de dispersión o nube de puntos está repartido en cuatro cuadrantes. El primer cuadrante se corresponde con los valores de *X* e *Y* mayores que sus medias. El segundo se corresponde con los valores de *X* menores que su media y con los de *Y* mayores que la suya. En el tercero están situados los valores de *X* e *Y* menores que sus medias respectivas. Finalmente, en el cuarto aparecen los valores de *X* mayores que su media y los de *Y* menores que la suya.

De acuerdo con esta distribución de puntos del diagrama, si ahora definimos las desviaciones de *X* e *Y* como $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$, resulta que

a) el producto $x_i y_i$ del primer cuadrante será positivo
b) el producto $x_i y_i$ del segundo cuadrante será negativo

c) el producto $x_i y_i$ del tercer cuadrante será positivo
d) el producto $x_i y_i$ del cuarto cuadrante será negativo.

Teniendo en cuata esos resultados, resulta que $\Sigma_i x_i y_i$ sirve como medida de covariación entre X e Y. Esto es así porque si esa suma es positiva, la mayor parte de los puntos estarán en los cuadrantes I y III, con lo que la relación será directa. Por el contrario, si la mayoría de los puntos están en los cuadrantes II y IV, la suma será negativa y la relación será inversa. En cambio si los puntos están muy repartidos entre los cuatro cuadrantes, la suma será pequeña, tendente a cero, lo que nos informará de que no hay relación lineal alguna.

Pero ese indicador del grado de asociación lineal entre dos variables adolece de dos defectos. Por un lado bastaría con cambiar el número de pares de valores de X e Y para que el mismo fuera distinto. Por otro, el mismo viene influido por las unidades de medida de *X* e *Y*. La forma de corregir estos inconvenientes es promediar la suma (se elimina el primer problema) y expresarla en términos de la desviación estándar de *X* y de *Y*. El resultado es

$$r = \frac{\dfrac{\sum_i x_i y_i}{N}}{S_X S_Y} = \frac{S_{XY}}{S_X S_Y} \qquad (3.7)$$

que se conoce como coeficiente de correlación lineal.

Al numerador del coeficiente de correlación se le llama covariancia (S_{XY}), siendo S_X la desviación estándar de *X* y S_Y la de *Y*. Como las expresiones de cálculo de las desviaciones estándares las conocemos, habrá que dar ahora la correspondiente a la covariancia.

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) n_i}{N} = \frac{\sum_i X_i Y_i n_i}{N} - \frac{\sum_i X_i n_i}{N} \frac{\sum_i Y_i n_i}{N} \qquad (3.8)$$

Mediante el coeficiente de correlación lineal lo que se busca es un número que indique, de forma objetiva, el grado de variación lineal conjunta entre las dos variables. El signo de este coeficiente puede ser positivo o negativo, según cual sea el de la covariancia. Los valores de este coeficiente oscilan entre menos uno y más uno. La forma de interpretar el significado de esos valores es la siguiente:

a) Si $r = 1$, la correlación lineal es perfecta y directa, o sea, la nube de puntos se sitúa sobre una línea recta creciente.

b) Si $r = -1$, la correlación lineal es perfecta y inversa, o sea, la nube de puntos se sitúa sobre una linera recta decreciente.

c) Si $r = 0$, no existe relación lineal, bien porque no exista covariación entre las variables o porque ésta no sea lineal. En este caso decimos que las variables están incorrelacionadas linealmente, lo que no significa que necesariamente sean independientes. Si el coeficiente de correlación lineal es cero, entonces las variables puede que sean independientes o bien que no lo sean y que presenten otro tipo de covariación distinto al lineal. En cambio si las variables son independientes, entonces el coeficiente de correlación lineal será siempre cero.

d) En los demás casos se puede hablar de una correlación débil o fuerte según que el valor de r esté próximo a 0 o a ± 1.

En cuanto a las propiedades del coeficiente de correlación lineal, hay que indicar que el mismo es invariante frente a cambios de origen y de escala. Para probar que esta afirmación es cierta se estudiará el comportamiento de la covariancia frente a cambios de origen y de escala en las variables X e Y, pues ya se sabe cual es la respuesta de la desviación estándar frente a este tipo de cambios. Supóngase que se definen las siguientes variables: $X' = h + kX$ e $Y' = f + gY$. Entonces:

$$S'_{xy} = \frac{\sum_i \left(x'_i - \overline{x'}\right)\left(y'_i - \overline{y'}\right)}{N} = \frac{\sum_i \left(h + kx_i - h - k\overline{x}\right)\left(f + gy_i - f - g\overline{y}\right)}{N} =$$

$$= \frac{kg\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{N} = kgS_{xy}$$

Así pues se comprueba que, al igual que la variancia, la covariancia solo se ve afectada por cambios de escala, por lo que el coeficiente de correlación resulta invariante a los cambios de origen y de escala.

Para finalizar esta exposición sobre el coeficiente de correlación lineal, hay que señalar que para el cálculo del mismo no se asume ningún tipo de relación de causalidad. Por otro lado debe quedar claro que este coeficiente lo que mide es la intensidad de la relación lineal entre variables. Así un coeficiente *r = 0,8* indica una covariación más fuerte que *r= 0,4*, pero ello no implica que la covariación lineal en el primer caso sea doble que en el segundo.

Ejemplo 4. *Obtener el coeficiente de correlación lineal entre las variables X e Y si los valores observado de las mismas son los siguientes:*

X_i	Y_i
3	3
10	9
9	10
1	4
2	1
4	2
6	5
5	6
7	7
7	9

Para calcular el coeficiente de correlación pedido es aconsejable ampliar la tabla anterior con tres columnas adicionales como las que aparecen en la siguiente tabla:

x_i	y_i	x_i^2	y_i^2	x_iy_i
3	3	9	9	9
10	9	100	81	90
9	10	81	100	90
1	4	1	16	4
2	1	4	1	2
4	2	16	4	8
6	5	36	25	30
5	6	25	36	30
7	7	49	49	49
7	9	49	81	63
54	56	370	402	375

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i}{N}\frac{\sum_i y_i n_i}{N} = \frac{375}{10} - \frac{54}{10}\frac{56}{10} = 7{,}26$$

$$S_x = \sqrt{\frac{\sum_i x_i^2 n}{N} - \left(\bar{x}\right)^2} = \sqrt{\frac{370}{10} - \left(\frac{54}{10}\right)^2} = 2{,}8$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n}{N} - \left(\bar{y}\right)^2} = \sqrt{\frac{402}{10} - \left(\frac{56}{10}\right)^2} = 2{,}97$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{7{,}26}{(2{,}8)(2{,}97)} = 0{,}873$$

Ejemplo 5. *Obtenga el coeficiente de correlación lineal para las variables que se recogen en la tabla siguiente.*

		Y			
		1	**2**	**3**	**4**
X	**1**	10	8	5	3
	2	7	12	6	3
	3	6	8	8	4
	4	1	4	5	10

En este caso, se trata de obtener el coeficiente de correlación cuando las frecuencias de los distintos pares de valores de las variables no son unitarias y, además, todos esos pares tienen frecuencias distintas de cero, cosa que no ocurría en el Ejemplo 4. Para calcular la correlación existente entre X e Y, es aconsejable, cuando se tiene una distribución de frecuencias como la presente, determinar previamente las marginales y después dar esa tabla de doble entrada en forma de pares de valores. Todo ello nos lleva a que:

x_i	n_i	$x_i n_i$	$x_i^2 n_i$
1	*26*	*26*	*26*
2	*28*	*56*	*112*
3	*26*	*78*	*234*
4	*20*	*80*	*320*
Total	*100*	*240*	*692*

y_i	n_i	$y_i n_i$	$y_i^2 n_i$
1	*24*	*24*	*24*
2	*32*	*64*	*128*
3	*24*	*72*	*216*
4	*20*	*80*	*320*
Total	*100*	*240*	*688*

$$S_x = \sqrt{\frac{\sum_i x_i^2 n}{N} - \left(\overline{x}\right)^2} = \sqrt{\frac{692}{100} - \left(\frac{240}{100}\right)^2} = 1{,}077$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n}{N} - \left(\overline{y}\right)^2} = \sqrt{\frac{688}{100} - \left(\frac{240}{100}\right)^2} = 1{,}058$$

x_i	y_i	n_{ij}	$x_i y_i n_i$
1	1	10	10
1	2	8	16
1	3	5	15
1	4	3	12
2	1	7	14
2	2	12	48
2	3	6	36
2	4	3	24
3	1	6	18
3	2	8	48
3	3	8	72
3	4	4	48
4	1	1	4
4	2	4	32
4	3	5	60
4	4	10	160
Total		100	617

$$S_{XY} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i \sum_i y_i n_i}{N \quad N} = \frac{617}{100} - \frac{240}{100}\frac{240}{100} = 0{,}41$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{0{,}41}{(1{,}077)(1{,}058)} = 0{,}36$$

3.4 Regresión

El segundo procedimiento que se indicó que podía utilizarse para cuantificar la covariación entre dos variables era el análisis de regresión. La aplicación del mismo se limitará, inicialmente, al caso de dependencia causal unilateral entre dos variables.

El objetivo que se busca con el análisis de la regresión es determinar una función del tipo *y=f(x)* que relacione a estas dos variables y nos indique la forma en varían conjuntamente. Pero esta función, que se intenta cuantificar mediante el análisis de la regresión, será una

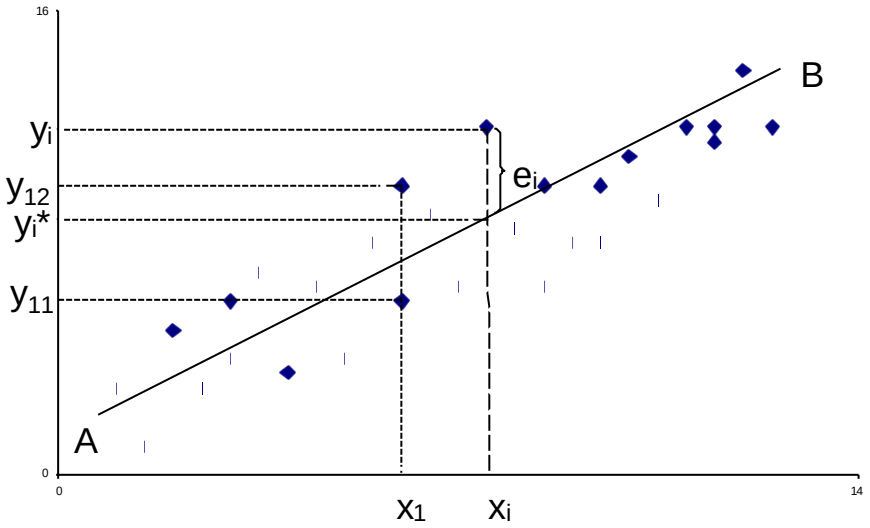
línea que intentará resumir toda la nube de puntos del diagrama de dispersión. Como tal tendrá un carácter de línea media, y esta línea nos medirá la dependencia estadística existente entre las variables. Este tipo de dependencia es distinta a la dependencia funcional o exacta. La diferencia entre las mismas radica en que en el primer caso, aunque las variables estén fuertemente relacionadas, las observaciones suelen tener una componente aleatoria que les impide que la nube de puntos aparezca exactamente distribuida a lo largo de una línea. Pero esa falta de alineación perfecta no impide que esos puntos tiendan a agruparse con mayor o menor intensidad en torno a esa línea "ideal" o media de la que se ha hablado.

Pues bien, el análisis de regresión consiste en obtener esa línea "ideal" o media, línea de regresión, hacia la cual tienden los puntos de un diagrama de dispersión. De lo que se trata, en realidad, es de determinar la dependencia exacta que se haya contenida en la dependencia estadística observada mediante la eliminación de los factores aleatorios.

Para centrar un poco estas ideas se hará uso de la Figura 6. Admitamos de entrada que esa línea media es conocida y que es la que se ha representado en el mismo como AB[1]. En ese gráfico podemos comprobar como para un determinado valor de X (x_1) observado, la variable Y puede tomar, y de hecho los toma en este caso, más de una valor (y_{11} e y_{12}), mientras que por la línea de regresión le correspondería solo uno (y^*_1). Este paso de la dependencia estadística a la dependencia exacta implica que a cada valor de la variable independiente le asignemos uno solo de la variable dependiente. Ese valor de la variable dependiente, dado por la línea de regresión, tiene categoría de valor medio, pues como ya hemos indicado, la línea de regresión tiene ese carácter de línea media.

[1] Pese a que en el gráfico la línea media o línea de regresión se ha representado como una recta, la misma puede ser una curva cualquiera.

Figura 6. Diagrama de dispersión



Mediante este gráfico también es posible comprobar como cada valor de y_i observado se puede descomponer en dos partes. Una de ellas viene dada por el valor de la línea de regresión, $y_i^*=f(x_i)$, y la otra sería la diferencia entre el valor observado y el asignado por nuestra relación funcional exacta a la que llamaremos error o residuo, e_i. Formalmente tendríamos:

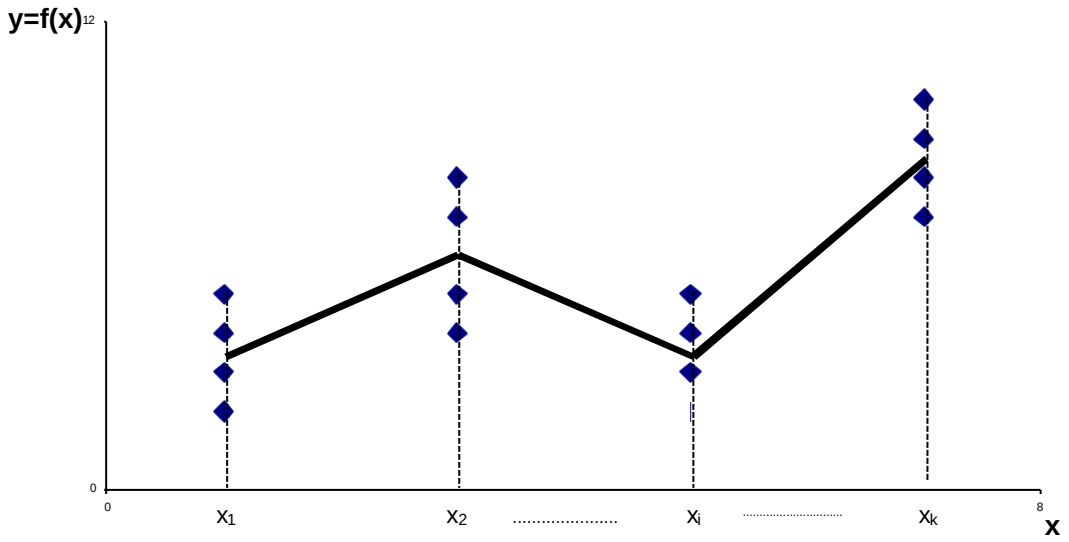
$$y_i = f(x_i) + e_i = y_i^* + e_i \,. \qquad (3.9)$$

En consecuencia el análisis de regresión lo que persigue es obtener los valores medios y^*_i de la variable dependiente que corresponden a los valores x_i observados.

El siguiente paso en el análisis de la regresión es definir los procedimientos que nos permitan obtener esa línea media que es la línea de regresión. No vamos a entrar a describir todos los posibles métodos que existen para determinar esa línea de regresión. Solo vamos a mencionar tres. El primero es el más sencillo y consiste en trazar la línea que más se ajuste a la nube de puntos. Este procedimiento gráfico, frente a su sencillez, tiene en su contra la falta de rigor.

Un segundo procedimiento consiste en sustituir todos los valores de *Y*, para un valor dado de X (x_i), por su media. Se trataría de una media condicional, $\left(x_i ; y \neq x_i \right)$ y habría tantas medias como valores tome la variable independiente. Con la unión de esos valores medios se tendría la línea de regresión.

Figura 7. Línea de regresión



El tercer método es aquel que hace uso de una función matemática para explicar la dependencia exacta existente de forma implícita entre las dos variables observadas. Haciendo uso de los símbolos ya utilizados, esa función, que es la línea de regresión, es

$$y^*_i = f(x_i) \qquad (3.10)$$

Esta relación, a parte de decirnos que la variable *Y* depende de la variable *X*, y de servir para describir la relación causal exacta, permite realizar predicciones de la variable dependiente conocidos los valores de la variable independiente. Pero esas predicciones tienen un carácter de valores medios, pues los errores o residuos son impredecibles.

El siguiente paso supone la elección de la función matemática *f(x)* que ha de ser nuestra línea de regresión. Se trata pues de elegir aquella función *f(x)* que describa de la forma

más adecuada la dependencia entre las variables. A esas funciones se les denomina genéricamente como modelos. Los modelos más sencillos son los siguientes:

a) Modelo lineal : $y^* = a + bx$.

b) Modelo parabólico de segundo grado: $y^* = a + bx + cx^2$.

c) Modelo potencial: $y^* = Ax^b$.

d) Modelo exponencial: $y^* = AB^x$.

e) Modelo hiperbólico: $y^* = a/x$

En cada caso, la elección de uno de ellos dependerá de lo que digan los datos (análisis empírico) o de lo que indique la teoría. En todos esos modelos hemos introducido, además de las variables independiente y dependiente, los símbolos *a, b, c, A y B*. Los mismos reciben el nombre de coeficientes ó parámetros. Una vez seleccionado el modelo que se ajusta a la línea de puntos o que responde a una teoría existente, lo que debemos realizar a continuación es definir un método que nos permita cuantificar esos coeficientes a partir de los datos observados. El procedimiento más utilizado es el denominado método de los mínimos cuadrados ordinarios (MCO).

3.4.1 Método de los mínimos cuadrados.

Este método consiste en determinar unos valores para los coeficientes ó parámetros de la función seleccionada, $y_i^* = f(x_i)$, con la condición de que haga mínima la suma de los errores al cuadrado, conforme se definieron en (3.9) ($e_i = (y_i - y_i^*)$), es decir:

$$\sum_i e_i^2 = \sum_i \left(y_i - y_i^*\right)^2 = mínimo \qquad (3.11)$$

De forma más general esta expresión se puede dar como

$$\sum_i e_i^2 n_i = \sum_i \left(y_i - y_i^*\right)^2 = mínimo \qquad (3.12)$$

es decir, admitir frecuencias mayores que uno para los distintos pares de valores de X e Y. Con este planteamiento, y para el caso de un modelo lineal, la función a minimizar sería

$$\theta(a,b) = \sum_i e_i^2 n_i = \sum_i \left(y_i - y_i^*\right)^2 n_i = \sum_i \left(y_i - a - bx_i\right)^2 n_i \qquad (3.13)$$

Si a esta función le aplicamos la primera condición de mínimo se llega al siguiente sistema:

$$\frac{\partial \theta(a,b)}{\partial a} = -2\sum_i \left(y_i - a - bx_i\right) n_i = 0 \qquad (3.14)$$

$$\frac{\partial \theta(a,b)}{\partial b} = -2\sum_i \left(y_i - a - bx_i\right) x_i n_i = 0 \qquad (3.15)$$

Estas dos ecuaciones se pueden expresar como:

$$\sum_i e_i n_i = 0 \qquad (3.16)$$

$$\sum_i e_i x_i n_i = 0 \qquad (3.17)$$

A partir de este sistema de ecuaciones se llega a otro donde se aprecia, de forma más clara, que las soluciones del mismo son funciones de los valores observados de las variables:

$$\sum_i y_i n_i = Na + b\sum_i x_i n_i \qquad (3.18)$$

$$\sum_i y_i x_i n_i = a\sum_i x_i n_i + b\sum_i x_i^2 n_i \qquad (3.19)$$

A estas ecuaciones se les conoce con el nombre de ecuaciones normales. A partir de ellas se pueden obtener unas fórmulas de cálculo de los parámetros de la recta. Esas fórmulas son:

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{S_{xy}}{S_x^2} \qquad (3.20)$$

$$a = \bar{y} - b\bar{x} \qquad (3.21)$$

Al parámetro *b* del modelo lineal se le llama también coeficiente de regresión. Este coeficiente mide la tangente o pendiente de la recta. Su signo será el de la covariancia, que a su vez le daba el signo al coeficiente de correlación. Así pues cuando la relación es directa, el signo será positivo y cuando al relación es negativa o inversa el signo será negativo. Este coeficiente mide la variación de la variable *Y* frente un cambio unitario de la variable *X*. También se puede interpretar como la razón de una progresión aritmética.

Al parámetro *a* del modelo lineal se le conoce como término independiente u ordenada en el origen. Nos daría el valor de la variable dependiente cuando la independiente valiese cero.

De igual forma que se han obtenido las ecuaciones normales para este modelo lineal de dos variables (*Y*, *X*), también es posible llegar a un sistema de ecuaciones normales para cada uno de los modelos indicados anteriormente o para un modelo lineal donde el número de variables explicativas sea mayor que uno. Cada sistema tendrá tantas ecuaciones como parámetros existan en la línea de regresión. Para el caso del modelo parabólico las ecuaciones normales son:

$$\sum_i y_i n_i = Na + b\sum_i x_i n_i + c\sum_i x_i^2 n_i$$

$$\sum_i y_i x_i n_i = a\sum_i x_i n_i + b\sum_i x_i^2 n_i + c\sum_i x_i^3 n_i$$

$$\sum_i y_i x_i^2 n_i = a\sum_i x_i^2 n_i + b\sum_i x_i^3 n_i + c\sum_i x_i^4 n_i$$

A este modelo se le puede considerar como un modelo lineal de tres variables, una dependiente (*Y*) y dos explicativas o independientes (*X* , X^2), pues el mismo se podría haber formulado como: $y_i^* = a + bx_1 + cx_2$, donde $x_1=x$ y $x_2=x^2$. Tanto a este modelo, como a los otros señalados, se les conoce como modelos no lineales pero linealizables mediante las transformaciones adecuadas. Así, si en el modelo potencial, que como se ha visto viene dado por $y^* = Ax^b$, se toman logaritmos, entonces el nuevo modelo es lineal, y vendría dado por: $lny^* = lnA + blnx = a' + bx'$. En este modelo, el parámetro *b* es la elasticidad de *Y* con respeto de *X*, que es, además, constante, pues:

$$E = \frac{\partial y}{\partial x}\frac{x}{y} = Abx^{b-1}\frac{x}{Ax^b} = b \qquad (3.22).$$

De forma similar se podría proceder con el modelo exponencial $y^*=AB^x$. En este caso la linealización del mismo se obtendría, también, mediante el uso de logaritmos: $lny^*=lnA+xlnB=a'+b'x$. Ahora en este modelo, el coeficiente *b* es la razón de una progresión geométrica y a partir del mismo puede obtenerse la tasa de crecimiento de la variable *Y*.

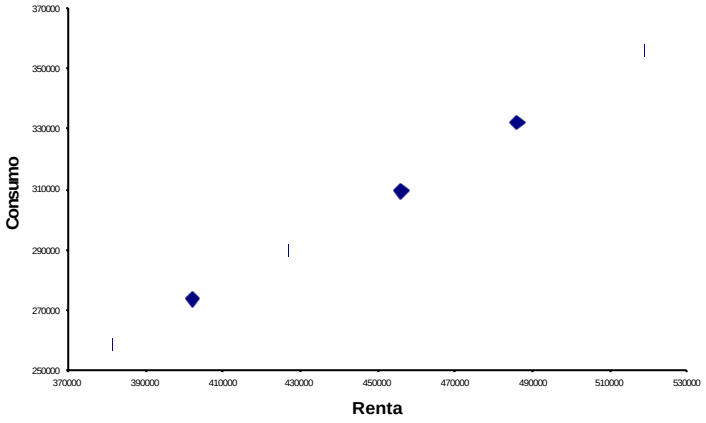
Ejemplo 6. *En la tabla siguiente se recoge la evolución, para el periodo 1999-2000, de dos de las principales macromagnitudes de la Economía de España. La Renta Nacional Disponible Neta a precios corrientes así como el Gasto en Consumo Final de los Hogares, expresadas las dos en miles millones de euros. Con estos datos: a) ajuste la función agregada de consumo, dando una interpretación del significado de los coeficientes obtenidos; b) obtenga la elasticidad del consumo para una renta de 450 miles de millones de euros así como la propensión media al consumo para esa renta.*

	Renta	*Consumo*
1995	*381,5*	*258,6*
1996	*402,3*	*273,6*
1997	*426,9*	*289,7*
1998	*456,1*	*309,3*
1999	*486,0*	*331,8*
2000	*519,0*	*356,2*

Fuente: Web INE

a) Antes de realizar ajuste alguno hay que preguntarse por la relación funcional de la función agregada de consumo, pues la Teoría Económica, de entrada, solo dice que el consumo de las familias es una función del la renta, sin especificar mucho más. Sin embargo, la teoría keynesiana señala que la relación entre ambas variables es de tipo lineal. En cualquier caso parece aconsejable realizar un análisis gráfico exploratorio previo que confirme o desmienta ese planteamiento teórico. A tal efecto se ha realizado la Figura 8 donde se aprecia que, al menos a corto plazo, el planteamiento keynesiano no es del todo erróneo, pues las seis parejas de valores están casi alineadas. Esto nos permite ensayar un ajuste lineal. Este resultado se puede confirmar calculando el coeficiente de correlación lineal.

Figura 8. Relación entre la renta disponible y el consumo de los hogares en España. Periodo 1995-2000



Los cálculos necesarios para su realización son los que aparecen a continuación.

	Renta (X)	Consumo (Y)	X^2	XY	Y^2
1995	381,534	258,647	145568,2	98682,6	66898,3
1996	402,283	273,561	161831,6	110048,9	74835,6
1997	426,908	289,675	182250,4	123664,6	83911,6
1998	456,102	309,279	208029,0	141062,8	95653,5
1999	486,049	331,825	236243,6	161283,2	110107,8
2000	518,999	356,225	269360,0	184880,4	126896,3
Total	**2671,875**	**1819,212**	**1203282,9**	**819622,5**	**558303,1**

$$S_{xy} = \frac{\sum_i x_i y_i}{N} - \frac{\sum_i x_i \sum_i y_i}{N \quad N} = \frac{819622,5}{6} - \frac{2671,9}{6}\frac{1819,2}{6} = 1584,1$$

$$S_x = \sqrt{\frac{\sum_i x_i^2}{N} - \left(\bar{x}\right)^2} = \sqrt{\frac{1203282,9}{6} - \left(\frac{2671,9}{6}\right)^2} = 47,4$$

$$S_y = \sqrt{\frac{\sum_i y_i^2}{N} - \left(\bar{y}\right)^2} = \sqrt{\frac{558303,1}{6} - \left(\frac{1819,2}{6}\right)^2} = 33,5$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{1584,1}{(47,4)(33,5)} = 0,99966898$$

Como puede observarse, la relación lineal entre estas dos variables es muy fuerte.

$$b = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2} = \frac{1584,1}{2243,9} = 0,706$$

$$a = \bar{y} - b\bar{x} = 303,2 - (0,706)(445,3) = -11,2$$

Todos estos cálculos nos llevan a que, finalmente, la función lineal ajustada pueda expresarse como:

$$y_i = y^*_i + e_i = -11,2 + 0,706x_i + e_i$$

En este ajuste, donde la variable dependiente es el consumo agregado y la independiente la renta disponible, a la pendiente de la recta se le conoce como propensión marginal al consumo, indicando la proporción que de cada unidad monetaria de renta se dedica al consumo. En este contexto se habla de proporción porque se asume la identidad fundamental de la renta, según la cual una unidad de renta cualquiera solo puede destinarse al consumo o al ahorro. Así, para esta función agregada de consumo se tiene que la diferencia entre la unidad y la propensión marginal al consumo es la propensión marginal al ahorro.

A la ordenada en el origen o término independiente de esta función ajustada se le conoce como consumo autónomo. Es decir, se trataría de un consumo mínimo o de subsistencia correspondiente a niveles de renta nulos. En nuestro caso el signo de este coeficiente es negativo, lo que económicamente carece de sentido, pues el consumo deberá ser siempre mayor o igual que cero. La razón de este "absurdo" económico hay que buscarla en que en este ejemplo la ordena en el origen, desde el punto de vista estadístico, es una predicción, pues nos da el valor del consumo bajo el supuesto de que la renta fuera nula. Pero esos valores de la renta no han sido observados. Esto nos lleva a plantearnos nuevas cuestiones tales como la calidad del ajuste y la fiabilidad de las posibles predicciones que se puedan realizar con un modelo de regresión. Pero esto son cuestiones que se irán respondiendo en los siguientes epígrafes de es capitulo.

b) Sabemos que la elasticidad se define como

$$E = \frac{\partial y}{\partial x}\frac{x}{y}$$

En nuestro caso tendremos que:

$$E = \frac{\partial y}{\partial x}\frac{x}{y} = (0{,}706)\ \frac{450}{306{,}5} = 1{,}0365$$

lo que significa que cuando la renta varía en 1%, en un entorno próximo a los 450 miles de millones de euros, el consumo cambia en 1,0365%.

A su vez la propensión media al consumo se define como:

$$PMC = \frac{y^*}{x} = \frac{306{,}5}{450} = 0{,}6811$$

3.4.2. Regresión multivariante.

En los epígrafes anteriores se ha hablado de distribuciones bivariantes, en las que se estudiaba la distribución conjunta de dos variables, las cuales eran el resultado de observar dos caracteres de una población. Pero los caracteres observables de una población no tienen porqué limitarse a solo dos. Mas bien al contrario. Lo normal es que puedan observarse más de dos. En estos casos se tendría, también, una distribución de frecuencias conjuntas cuya representación mediante una tabla de doble entrada se hace difícil (caso de tres variables) o no puede realizarse (para más de tres variables). Pero estas dificult ades que plantea su tabulación no impiden que la técnica del análisis de regresión vista anteriormente no le sea aplicable. Por el contrario, es precisamente para este tipo de situaciones donde la regresión se convierte en un instrumental de análisis realmente potente. Pero cuando lo que se pretende ajustar es una función a una nube de puntos de más de dos dimensiones, entonces el procedimiento descrito anteriormente para determinar los coeficientes de la línea de regresión, basado en la obtención de un sistema de ecuaciones (ecuaciones normales), resulta poco operativo, pues la resolución del mismo, cuando se tienen muchas incógnitas (muchos coeficientes de regresión), se hace tedioso por el elevado número de ecuaciones del que puede constar. En estos casos lo que se hace es recurrir al álgebra matricial. Para centrar un poco las ideas, supongamos que tenemos una variable *Y* que depende de *k* variables explicativas *X*. Para estas *k+1* variables (las *k* independientes más ordenada en el origen) se realizan *N* observaciones. Si admitimos que la relación funcional entre la variable dependiente y las independientes o explicativas es de tipo lineal, entonces tendríamos la función:

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots\ldots\ldots\ldots + b_k x_{ki} + e_i = y_i^* + e_i \qquad (3.23)$$

Pero este modelo, en términos matriciales, y para todas las observaciones realizadas se puede expresar como:

$$\mathbf{y = Xb + e = y^* + e} \qquad (3.24)$$

donde **y** es un vector de dimensión *Nx1,* correspondiente a las *N* observaciones de la variable *Y*, **X** es una matriz *Nx(k+1)* correspondiente a las *N* observaciones de las *k+1* variables (las *k* variables explicativas más un vector columna de unos correspondiente al

términos independiente *a*), **b** es un vector *(k+1)x1* de coeficientes y **e** es el vector *Nx1* de errores. De forma expandida el modelo quería como:

$$y_1 = a + b_1 x_{11} + b_2 x_{21} + \ldots\ldots\ldots\ldots + b_k x_{k1} + e_1$$
$$y_2 = a + b_1 x_{12} + b_2 x_{22} + \ldots\ldots\ldots\ldots + b_k x_{k2} + e_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_N = a + b_1 x_{1N} + b_2 x_{2N} + \ldots\ldots\ldots\ldots + b_k x_{kN} + e_N$$

O bien como:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots\ldots & x_{k1} \\ 1 & x_{12} & \ldots\ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1N} & \ldots\ldots & x_{kN} \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ \ldots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ e_3 \end{bmatrix} = \mathbf{Xb} + \mathbf{e} \qquad (3.25)$$

Ahora, si lo que se pretende obtener es el vector de coeficientes **b** por mínimos cuadrados, entonces hay seleccionar aquel **b** que minimice la suma de cuadrados de los errores dada por:

$$\mathbf{e'e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) = \mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'Xb} \qquad (3.26)$$

Si esta suma de cuadrados la derivamos respecto del vector **b** se obtiene la condición necesaria para que esa función sea mínima:

$$\partial(\mathbf{e'e})/\partial\mathbf{b} = -2\mathbf{X'y} + 2\mathbf{X'Xb} \qquad (3.27)$$

De este resultado se deduce que la condición necesaria de mínimo buscada es:

$$\mathbf{X'Xb} = \mathbf{X'y} \qquad (3.28)$$

Por lo que el vector de coeficientes **b** vendrá dado por:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad (3.29)$$

De la condición necesaria para minimizar la suma de cuadrados se deduce que:

$$\mathbf{X'Xb} = \mathbf{X'y} = \mathbf{X'Xb} + \mathbf{X'e} \qquad (3.30)$$

Es decir, que $\mathbf{X'e} = \mathbf{0}$. Este es el mismo resultado que se obtuvo en el caso de dos variables, donde se vio que $\sum_i e_i n_i = 0$ y que $\sum_i e_i x_i n_i = 0$. Estos resultados nos llevan también a que: $\sum_i y_i = \sum_i y_i^*$ en el caso multivariante, pues la suma de los errores es cero, y, en consecuencia, la media de los valores observados de *y* es igual a la media de los valores ajustados *y**.

Ejemplo 7. *Para un conjunto de empresas, de características similares, se han obtenido los siguientes datos de producción y costes medios totales. Ajuste el modelo más adecuado a esos datos en que los costes sean función de la producción.*

Producción	*Costes medios*
2	*7*
7	*3*
5	*4*
6	*3.5*
4	*5.5*
8	*3.5*
3	*5*
10	*4*
12	*5*
15	*6*

En primer lugar hay que seleccionar la forma funcional que relaciona los costes medios de una empresa con sus niveles de producción. Para ello se puede recurrir a dos vías distintas. Por un lado está el análisis gráfico y por otro está la información que pueda suministrar la teoría (en nuestro caso la teoría económica). En general, será siempre preferible recurrir a la segunda fuente de información, pues el método gráfico nos puede llevar a cuantificar relaciones funcionales que se ajusten muy bien a los datos observados cuando, en realidad, se está trabajando con variables que no están relacionadas entre si. Se trataría del caso de covariación casual o espúrea. Al recurso gráfico hay que acudir

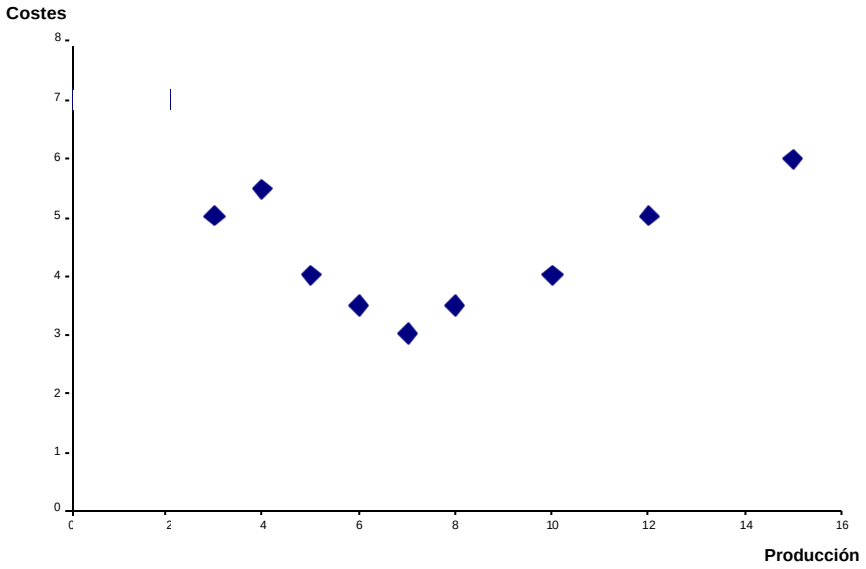
cuando no existe teoría alguna que nos informe sobre la posible relación existente entre esas variables.

Según los criterios establecidos en el párrafo anterior, lo primero que habría que hacer es indagar, desde la microeconomía, que tipo de relación hay entre los costes medios de una empresa y su producción. En este caso, la teoría nos dice que los costes medios decrecen inicialmente (debido a la caída de los costes medios fijos) hasta que se alcanza un cierto nivel de producción y después crecen. Así pues sería poco razonable pensar en un modelo lineal como el más adecuado para este ejemplo. Para confirmar lo que nos dice la teoría se puede recurrir, de forma complementaria, al análisis gráfico. En este caso, los datos de esas empresas está representados en la Figura 9.

Como puede comprobarse, y según indicaba la teoría, la relación no es lineal. Sin embargo, y a titulo de mero ejercicio, se va a proceder al ajuste lineal de estos datos mediante mínimos cuadrados. En primer lugar se calculará el coeficiente de correlación lineal, el cual nos indicará si la covariación lineal entre esas variables es fuerte o débil. Para obtener este coeficiente, y los demás que iremos obteniendo, resulta aconsejable construir la tabla siguiente:

xy	x^2y	x^2	x^3	x^4
14	28	4	8	16
21	147	49	343	2401
20	100	25	125	625
21	126	36	216	1296
22	88	16	64	256
28	224	64	512	4096
15	45	9	27	81
40	400	100	1000	10000
60	720	144	1728	20736
90	1350	225	3375	50625
331	**3228**	**672**	**7398**	**90132**

Figura 9. Relación costes-producción



En la misma aparece recogida toda la información necesaria para la resolución de este ejemplo.

El coeficiente de correlación lineal, como sabemos, es igual al cociente de la covariancia entre el producto de las desviaciones estándares.

$$S_{xy} = \frac{\sum_i x_i y_i n_i}{N} - \frac{\sum_i x_i n_i \sum_i y_i n_i}{N\ N} = \frac{331}{10} - \frac{72}{10}\frac{46,5}{10} = -0,38$$

$$S_x = \sqrt{\frac{\sum_i x_i^2 n}{N} - \left(\bar{x}\right)^2} = \sqrt{\frac{672}{10} - \left(\frac{72}{10}\right)^2} = 3,919$$

$$S_y = \sqrt{\frac{\sum_i y_i^2 n}{N} - \left(\bar{y}\right)^2} = \sqrt{\frac{230,75}{10} - \left(\frac{46,5}{10}\right)^2} = 1,205$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{-0,38}{(3,919)(1,205)} = -0,08$$

Como puede verse, la relación lineal entre estas dos variables es muy débil y el signo del coeficiente de correlación es intranscendente en este caso, pues la curva tiene dos ramas, una decreciente y la otra creciente.

Si a pesar de todos los indicios detectados hasta ahora contra la relación lineal se siguiera adelante con este tipo de ajuste, los resultados serían los siguientes:

$$b = \frac{Cov(x, y)}{Var(x)} = \frac{S_{xy}}{S_x^2} = \frac{-0{,}38}{15{,}36} = -0{,}0247$$

$$a = \bar{y} - b\bar{x} = 4{,}65 - (-0{,}0247)(7{,}2) = 4{,}828$$

$$y_i = y^*_i + e_i = 4{,}828 - 0{,}0247x_i + e_i$$

Este ajuste lineal, como puede comprobarse por el signo negativo del coeficiente de regresión b, solo recoge el tramo decreciente de la curva, aquel donde los costes fijos medios prevalecen sobre los costes variables medios. En cambio no es capaz de detectar la rama creciente de la curva.

Si en lugar de ajustar una función lineal se trabajara con una función parabólica, entonces el sistema de ecuaciones normales asociado a ese modelo sería el siguiente:

$$46{,}5 = 10a + 72b + 672c$$
$$331 = 72a + 672b + 7398c$$
$$3228 = 672a + 7398b + 90132c$$

con lo que los parámetros o coeficientes calculados son:

$$a = 8{,}616 \quad b = -1{,}22 \quad c = 0{,}07177$$

de forma que el modelo ajustado quedaría:

$$y = y_i^* + e_i = 8{,}616 - 1{,}22\, x_i + 0{,}07177\, x_i^2 + e_i$$

Al mismo resultado se habría llegado si en lugar de resolver ese sistema de ecuaciones se hubiera trabajado en términos matriciales. En este caso la matriz ***X*** *y el vector* ***y*** *vienen dados por:*

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 7 & 49 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 4 & 16 \\ 1 & 8 & 64 \\ 1 & 3 & 9 \\ 1 & 10 & 100 \\ 1 & 12 & 14 \\ 1 & 15 & 225 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ 3 \\ 4 \\ 3{,}5 \\ 5{,}5 \\ 3{,}5 \\ 5 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

y a partir de ellas se obtiene que:

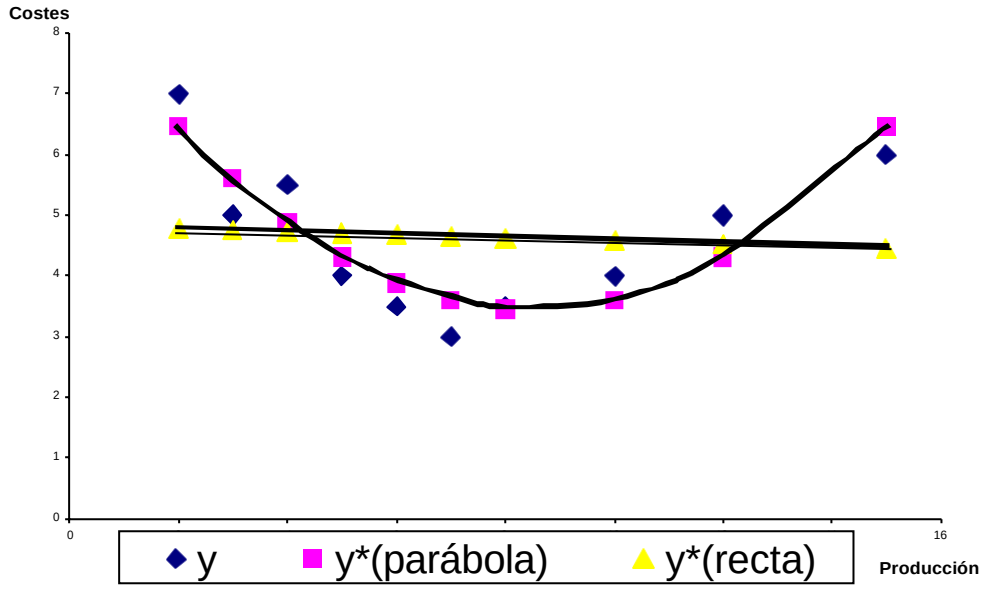
$$(\mathbf{X'X}) = \begin{bmatrix} n & \sum_i x_{1i} & \sum_i x_{2i} \\ \sum_i x_{1i} & \sum_i x_{1i}^2 & \sum_i x_{1i} x_{2i} \\ \sum_i x_{2i} & \sum_i x_{1i} x_{2i} & \sum_i x_{2i}^2 \end{bmatrix} = \begin{bmatrix} 10 & 72 & 672 \\ 72 & 672 & 7398 \\ 672 & 7398 & 90132 \end{bmatrix}$$

$$\mathbf{X'y} = \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_{1i} \\ \sum_i y_i x_{2i} \end{bmatrix} = \begin{bmatrix} 46{,}5 \\ 331 \\ 3228 \end{bmatrix}$$

donde x_1 son los costes y x_2 son los costes al cuadrado.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1,63816 & -0,42595 & 0,02275 \\ -0,42595 & 0,12619 & -0,00718 \\ 0,02275 & -0,00718 & 0,00043 \end{pmatrix} \begin{pmatrix} 46,5 \\ 331 \\ 3228 \end{pmatrix} = \begin{pmatrix} 8,6162 \\ -1,2207 \\ 0,0718 \end{pmatrix}$$

Figura 10. Valores observados y ajustados



3.4.3 Variancia residual y coeficiente de determinación.

Si ahora quisiéramos analizar cual de los dos modelos se ajusta mejor a los datos de partida se podría proceder a la inspección gráfica (Figura 10) o a calcular los errores que se cometen con los dos modelos.

Como puede apreciarse con estos resultados, tanto gráfico como numérico, la parábola representa un ajuste mejor que la línea recta, pues los errores que se comenten con la recta son mayores que con la parábola. Pero hay que fijarse no en los errores de cada uno de los modelos, pues como sabemos la suma de ellos es siempre cero de acuerdo con la primera de las ecuaciones normales, sino en el cuadrado de los mismos que es lo que se pretende hacer mínimo. En este sentido diremos que un modelo es tanto mejor

cuanto menor sea la suma de los cuadrados de sus errores, y más concretamente la media de esa suma. Este es precisamente el criterio que habrá que usar para medir la bondad o representatividad de un modelo.

		Modelo lineal			Modelo parabólico		
x_i	y_i	y^*	e_i	e_i^2	y^*	e_i	e_i^2
2	7	4,7786	2,2214	4,9344	6,4619	0,5381	0,2896
3	5	4,7539	0,2461	0,0606	5,6000	-0,6000	0,3600
4	5,5	4,7292	0,7708	0,5942	4,8817	0,6183	0,3823
5	4	4,7044	-0,7044	0,4962	4,3069	-0,3069	0,0942
6	3,5	4,6797	-1,1797	1,3917	3,8756	-0,3756	0,1411
7	3	4,6549	-1,6549	2,7389	3,5879	-0,5879	0,3457
8	3,5	4,6302	-1,1302	1,2774	3,4438	0,0562	0,0032
10	4	4,5807	-0,5807	0,3372	3,5861	0,4139	0,1714
12	5	4,5313	0,4688	0,2197	4,3025	0,6975	0,4865
15	6	4,4570	1,5430	2,3808	6,4537	-0,4537	0,2058
			0,0000	**14,4310**		**0,0000**	**2,4797**

Esta forma de proceder descansa en dos argumentos. El primero es el indicado antes, es decir, la selección del modelo adecuado ha de realizarse en función de que la media del cuadrado de sus errores sea mínima. En segundo lugar hay que recordar que la línea de regresión (y^*) se ha definido como una línea media que trata de resumir toda la nube de puntos. Pues bien, como tal media es aconsejable acompañarla de una indicador que mida su representatividad o bondad de ajuste, al igual que se hacía con la media aritmética y la variancia. La mayor o menor bondad dependerá de que las desviaciones de los valores observados de Y con respecto a los que se obtienen mediante la línea de regresión sean pequeñas o grandes. Si esas desviaciones son pequeñas la bondad será alta. Por el contrario, si las desviaciones son grandes la bondad será pequeña. Lo ideal es que esas desviaciones fueran siempre nulas. Pero como la suma de esas desviaciones es siempre nula (la suma de esas desviaciones es simplemente la suma de los errores) entonces la media de esa suma no nos sirve. Por esa razón, y de forma similar a como se procedió con la media y la variancia, el indicador de la bondad del ajuste se basará en la media del cuadrado de los errores. Este promedio se define como la media cuadrática de las desviaciones de los valores observados de Y respecto de sus valores "medios" Y^*. Esta media cuadrática, que sería una variancia, quedaría como:

$$S_e^2 = \frac{\sum_i \left(y_i - y_{i}^{*}\right)_2}{N} = \frac{\sum_i e_i^2}{N} \qquad (3.31)$$

A esta media cuadrática se le llama variancia residual. Se le llama así porque a los errores e_i se les conoce también como residuos. Esta variancia no es la misma que la ya definida para Y (S_y^2), pues en un caso las diferencias se toman con respecto a $\bar{Y}$ y no con respecto a Y*.

La anter ior definición de variancia residual es válida para cualquier ajuste, sin importar el tipo de modelo, lineal o no. Ahora bien, el cálculo de la misma dependerá del modelo con el que se esté trabajando. En el caso lineal para dos variables la variancia residual es:

$$S_e^2 = \frac{\sum_i \left(y_i - y_i^*\right)^2}{N} = \frac{\sum_i e_i^2}{N} = \frac{\sum_i e_i \left(y_i - a - bx_i\right)}{N} = \frac{\sum_i e_i y_i - a\sum_i e_i - b\sum_i e_i x_i}{N} = \frac{\sum_i e_i y_i}{N} =$$

$$= \frac{\sum_i \left(y_i - a - bx_i\right) y_i}{N} = \frac{\sum_i y_{i}^{2} - a\sum_i y_i - b\sum_i x_i y_i}{N} \qquad (3.32)$$

De forma similar se llega a que la variancia residual para el caso de la parábola es:

$$S_e^2 = \frac{{}_i \sum_i \sum_i \sum_i}{N} \qquad (3.33)$$

Finalmente, esa variancia, para el caso lineal multivariante, viene dada por:

$$S_e^2 = \frac{\mathbf{e'e}}{N} = \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'Xb}}{N} = \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'X}\left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}}{N} =$$

$$= \frac{\mathbf{y'y} - 2\mathbf{b'X'y} + \mathbf{b'X'y}}{N} = \frac{\mathbf{y'y} - \mathbf{b'X'y}}{N} \qquad (3.34)$$

A la raíz cuadrada de esta variancia se le conoce como error estándar del ajuste y es el equivalente a la desviación estándar ya definida previamente. Como tal, nos da el tamaño medio de los errores del ajuste. Las unidades de medida de este error estándar son las de la variable *Y*. Pero el hecho de que este indicador de los errores del ajuste no sea adimensional impide realizar comparaciones cuando se trabaja con variables dependientes de distinta naturaleza. Este problema se resolvió haciendo uso del coeficiente de variación cuando se trabajaba con una sola variable. Ahora, en el contexto de la regresión, el problema se resuelve recurriendo a lo que se conoce como coeficiente de determinación**.** Este coeficiente se utiliza para estudiar la representatividad de la línea de regresión o bondad del ajuste.

Debe recordarse que el objetivo en el análisis de la regresión es explicar las variaciones observadas en la variable *Y* mediante las variaciones de la variable explicativa *X*. Como ya se ha podido comprobar, la variable *X* no es capaz, por si sola, de explicar todas las variaciones de la *Y*, por lo que se admite la posibilidad de cometer un error, de manera que los valores de *Y,* como se indicó en (3.9) y (3.24), se pueden descomponer en dos términos:

$$y_i = y^*_i + e_i. \qquad (3.35)$$

Partiendo de esa relación se va a demostrar que se cumple la siguiente igualdad:

$$S_y^2 = S^2_{y^*} + S_e^2 \qquad (3.36)$$

La demostración es como sigue. En (3.35) se le resta la media de *Y* en ambos lados de la igualdad:

$$(y_i - \bar{y}) = (y^*_i - \bar{y}) + e_i \qquad (3.37)$$

A continuación se eleva al cuadrado (3.37) y se suma para las *N* observaciones resultado que:

$$\sum_i \left(y_i - \bar{y}\right)^2 = \sum_i \left(\left(y_i^* - \bar{y}\right) + e_i\right)^2 = \sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i e_i^2 + 2\sum_i \left(y_i^* - \bar{y}\right)e_i =$$
$$= \sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i e_i^2 \qquad (3.38)$$

pues $2\sum_i \left(y_i^* - \bar{y}\right)e_i$ vale cero. Si ahora dividimos todo por *N* se llega a:

$$S_y^2 = \frac{\sum_i \left(y_i - \bar{y}\right)^2}{N} = \frac{\sum_i \left(y_i^* - \bar{y}\right)^2}{N} + \frac{\sum_i e_i^2}{N} = S_{y^*}^2 + S_e^2 \qquad (3.39)$$

y este es precisamente el resultado al que se quería llegar.

Esta descomposición dada en (3.39) se puede interpretar como que la variancia total ($S^2)_y$ es la suma de la variancia explicada por el modelo ($S^2)_{y^*}$ más la variancia residual ($S^2)_e$. A partir de esta relación entre variancias, y dado que todas ellas serán siempre no negativas, si dividimos ambos miembros de la igualdad por al variancia total y ordenamos términos, tendremos que:

$$\frac{S_{y^*}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \qquad (3.40)$$

es decir, que la proporción de variancia explicada por el modelo respecto del total será igual a la unidad menos la proporción no explicada por el modelo. Pues bien, a esa proporción explicada por el modelo se le conoce como coeficiente de determinación (R^2), es decir:

$$R^2 = \frac{S_{y^*}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \qquad (3.41)$$

Como se ha indicado antes, al ser la variancias magnitudes no negativas, el coeficiente de determinación será siempre mayor o igual que cero. Por otro lado, como la variancia de *Y**

es una parte de la variancia total, resultará que *los valores del coeficiente de determinación estarán comprendidos siempre entre cero y uno.* El valor cero lo tomará cuando la variancia explicada por el modelo sea cero, en cuyo caso diremos que el modelo seleccionado para el ajuste no es el adecuado, pues las variaciones de *X* no explican ninguna de las variaciones de *Y*. Por el contrario, cuando toma el valor uno ello implica que la variancia residual es nula. En tal caso la dependencia estadística se convierte en dependencia exacta. Esta es la situación que se da cuando todas las variaciones de la variable dependiente quedan perfectamente explicadas por las variaciones de la variable explicativa.

Todo lo anterior nos lleva a decir que el coeficiente de determinación da la proporción de la variación total de la variable dependiente que viene explicada por el modelo o la variable explicativa. Así cuando el coeficiente de determinación tome valores próximos a la unidad diremos que la bondad del ajuste es muy buena o que el modelo seleccionado para el ajuste es representativo, ya que explica una elevada proporción de las variaciones de la variable dependiente. Por todo ello resulta aconsejable que cada vez que realicemos un ajuste lo acompañemos de una medida de su bondad, y esta medida es el coeficiente de determinación.

El coeficiente de determinación definido para el caso lineal de dos variables se puede generalizar al modelo de *k* variables. También en este caso la variancia total se puede descomponer en la suma de la variancia explicada por el modelo más la variancia residual. Concretamente, el coeficiente de determinación vendrá dado por:

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{\mathbf{y'y} - \mathbf{b'X'y}}{\mathbf{y'y} - N\bar{y}^2} = \frac{\mathbf{b'X'y} - N\bar{y}^2}{\mathbf{y'y} - N\bar{y}^2} \qquad (3.42)$$

Ejemplo 8. *Obtener el coeficiente de determinación para los ajustes realizados en el Ejemplo 7.*

a) Para el caso de la recta:

$$S_e^2 = \frac{\sum_i i \quad \sum_i i \quad \sum_i i\, i}{N} = \frac{230\,,75 - 4{,}828\,(46{,}5) + 0{,}0247\,(331)}{10} = 1{,}44237$$

$$S_y^2 = 1{,}452499$$

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 0{,}00697$$

y = 4,828 – 0,0247x;* $R^2 = 0{,}00697$

b) Para el caso de la parábola el resultado es

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 0{,}829$$

y = 8,616 – 1,22x + 0,07177x²;* $R^2 = 0{,}829$

Como se puede comprobar el modelo lineal es una mala elección, pues no llega a explicar ni siquiera el 1% de las variaciones de la variable dependiente. En cambio la parábola explica más del 82%, lo que nos permite calificarlo como un modelo bastante adecuado. Es decir, en este caso la bondad del ajuste es bastante buena, aunque mejorable, pues hay más de un 17% de variaciones no explicadas por el modelo.

A continuación vamos a mostrar una relación muy interesante que existe entre el coeficiente de determinación y el de correlación para el caso lineal. En esta situación se cumple que el coeficiente de determinación es igual al de correlación al cuadrado, es decir, $R^2 = r^2$. Para comprobar que esto es cierto debemos recordar que:

$$R^2 = \frac{S^2_{y^*}}{S^2_y} = \frac{\sum_i \left(y^*_i - \bar{y}\right)^2}{N} = \frac{\sum_i \left(a + bx_i - a - b\bar{x}\right)^2}{N} = \frac{b^2 \sum_i \left(x_i - \bar{x}\right)^2}{N} = b^2 \frac{S^2_x}{S^2_y} = \frac{S^2_{xy}}{S^2_x S^2_y} S^2_x = \frac{S^2_{xy} S^2_x}{\left(S^2_x\right)^2 S^2_y} =$$

$$= \frac{S^2_{xy}}{S^2_x S^2_y} = r^2. \qquad (3.43)$$

Ejemplo 9. *Para el modelo de regresión lineal $y_i = f(x_i) + e_i = y_i^* + e_i = a + bx_i + e_i$, demuestre que:*

a) $e_i = y_i - \bar{y} - b(x_i - \bar{x})$

b) $\sum_i e_i = 0$

c) $\sum_i e_i^2 = \sum_i \left(y_i - \bar{y}\right)^2 - b^2 \sum_i \left(x_i - \bar{x}\right)^2$

d) $y_i^* - \bar{y} = b\left(x_i - \bar{x}\right)$

e) $\sum_i e_i \left(x_i - \bar{x}\right) = 0$

Demostración:

a) $e_i = y_i - y_i^* = y_i - \left(a + bx_i\right) = y_i - \left(\bar{y} - b\bar{x} + bx_i\right) = y_i - \bar{y} - b\left(x_i - \bar{x}\right)$

b) $\sum_i e_i = \sum_i y_i - N\bar{y} - b\sum_i \left(x_i - \bar{x}\right) = \sum_i y_i - N\bar{y} = \sum_i y_i - \sum_i y_i = 0$

c) A partir del resultado obtenido en (3.38) y de uno de los pasos intermedios de (3.43) queda demostrada la igualdad planteada en este apartado.

d) $y_i^* - \bar{y} = a + bx_i - \bar{y} = \bar{y} - b\bar{x} + bx_i - \bar{y} = b\left(x_i - \bar{x}\right)$

e) Este resultado es una consecuencia inmediata de (3.16) y (3.17), pues la suma dada en este apartado se puede poner como:

$$\sum_i e_i \left(x_i - \bar{x}\right) = \sum_i e_i x_i - \bar{x}\sum_i e_i = 0$$

Ejemplo 10. *Analice la veracidad de los siguientes resultados cuando se trabaja con un modelo de regresión lineal de dos variables.*

a) $r_{xy}= 0,5 \quad b=0,7 \quad R^2=0,9$

b) $b=0,7 \quad S_{xy} = -10$

c) $S_y^2 = 15 \quad S^2_{y^*}=20$

d) $S_y^2 = 15 \quad S^2_{y^*} = 10 \quad S_e^2 = 5$

a) Estos resultados no pueden ser ciertos pues, aunque el signo del coeficiente de correlación lineal y el de la pendiente de la recta coinciden, dado que tienen el mismo numerador (la covariancia) y sus denominadores serán siempre no negativos, sin embargo no se cumple la relación que existe entre el coeficiente de correlación y el de determinación, que, como se ha demostrado, es: $R^2 = r^2$.

b) Este resultado tampoco es posible. El signo de b y el de S_{xy} *debe ser siempre el mismo.*

c) Teniendo en cuenta que en un modelo de regresión lineal de dos variables se cumple (3.36) y que la variancia no puede ser negativa, resulta que este resultado es imposible. Además la variancia explicada por el modelo nunca podrá ser mayor que la total.

d) Este resultado si es cierto, pues satisface la relación (3.36).

Ejemplo 11. *A veces, la relación entre las variables es de tal naturaleza que la línea de regresión debe pasar por el origen. En tales casos lo que se hace es imponer que la ordenada en el origen sea nula. Esto es equivale a introducir una restricción al modelo lineal de dos variables. Cuando se procede de esta forma algunos de los resultados obtenidos anteriormente dejan ser válidos. Veamos a continuación como afecta esta restricción a los resultados obtenidos en epígrafes anteriores.*

Para empezar hay que señalar que nuestro modelo es ahora el siguiente:

$$y_i = y^*_i + e_i = bx_i + e_i.$$

Como en este modelo solo hay un parámetro desconocido, entonces la aplicación del método de los mínimos cuadrados nos llevará a una sola ecuación normal

$$\theta(b)=\sum_i e_i^2=\sum_i \left(y_i - y_i^*\right)^2 = \sum_i \left(y_i - bx_i\right)^2$$

$$\frac{\partial \theta(b)}{\partial b}=-2\sum_i \left(y_i - bx_i\right)x_i = 0$$

De estos resultados se deduce que ahora:

$$b = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Además, como solo se trabaja con una ecuación normal, la segunda, resulta que la primera ni siquiera se cumple. Es decir, la suma de los errores ya no es cero, como comprobaremos a continuación.

$$\sum_i e_i = \sum_i \left(y_i - y^*\right) = \sum_i y_i - b\sum_i x_i$$

Pero para que esta expresión fuera cero es necesario que se cumpla que

$$b = \frac{\sum_i y_i}{\sum_i x_i}$$

lo cual no es cierto, como se ha visto en líneas anteriores. Así pues, la suma de los errores en este tipo de modelo lineal que estamos analizando no es cero. Una consecuencia de este resultado es que la media de los errores no es nula y además la media de la variable dependiente no será igual a la media de la línea de regresión. Es decir:

$$\bar{e} \neq 0 \qquad\qquad \overline{y} \neq \overline{y^*}$$

Además, todo lo anterior lleva a que (3.36) tampoco se cumpla. En este caso solo es cierto que:

$$\sum_i y_i^2 = \sum_i y_i^{*2} + \sum_i e_i^2$$

Para comprobar que esto es cierto vamos a desarrollar el último sumando.

$$\sum_i e_i^2 = \sum_i \left(y_i - y_i^*\right)^2 = \sum_i y_i^2 + \sum_i y_i^{*2} - 2\sum_i y_i y_i^* = \sum_i y_i^2 + b^2 \sum_i x_i^2 - 2b\sum_1 y_i x_i =$$

$$= \sum_i y_i^2 + b^2 \sum_i x_i^2 - 2b^2 \sum_1 x_i^2 = \sum_i y_i^2 - b^2 \sum_i x_i^2 = \sum_i y_i^2 - \sum_i (bx_i)^2 =$$

$$\sum_i \quad \sum_i$$

Para llegar a este resultado se ha hecho uso de que $y^*_i = bx_i$ *y de que* $b = \dfrac{\sum_i y_i x_i}{\sum_i x_i^2}$.

Además, como (3.36) no se cumple, ahora no es aplicable la definición del coeficiente de determinación dada en (3.41) ni su relación con el coeficiente de correlación lineal que se mostró en (3.43).

3.4.4. Cambios de origen y de escala.

En este apartado se analizará como afectan los cambios de origen y de escala sobre las variables dependiente e independiente a los distintos coeficientes que se han definido y estudiado en este capítulo.

1º Efectos de cambios de origen y de escala en los parámetros de la línea de regresión, es decir, en la ordenada en el origen y en la pendiente.

a) Si se realiza un cambio de origen en la variable X, la ordenada el origen se ve afectada por ese cambio, pero no la pendiente.

Si se define $X' = X + k$, entonces se tiene que: $X = X' - k$, por lo que el modelo inicial $y_i^* = a + bx_i$ se transforma en: $y_i^* = a + bx_i = a + b(x' - k) = a - bk + bx' = a' - bx'$

b) Si se realiza un cambio de escala en la variable *X*, cambia la pendiente pero no la ordenada el origen.

Si se define *X' = kX*, entonces se tiene que: *X = X'/k* , por lo que el modelo inicial $y_i^* = a + bx_i$ se transforma en $y_i^* = a + bx_i = a + b(x'/k) = a + \left(\frac{b}{k}\right)x' = a - b'x'$

c) Si se realiza un cambio de ori gen en la variable *Y*, la ordenada el origen se ve afectada por ese cambio, pero no la pendiente.

Si se define *Y' = Y + k*, entonces el modelo inicial $y_i = y_i^* + e_i = a + bx_i + e_i$ se transforma en: $y_i' = y_i + k = a + bx_i + e_i + k = a' + bx_i + e_i$

d) Si se realiza un cambio de escala en la variable *Y*, cambia la pendiente y la ordenada el origen.

Si se define *Y' = kY*, entonces el modelo inicial $y_i = y_i^* + e_i = a + bx_i + e_i$ se transforma en: $y_i' = ky_i^* + ke_i = ka + kbx_i + ke_i = a' + b'x_i + e_i'$

Los resultados obtenidos anteriormente se podrían haber alcanzado también si se hubiera trabajado directamente con las expresiones de cálculo de la pendiente (3.20) y la ordenada en el origen (3.21).

2º Efectos de cambios de origen y de escala en las variables *X* e *Y* sobre los coeficientes de correlación y de determinación.

Estos coeficientes son invariantes frente a cambios de origen y de escala. Esta afirmación no es necesario comprobarla, pues para el coeficiente de correlación ya se hizo en el apartado 5.4, y para el de determinación resulta innecesario, dada la relación existente entre ambos.

Ejemplo 12. *Obtenga el valor de los coeficientes del ejemplo 6 si el consumo y la renta se expresaran en pesetas.*

En este caso se trata de un cambio de escala realizado simultáneamente sobre X e Y. Sabemos que tanto X como Y vienen expresadas en €. Para expresarlas en pesetas tendríamos que realizar los siguientes cambios: X' = kX. A su vez Y' = kY. Tanto en un caso como en otro sabemos que k = 166,386. Todo ello nos lleva a que de la relación inicial en euros dada por:

$$y_i = a + bx_i + e_i$$

se pase a la siguiente:

$$\frac{y_i'}{k} = a + b\frac{x_i'}{k} + e_i$$

de forma tal que la relación expresada en pesetas quedaría como:

$$y_i' = ka + bx_i' + ke_i$$

es decir, la propensión marginal al consumo no cambia, mientras que el consumo autónomo si que cambia.

3.4.5.- Predicción.

Adelantarse al futuro es y ha sido siempre un continuo deseo para el ser humano, pero a la vez inalcanzable. Muchas han sido las técnicas puestas al servicio de esa empresa. Pero ninguna de ellas consigue unos resultados aceptables sin un mínimo de "arte", porque la predicción es, en buena medida, un arte, incluso aunque las técnicas que se utilicen para la realización de esas predicciones sean, en términos estadísticos, robustas y potentes.

Como ya se ha indicado, el análisis de la regresión es una de esas técnicas que sirve para describir el comportamiento conjunto de dos variables y también para realizar predicciones. Pero lo que se predice con una línea de regresión ajustada son los valores medios de la variable dependiente, pues la componente errática de la dependencia

estadística no es predecible. Es, como se ha señalado en otro lugar, un símbolo de la ignorancia residual del estadístico.

Aún teniendo en cuenta estas limitaciones, los resultados de estos ejercicios de prospectiva deben tomarse con mucha cautela y la validez de los mismos estará sujeta a una serie de requisitos previos que han de tenerse en cuenta.

Las predicciones que se realizan a partir de la línea de regresión están condicionadas a los valores de la variable independiente. Esto hace que las mismas puedan agruparse en dos categorías distintas. Por un lado están las interpolaciones y por otro las extrapolaciones. Las primeras se corresponden con valores de la variable independiente pertenecientes al recorrido de valores observados de esta variable. En cambio, las segundas son las que se realizan cuando a X se le asignan valores fuera de su recorrido observado.

Para que la validez del primer tipo de predicciones sea aceptable será requisito necesario que el ajuste realizado sea bueno. Esto se puede medir en términos del coeficiente de determinación. Solo se debería depositar confianza en esos resultados cuando el valor de R^2 sea suficientemente alto.

Por lo que se refiere al segundo tipo de predicciones, la validez de las mismas estará condicionada a que el ajuste sea bueno (R^2 alto) y a que la relación cuantificada entre las variables X e Y en el recorrido de los valores observados de las mismas se mantenga incluso para aquellos que se alejen. Esta segunda condición es fundamental, pues la bondad del ajuste no es en absoluto suficiente para realizar pronósticos cuando nos alejamos de los valores observados.

Ejemplo 13. *A lo largo de 26 días se han observado los precios de venta (x) y las cantidades demandas (Y) de cierto producto. Los resultados de esas observaciones son los siguientes:*

Precio	*Cantidad*	*Precio*	*Cantidad*
3,5	*20*	*6,3*	*17*
3,8	*19*	*6,5*	*16,5*
4	*18*	*7,3*	*16*
4,9	*19*	*7,2*	*16*
4,5	*19*	*7*	*16,8*
5	*18*	*7,1*	*16,5*
5,2	*19*	*8*	*16*
5	*18*	*8,6*	*16*
5,3	*18*	*8,7*	*15,3*
5,7	*18*	*8*	*15,2*
6	*17*	*9*	*15,5*
6,2	*17*	*9,1*	*15*
6,7	*16*	*9*	*14,5*

Con esta información:

a) *Ajustar una función de demanda lineal.*
b) *Interpretar el significado de los coeficientes.*
c) *Obtener la elasticidad demanda-precio media.*
d) *Predecir la cantidades demandadas cuando los precios son 5,5 y 20.*

Ejemplo 14. *En unos grandes almacenes se ha observado que las compras de los clientes (expresadas en euros) de productos de marca blanca dependen de forma lineal del total de compras de los mismos. Con la información de 200 compras realizadas en un día se obtuvieron los siguientes resultados:*

b = 0,1023; r = 0,9886; Media de X = 181,5 €; Media de Y = 18,2 €; V(Y) = 95,76 €^{2}

1) *Obtener la recta de regresión.*
2) *Dar una interpretación del significado de los coeficientes obtenidos.*
3) *Estudiar la bondad del ajuste y cuantificar la variancia de la variable de dependiente no explicada por el modelo.*
4) *Obtener la elasticidad media.*

5) *Si admitimos que la relación ajustada se mantiene para cualquier volumen de compras, determine las ventas de productos de marca blanca que cabría esperar en un cliente que realiza una compra por valor de 400€.*

Ejemplo 15. *La dirección de un restaurante ha observado que el número de botellas (Y) de vino gran reserva que se sirven en una noche depende linealmente del gasto medio por persona expresado en euros (X). Para ver hasta que punto es cierto que hay ese tipo de relación se anotaron a lo largo de diez semanas el número de botellas diarias vendidas en las cenas así como el coste medio por persona de esas cenas. Los resultados de esos 70 días, de forma resumida, son los siguientes:*

Media de X = 50 ; Media de Y = 10; V(Y) = 20,25; V(X) = 325; COV(X,Y) = 80

Con estos datos:

a) *Hallar la recta de regresión.*
b) *Analizar la bondad del ajuste.*
c) *¿Cuál sería la demanda esperada de botellas si el coste medio por persona de la cena en una noche se eleva a 70 euros?*

Ejemplo 16. *En unos grandes almacenes se ha realizado una campaña publicitaria orientada a conseguir una mayor demanda por parte de sus clientes de productos de marcas blancas. Finalizada esta campaña, se observó, durante veinte días consecutivos, el volumen de comparas, medido en euros, que realizaron sus clientes de ese tipo de productos. Los resultados de esta observación son los que se dan en la tabla adjunta. Con estos datos realice un análisis del el impacto de la campaña publicitaria sobre la venta de productos de marca blanca, seleccionando para ello el modelo que considere más adecuado, estudiando la bondad del ajuste del mismo y valorando la capacidad predictiva de ese modelo.*

Días	Ventas
1	12000
2	11000
3	10500
4	9800
5	8000
6	8100
7	8000
8	7500
9	7200
10	6800
11	7000
12	7100
13	6800
14	6500
15	6200
16	6500
17	6100
18	6000
19	6200
20	6100

CAPITULO 4.- SERIES TEMPORALES

4.1 Introducción.

Hasta ahora todas las variables que se han estudiado tenían en común que, por lo general, nunca han estado fechadas, es decir no estaban vinculadas al tiempo en forma alguna y menos explícitamente. Se trataba de datos de corte transversal o atemporales. Sin embargo es muy frecuente, especialmente en el ámbito económico y en general en las ciencias sociales, que las observaciones de los caracteres de una población se realicen ligadas al tiempo o fechadas en instantes determinados del tiempo. Así, por ejemplo, una de los caracteres de una empresa susceptible de ser observado puede ser su volumen de ventas y podemos estar interesados en estudiar el comportamiento y evolución temporal de esa característica de la empresa. En este caso esa observación se realizará de forma repetida durante una serie de momentos del tiempo. Esa observación repetida en el tiempo da lugar a una serie temporal. En este sentido diremos que *una serie temporal, cronológica, histórica o de tiempo es una sucesión de observaciones cuantitativas de un fenómeno ordenadas en el tiempo*.

El análisis de series temporales, desde el punto de vista de su comportamiento, tanto pasado como futuro, requiere el uso de nuevas técnicas, pues las presentadas hasta el momento, aunque le son aplicables, no cubren las necesidades que surgen en el tratamiento de este tipo de datos.

Desde el momento que los valores de una serie temporal van ligados a instantes del tiempo, entonces, podemos decir que el análisis de una serie implica el manejo conjunto de dos variables, siendo una de ellas nuestra serie temporal y la otra los intervalos o instantes del tiempo sobre los cuales se han realizado las observaciones. Hay que señalar que esa observación sincronizada de la variable en el tiempo implica que los valores de la misma han de estar perfectamente ordenados, de igual modo que los intervalos del tiempo lo están.

Esas observaciones de una variable cuantitativa pueden estar referidas, como ya se ha señalado, a un instante del tiempo o a un intervalo del mismo, dando lugar a dos tipos de magnitudes. En el primer caso hablaremos de magnitudes stocks o niveles. En el segundo se habla de flujos. Una variable stock es el número de empleados de una empresa en un instante del tiempo, a final de cada mes, mientras que un flujo serían las ventas de esa empresa a lo largo de ese mes, por ejemplo. La diferencia

entre una y otra es que la primera no es sumable para los distintos instantes de un intervalo, pues se incurriría en duplicaciones de los valores de esa magnitud. En cambio, el segundo tipo de magnitud si es sumable o acumulable a lo largo de un periodo o intervalo de tiempo. Para este segundo tipo, los intervalos para los que se acumulan deben ser siempre de igual amplitud. Es decir, si se dan datos de ventas de una empresa, estos deberán ser siempre mensuales, trimestrales, etc, pero lo que no nunca deberá hacerse es intentar trabajar con una serie que mezcle datos semanales con mensuales o referidos a cualquier otro periodo temporal.

Este requisito lleva implícita la idea de homogeneidad. Para que el análisis de una serie temporal nos conduzca a conclusiones acertadas no basta con utilizar las técnicas apropiadas, sino que será imprescindible que esos datos sean comparables y no lo serán nunca si no son homogéneos. Si cada año cambia la metodología de observación, se cambian las definiciones, se modifica la población de referencia, etc, el resultado será una serie temporal compuesta por un conjunto de valores no comparables porque son muy heterogéneos. Esta falta de homogeneidad se pierde, de una forma natural, con el transcurso del tiempo, de manera que cuando las series son muy largas no hay garantía de que los datos iniciales y finales sean comparables. Pero esta necesidad de que las series no sean muy largas, para que sus datos no pierdan la deseable homogeneidad, entra en contradicción con el objetivo más elemental de la Estadística que es el de detectar regularidades en los fenómenos de masas.

Lo que se pretende con una serie es describir y predecir el comportamiento de un fenómeno que cambia en el tiempo. Esas variaciones que experimenta una serie temporal pueden ser de naturaleza doble. Por un lado las variaciones pueden ser evolutivas o estacionarias. Diremos que las variaciones son evolutivas cuando el valor medio de la serie cambia, no permanece fijo a lo largo del tiempo, mientras que las variaciones estacionarias son aquellas en las su valor medio no cambia, aunque sufra oscilaciones en torno a ese valor medio fijo o constante[1]. Esta clasificación de las variaciones de una serie permite hablar de series evolutivas y estacionarias.

[1] Esta forma de definir una serie estacionaria es solo una aproximación al concepto de la misma. No basta con que el valor medio no cambie a lo largo del tiempo. También es preciso que la variabilidad de la misma sea constante.

4.2 Componentes de una serie temporal.

La forma más sencilla de iniciar el análisis de una serie temporal, al igual que se ha venido haciendo con datos de corte transversal, es mediante su representación gráfica. Para ello se hará uso de un sistema cartesiano en el que los valores o periodos de tiempo se llevan al eje de abscisas y los valores de la serie, y_t, se llevan al eje de ordenadas. El resultado es un diagrama de dispersión, con la particularidad de que el eje de abscisas se reserva siempre a la misma variable: el tiempo.

Mediante este tipo de representación se pueden detectar las características más sobresalientes de una serie, tales como el movimiento a largo plazo, la amplitud de las oscilaciones, la posible existencia de ciclos, los puntos de ruptura, la presencia de valores atípicos o anómalos, etc. Un ejemplo de este tipo de gráficas es el que aparece en la Figura 1, donde se ha representado la serie que recoge el paro registrado en España para un periodo de cinco años con datos mensuales. Estos datos son los que se dan en la Tabla 1.

Una vez iniciado el proceso de descripción de una serie y superado el primer paso que consiste en su representación gráfica, para poder llegar a conclusiones más definitivas respecto del comportamiento de la serie, es conveniente recurrir a otras técnicas que superen el mero análisis gráfico.

Tabla 1. **Evolución del paro registrado en España (Miles de parados).**

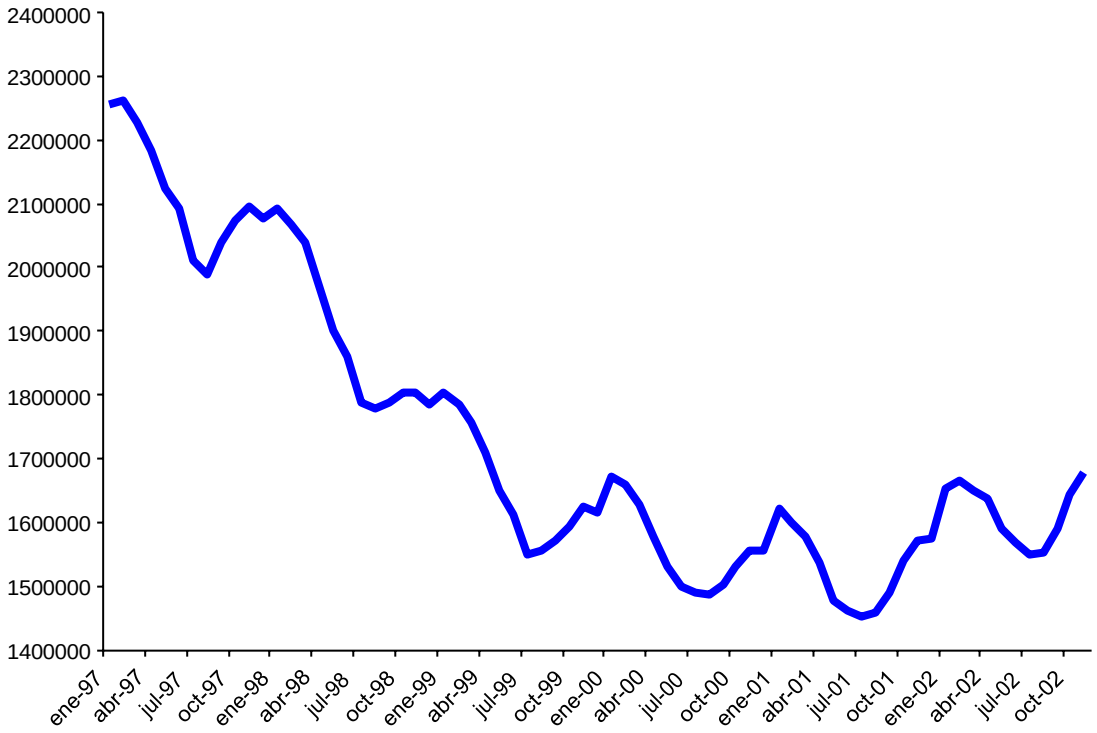
	1997	1998	1999	2000	2001	2002
Enero	2256,5	2091,3	1804,2	1670,6	1620,7	1651,7
Febrero	2262,7	2067,8	1783,9	1659,8	1598,9	1666,0
Marzo	2227,5	2039,1	1757,2	1628,5	1578,5	1649,0
Abril	2181,7	1968,0	1708,0	1578,9	1535,1	1636,3
Mayo	2123,8	1902,2	1649,1	1531,2	1478,1	1589,0
Junio	2091,9	1860,6	1612,5	1500,1	1460,6	1567,4
Julio	2009,2	1786,1	1551,0	1488,8	1451,5	1548,4
Agosto	1989,0	1777,1	1554,5	1487,6	1459,0	1552,0
Septiembre	2040,1	1788,4	1570,0	1501,4	1488,6	1590,3
Octubre	2072,9	1803,7	1591,7	1530,1	1540,0	1641,7
Noviembre	2093,9	1804,5	1623,7	1556,9	1572,8	1678,0
Diciembre	2075,7	1785,7	1613,8	1556,4	1574,8	1688,1

Fuente: Servidor web del INE

El objetivo del análisis de series temporales es doble. Por un lado se busca explicar las variaciones observadas en la serie en el pasado, tratando de determinar si responden a un determinado patrón de comportamiento. Y por otro, si se consigue definir ese patrón o modelo, se intentará predecir el comportamiento futuro de la misma.

Para alcanzar este doble objetivo se utiliza una metodología bastante consolidada, según la cual se admite que la serie temporal es una función del tiempo: $y_t = f(t)$. Bajo este esquema, la serie sería una variable dependiente y el tiempo una independiente o explicativa. Sin embargo, es necesario dejar bien claro que el tiempo, en si, no es una variable explicativa, es simplemente el "soporte" o escenario en el que se realiza o tiene lugar la serie temporal. El tiempo no sirve para explicar el comportamiento de la serie. A esta forma de abordar el estudio de una serie temporal se le conoce como enfoque clásico, frente al causal, según el cual, cualquier serie, como variable que es, puede ser explicada por otra u otras series.

Figura 1. Evolución del paro registrado en España



Desde este punto de vista, cualquier serie temporal se supone que es el resultado de cuatro componentes: tendencia, variaciones estacionales, variaciones cíclicas y variaciones residuales o accidentales. Pero esta descomposición de la serie no deja de ser un procedimiento diseñado para que el estudio de la misma resulte más fácil, pues esas componentes no siempre existen. Así cuando se trabaja con datos anuales la serie no puede presentar estacionalidad. A su vez las variaciones cíclicas son una componente ligada especialmente a las variables de tipo económico, pero que en variables de otra naturaleza puede que no esté presente. Estas componentes se definen en la forma siguiente:

1ª Tendencia (T). De forma amplia podemos definir la tendencia como aquella componente que recoge el comportamiento de la serie a largo plazo. Para poder detectarla es necesario que la serie conste de un número de observaciones elevado, a lo largo de muchos años, para que se puede determinar si la serie muestra un movimiento a largo plazo que responda a una determinada ley de crecimiento, decrecimiento o estabilidad. Ese comportamiento tendencial puede responder a distintos perfiles: lineal, exponencial, parabólico, logístico, etc. Para el ejemplo del paro registrado, puede observase que la tendencia de esa serie a lo largo de esos seis años (este periodo de tiempo no es muy largo para hablar de tendencia a largo plazo) es prácticamente una línea recta con pendiente negativa, aunque el ritmo de decrecimiento no solo se reduce al final del periodo sino que a lo largo de 2002 parece tener lugar un incipiente cambio de tendencia. O sea, que se tiene una serie que es, básicamente, decreciente para el periodo considerado. Mediante la tendencia se puede ver si la serie es estacionaria o evolutiva. Al considerar estos movimientos a largo plazo, prescindiremos de las variaciones a corto y medio plazo.

2ª Variaciones estacionales (VE). Son movimientos de la serie que se repiten de forma periódica. La razón de estas variaciones se basa en causas de tipo climatológico (producción, turismo, etc.) o de ordenación del tiempo (los días de la semana condiciona el comportamiento de ciertas series). La periodicidad generalmente es el año, aunque puede ser el mes, la semana o incluso el día. En el ejemplo de la Figura 1 se observa un patrón de estacionalidad bastante bien definido: el paro registrado desciende notablemente en los meses estivales y el resto del año se mantiene en niveles más elevados, salvo en el mes de diciembre que, de forma sistemática es algo más reducido que en los anteriores y posteriores.

3ª Variaciones cíclicas (C) Esta componente tiene un marcado carácter económico, pues suele ser el resultado de la sucesión de las fases expansivas y recesivas de la economía. Son movimientos a plazo medio, periodos superiores al año, que se repiten de forma casi periódica, aunque no son tan regulares como las variaciones estacionales. Esta componente resulta difícil de aislar, pues ocurre, con frecuencia, que se pueden superponer ciclos de distintos periodos o amplitudes. La amplitud es el número de años que dura un ciclo completo. En nuestro ejemplo no se detecta de forma clara la presencia de ciclos, bien sea porque el periodo de tiempo estudiado sea muy corto o porque realmente no hay ciclos, aunque lo más verosímil en este caso sea la primera razón, pues el empleo responde a los ciclos de la economía.

4ª Variaciones accidentales (R). Esta componente no responde a ningún patrón de comportamiento, sino que es el resultado de factores fortuitos o aleatorios que inciden de forma aislada y no permanente en una serie. Estos factores pueden ser de índole muy diversa tales, como inundaciones, huelgas y otras similares.

La interacción de estas cuatro componentes genera la serie temporal. La forma en que se combinen puede ser muy variada, pero tradicionalmente se ha optado por dos modelos distintos. El aditivo y el multiplicativo, aunque en algunas ocasiones se mezclan ambos. Según el modelo que se adopte, la serie temporal será:

$$y_t = T_t + VE_t + C_t + R_t \qquad (4.1)$$

en el caso del modelo aditivo, y

$$y_t = (T_t)(VE_t)(C_t)(R_t) \qquad (4.2)$$

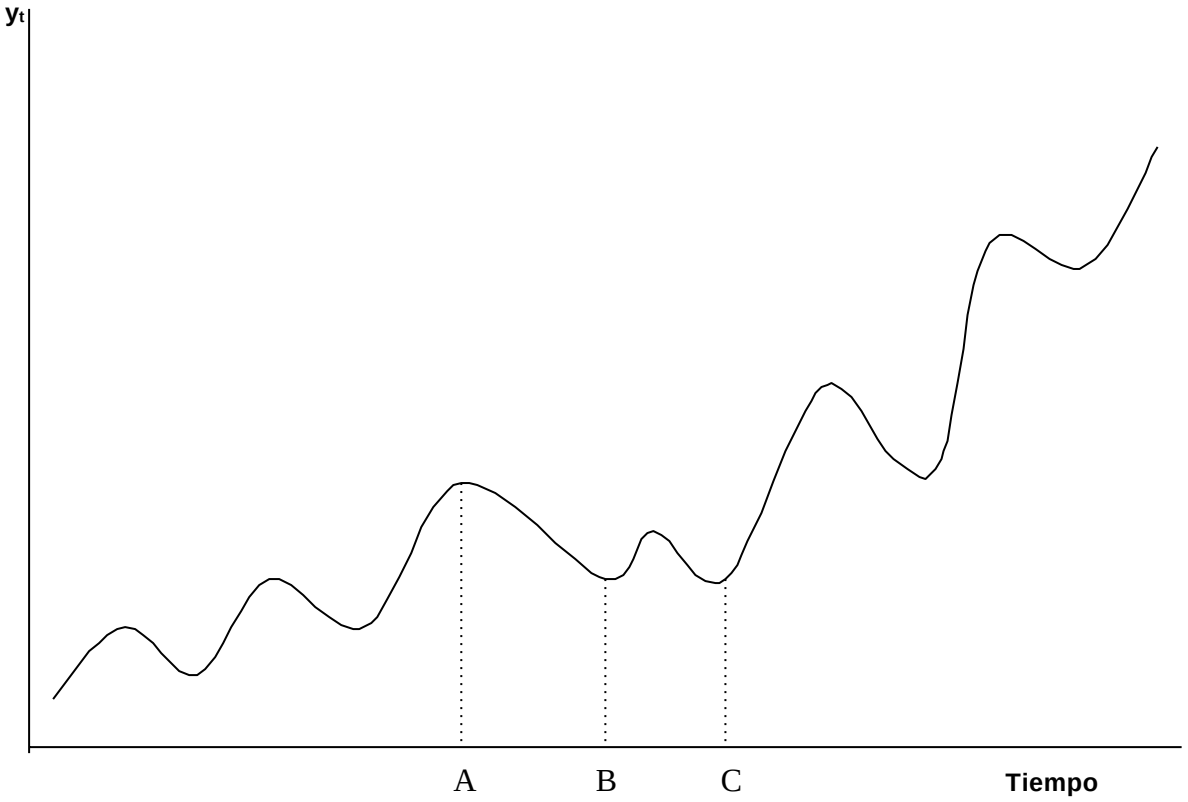
si se admite un esquema multiplicativo.

Frente a este tratamiento clásico de las series temporales, también se puede optar por otro enfoque de tipo causal, donde las variaciones de una serie podrían explicarse mediante las de otro conjunto de series temporales.

4.3 Obtención de la tendencia.

Para aislar esta primera componente de una serie se pueden utilizar distintos métodos alternativos. Pero con independencia del que se utilice, debe quedar bien claro que, tal y como se ha definido la misma, el periodo de información necesario debe ser lo suficientemente largo para evitar identificar como tendencia otros movimientos distintos de la serie.

Figura 2.



Para entender esta idea se puede hacer uso de la Figura 2. En la misma se ha representado una serie ficticia que muestra un perfil creciente, por lo que se puede concluir que la tendencia de la misma, como movimiento a largo plazo, no es decreciente, sino más bien lo contrario. Sin embargo, si nos hubiéramos limitado al periodo de tiempo que va de los puntos A al B, la conclusión seria justo la contraria de la indicada antes. Además, como dentro de ese periodo de tiempo no se ha observado ninguna perturbación de la "tendencia", salvo que se tuviera información extra, no habría motivos para dudar de una tendencia lineal decreciente. Al mismo resultado se habría llegado si el periodo de observación hubiera sido desde A hasta C, solo que

ahora se diría que ha habido un ciclo por medio. Pero, tanto en un caso como en otro, las conclusiones serían poco afortunadas, pues se habría detectado de forma correcta la tendencia pero circunscrita a un periodo de tiempo muy corto, lo que entra en contradicción con la definición misma de tendencia, que se asocia con periodos de tiempo largos.

El problema es que el concepto de largo plazo va íntimamente ligado a la naturaleza de la variable, por lo que la longitud de esos periodos no siempre es comparable. Algo similar a lo descrito en este párrafo le ocurre a la serie recogida en la Figura 1, pues la tendencia que muestra el paro registrado en el periodo considerado es, en realidad, la manifestación de la fase expansiva del ciclo económico experimentado por la economía española a lo largo de la segunda mitad de los años noventa.

A continuación vamos a comentar algunos de los métodos más habituales en la determinación de la tendencia.

4.3.1 Análisis gráfico.

Se trata del método más simple para la obtención de la tendencia. Pero en su sencillez está su debilidad, pues al no hacer uso de ningún procedimiento analítico que garantice, tanto la objetividad del resultado como la posibilidad de que dos analistas distintos lleguen al mismo resultado, este procedimiento es impreciso y no garantiza fiabilidad alguna. Todo depende del conocimiento que de esa serie tenga el investigador que la esté trabajando. A grandes rasgos, este método pasa por la representación gráfica de la serie para posteriormente, bien a mano alzada o por cualquier otro procedimiento de características similares, obtener la tendencia. Lo que si es cierto es que, mediante esta vía sencilla, se podrá tener una idea preliminar de cual es la tendencia. Además la representación grafica de la serie es un paso previo aconsejable en el análisis de la misma.

4.3.2 Medias móviles.

Con este método lo que se hace es “suavizar” la serie promediando los valores de la misma para periodos de tiempo fijos pero que se desplazan a lo largo de todo el horizonte de la serie. El resultado de este proceso mecánico es la eliminación de los movimientos a corto y medio plazo así como las irregularidades debidas a factores no controlables ni predecibles. Es decir, a la serie se el quitan tres de sus componentes y se le deja solo la cuarta, la tendencia. La idea que subyace detrás de este método es

que la media de cualquier conjunto de valores sirve para eliminar la dispersión o variabilidad de la serie motivada por factores coyunturales o esporádicos.

Este método de suavizado consiste, como se ha indicado, en promediar la serie. Estos promedios serán las medias aritméticas de un conjunto *k* de valores consecutivos, con el requisito de que *k* sea inferior al total de observaciones. El procedimiento específico sería el siguiente. Supóngase que *k* es un entero impar. Entonces las sucesivas medias se obtendrían de forma siguiente:

$$y_t^* = \frac{\sum_{i=-\frac{k-1}{2}}^{\frac{k-1}{2}} y_{t+i}}{k} = \frac{y_{t-\frac{k-1}{2}} + y_{t-\frac{k-1}{2}+1} + y_{t-\frac{k-1}{2}+2} + \ldots\ldots + y_t + \ldots\ldots + y_{t+\frac{k-1}{2}-2} + y_{t+\frac{k-1}{2}-1} + y_{t+\frac{k-1}{2}}}{k} \qquad (4.3)$$

A la media y_t^* se le llama centrada porque al ser impar el número de sumandos con el que se ha obtenido, la media resultante se le hace corresponder con la observación del momento *t*, que es el valor central de la suma. Según esta expresión de cálculo, la primera media que se puede calcular es la correspondiente al grupo de valores cuya observación central se corresponde al instante $t = \frac{k-1}{2}$, siendo en este caso la primera observación y_0, Una vez obtenida esa media, la siguiente se calcula para los *k* valores que tienen por observación central la del periodo *t+1*, y así sucesivamente. Esta forma de obtener medias eliminando la primera observación del grupo y añadiendo la siguiente es lo que el da el adjetivo de "móvil" a las mismas.

Par fijar un poco las ideas supongamos que *k=5*. En tal caso, las sucesivas medias móviles vendrían dadas por:

$$y_2^* = \frac{y_0 + y_1 + y_2 + y_3 + y_4}{5} \qquad y_3^* = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5} \qquad y_4^* = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$$

Ahora bien, si *k* fuera par, entonces la media de esos *k* valores no se correspondería con ninguno de los observados de la serie original, sino con el punto medio de los dos

centrales. Pero ese instante no es observable ($t = \frac{k-1}{2}$ no sería un entero), por lo que las medias calculadas de esta forma habría que promediarlas de dos en dos y de forma sucesiva para que el resultado si fuera una serie de valores (medias) centrados, es decir, que se correspondan con valores para periodos o instantes de tiempo observados[2].

Esta serie no centrada se obtendría mediante la expresión:

$$y^*_{t-0,5} = \frac{\sum_{i=-\frac{k}{2}}^{\frac{k}{2}-1} y_{t+i}}{k} \qquad \left(t = \frac{k}{2}, \frac{k}{2}+1, \ldots\ldots, n-\frac{k}{2} \right) \qquad (4.4)$$

Otra cuestión importante a la hora de calcular las medias móviles es determinar cuantas observaciones deben tomarse en cada caso. Si *k* es muy grande entonces el proceso de suavizado puede llegar a ser tan fuerte que se pierda más información de la deseada. Piense en la situación extrema de que *k* fuera igual al total de observaciones. En ese caso solo habría una media, por lo que el suavizamiento de la serie sería máximo, tanto que no habría ni tendencia ni componente alguna. Por esta razón *k* no debe ser demasiado grande, pues se podría incurrir en un suavizado excesivo. En general, cuanto mayor es *k*, menor será el número de términos de la serie suavizada resultante (se pierden observaciones al principio y al final de la serie). En conclusión, si se toma un grupo de observaciones muy alto se incurre en el peligro de perder información por dos vías: a) la serie se suaviza más de lo necesario, ocultando ciertos movimientos tedenciales; b) el número de términos de la nueva serie se reduce considerablemente, y perder datos nunca es bueno.

[2] Este proceso de tomar medias móviles en dos ocasiones par centrar la serie puede reducirse a una media móvil ponderada, donde el número de sumando es k+1 y todos los valores se ponderan doble, salvo el primero y el último cuya ponderación es la unidad, y esa suma se divide entre 2k. Así si k = 4, las dos primeras medias descentradas serían: $\frac{y_0+y_1+y_2+y_3}{4}$ y $\frac{y_1+y_2+y_3+y_4}{4}$. A partir de estas medias, el valor centrado sería: $\frac{\frac{y_0+y_1+y_2+y_3}{4}+\frac{y_1+y_2+y_3+y_4}{4}}{2} = \frac{y_0+2y_1+2y_2+2y_3+y_4}{8}$.

Por el contrario, si k es muy pequeño entonces no se conseguirán eliminar todas las perturbaciones ajenas a la tendencia. De forma similar a como se razonó antes, si *k=1*, entonces la serie original y la suavizada coinciden, con lo cual no se ha conseguido nada.

En algunos casos, ese valor de *k* es fácil de determinar. Así ocurre cuando la serie muestra un patrón de comportamiento que se repite de manera sistemática cada *k* periodos de tiempo. Tal sería el caso de la estacionalidad. Si se trabaja con datos mensuales y la serie está sometida a un esquema de estacionalidad que se repite todos los años, entonces la forma de suavizar esa serie y eliminar la componente estacional sería tomar una media móvil de doce meses (*k = 12*). A la serie resultante se le habrían eliminado dos componentes: la estacionalidad y las variaciones accidentales. Pero al ser *k* par, la serie resultante no estaría centrada, por lo que habría que volver a tomar medias móviles con *k = 2*.

Una vez que de la serie original se han eliminado esas dos componentes cabría preguntarse cómo proceder con las variaciones cíclicas. En este caso la elección de *k* es más difícil, pues los ciclos no son movimientos de la serie que se repitan con una periodicidad fija, como ocurre con la estacionalidad. En estos casos, si esa periodicidad no puede determinarse de forma clara y sin que perjudique notoramiente a los resultados, la mejor forma de proceder es trabajar con lo que ha dado en denominarse componente ciclo-tendencia.

En el esquema presentado hasta el momento, para la obtención de la tendencia mediante medias móviles, se ha trabajado con el supuesto de que los datos tenían una periodicidad inferior al año (semestres, cuatrimestres, trimestres, meses, etc.) lo que implicaba aceptar la posibilidad de que esa serie presentara estacionalidad. Ahora bien, si los datos fueran anuales entonces la estacionalidad quedaría descartada, pues las únicas componentes de la serie serían la tendencia, los ciclos y las variaciones accidentales. Esta nueva situación nos lleva a que sea poco verosímil que la serie presente un esquema repetitivo a lo largo del tiempo tan estable como presentaban las variaciones estacionales. Ante estas circunstancias se hace difícil saber cuál debiera ser el número adecuado de observaciones que debieran tomarse para calcular las medias móviles. La forma de salir de esta situación incómoda es obtener medias móviles de tres o cinco datos (número impar y pequeño) para de esa forma eliminar la componente accidental. Una vez que se ha procedido de esta forma, la serie suavizada resultante contiene una mezcla de ciclo-tendencia. Si la componente cíclica

fuera regular con periodos definidos y fijos, entonces la tendencia se obtendría aplicando una media móvil con un *k* igual a la longitud del ciclo. Pero es poco probable que los ciclos tengan ese comportamiento tan sistemático, por lo que quizás la mejor solución sea, como se indicó en el párrafo anterior, no manipular más los datos y trabajar con esa mixtura de componentes ciclo-tendencia.

Este método de obtención de la tendencia presenta, frente a su sencillez, algunos inconvenientes que deben ser señalados. Al igual que en el método gráfico, también aquí se introduce un cierto grado de subjetividad, pues la elección del número de observaciones a promediar queda a la elección del analista y, salvo que sea muy claro cual debe ser ese número (caso de la estacionalidad), esa decisión no siempre es la acertada, por lo que los valores de la componente tendencia variaran según quién los calcule. Por otro lado, esta forma de obtener la tendencia no permite alcanzar el objetivo de la predicción en el análisis de las series temporales, pues la tendencia obtenida mediante medias móviles no permite que se proyecte hacia el futuro.

Ejemplo 1. *Obténgase la tendencia de la serie de la Tabla 1 mediante medias móviles.*

En este caso, dado que los datos son mensuales y la serie muestra una clara componente estacional que se repite todos los años, el periodo de la media móvil debe ser de doce datos (doce meses). Pero al ser par el valor de k se deben tomar medias móviles en dos ocasiones. Primero con k=12 y después con k = 2, para de esta forma obtener una serie centrada, que será la tendencia, pues, como puede observarse, la serie original, para el conjunto de años considerado, no muestra componente cíclica clara. Los resultados de estas operaciones son los que aparecen en las Tablas 2 y 3. Adicionalmente, en la Figura 2, se ha vuelto a representar la serie original y la suavizada que recoge la tendencia. Esta última tiene menos observaciones que la primera. En este caso se ha perdido doce datos, seis al inicio y seis al final.

La forma en la que se han obtenido esos datos es la siguiente:

$$y^{*}_{97JUL} = \frac{\frac{y_{96\,ENE} + y_{96FEB} + \ldots\ldots + y_{97\,NOV} + y_{97DIC}}{12} + \frac{y_{96FEB} + y_{96MAR} + \ldots\ldots + y_{97DIC} + y_{97ENE}}{12}}{2} =$$

$$= \frac{\frac{2256,5 + 2262,7 + ... + 2093,9 + 2075,7}{12} + \frac{2262,7 + 2227,5 + ... + 2075,7 + 2091,3}{12}}{2} =$$

$$= \frac{2118,7 + 2105,0}{2} = 2111,9$$

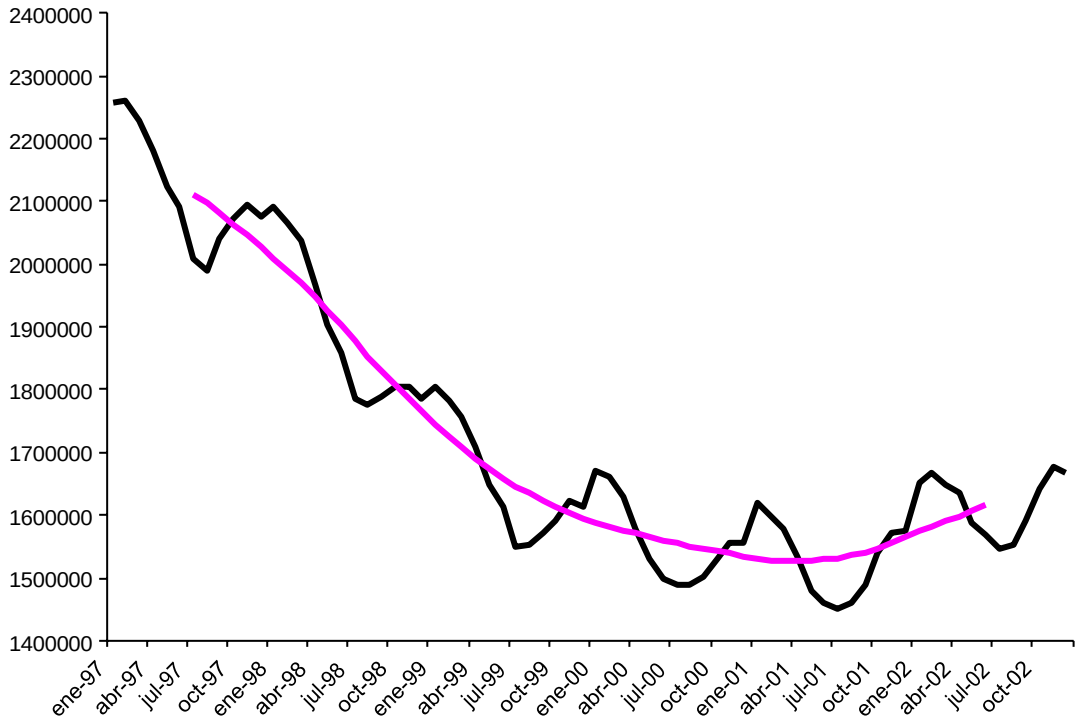
Tabla 2. **Medias móviles de doce meses no centradas de la serie del paro registrado en España**.

	1997	1998	1999	2000	2001	2002
Enero	,,	2017,5	1755,0	1589,5	1532,8	1570,5
Febrero	,,	1998,9	1735,4	1584,3	1529,7	1578,6
Marzo	,,	1981,2	1716,9	1578,7	1527,3	1586,3
Abril	,,	1960,3	1698,7	1573,0	1526,2	1594,8
Mayo	,,	1937,8	1681,0	1567,9	1527,0	1603,3
Junio	,,	1913,7	1665,9	1562,3	1528,3	1612,1
Julio	2118,7	1889,5	1651,6	1557,5	1529,9	1619,8
Agosto	2105,0	1865,6	1640,5	1553,4	1532,5	,,
Septiembre	2088,7	1842,0	1630,1	1548,3	1538,1	,,
Octubre	2073,0	1818,5	1619,4	1544,1	1543,9	,,
Noviembre	2055,2	1796,8	1608,7	1540,5	1552,4	,,
Diciembre	2036,8	1775,7	1598,8	1536,1	1561,6	,,

Tabla 3. **Tendencia del paro registrado en España obtenida por medias móviles. (Medias móviles centradas)**

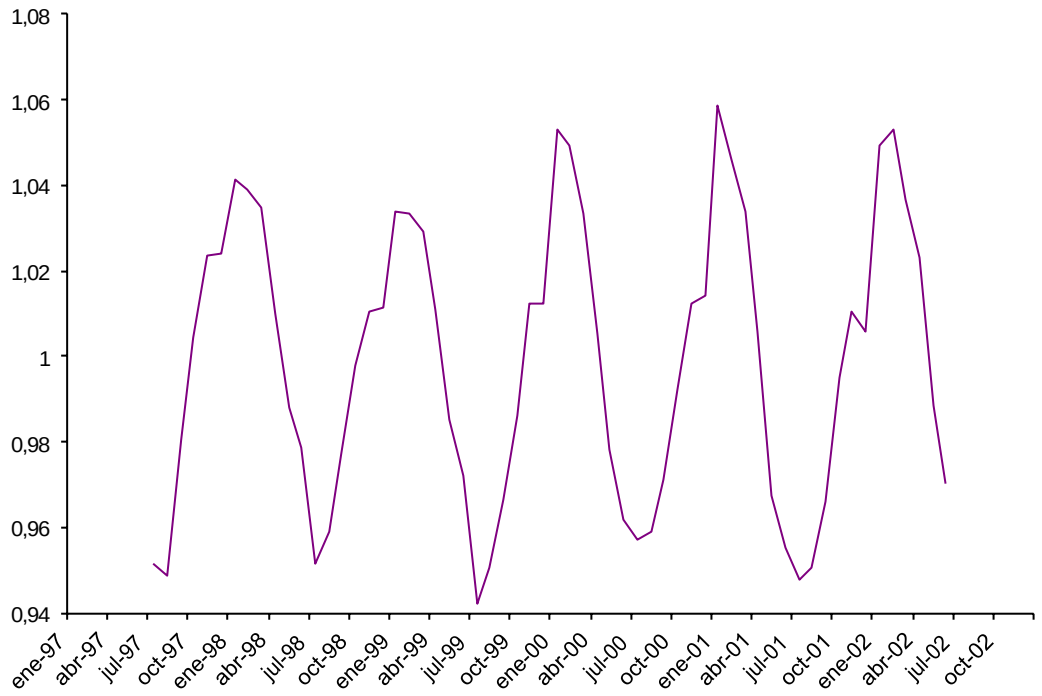
	1997	1998	1999	2000	2001	2002
Enero	,,	2008,2	1745,2	1586,9	1531,2	1574,6
Febrero	,,	1990,1	1726,2	1581,5	1528,5	1582,5
Marzo	,,	1970,7	1707,8	1575,9	1526,7	1590,6
Abril	,,	1949,0	1689,9	1570,4	1526,6	1599,1
Mayo	,,	1925,8	1673,5	1565,1	1527,7	1607,7
Junio	,,	1901,6	1658,8	1559,9	1529,1	1615,9
Julio	2111,9	1877,6	1646,1	1555,5	1531,2	,,
Agosto	2096,8	1853,8	1635,3	1550,8	1535,3	,,
Septiembre	2080,9	1830,2	1624,8	1546,2	1541,0	,,
Octubre	2064,1	1807,6	1614,0	1542,3	1548,2	,,
Noviembre	2046,0	1786,3	1603,7	1538,3	1557,0	,,
Diciembre	2027,1	1765,4	1594,1	1534,4	1566,1	,,

Figura 3. Tendencia del paro registrado obtendia por medias móviles.



Una vez que se ha obtenido la componente ciclo-tendencia y si se admite un esquema multiplicativo de generación de la serie, entonces el cociente entre estas dos componentes nos daría la estacionalidad y las variaciones accidentales. Los resultados de este cociente aparecen en la Tabla 7 y su representación gráfica en la Figura 4.

Figura 4. Serie de paro registrado corregida de la componente ciclo-tendencia.



4.3.3 Método analítico.

Con este procedimiento de obtención de la tendencia, lo que se pretende es seleccionar una función matemática que modelice de forma adecuada el comportamiento a largo plazo de la serie temporal objeto de estudio. Se trata, por tanto, de ajustar los datos observados a esa función, donde la variable a explicar o dependiente es la propia serie temporal y la independiente o explicativa es, ahora, el tiempo. Ni que decir tiene que el tiempo, como ya se ha indicado, no explica nada, sino que es un mero soporte en el que se mueve la serie. El procedimiento de ajuste que puede utilizarse no es único, aunque el más utilizado es el de los mínimos cuadrados.

En consecuencia, se trata de seleccionar aquella función del tiempo que minimice la suma de los cuadros de los errores. Pero aunque este es el criterio último, como paso previo para seleccionar esa función se puede recurrir a su representación gráfica, la cual nos informará de manera aproximada sobre la función que se debe ajustar. Otra alternativa es hacer uso del posible conocimiento de la naturaleza de esa serie o de lo que la teoría establezca.

Las funciones más utilizadas para modelizar series económicas son las siguientes:

- *Lineal*.

  Se trata de un modelo en el que la serie temporal se hace depender linealmente del tiempo y que responde a aquellas magnitudes que presentan unas variaciones constantes en periodos sucesivos. La forma general del mismo es:

$$y = y_t^* + e_t = a + bt + e_t \qquad (4.5)$$

  donde: y_t, la serie original, se descompone en la tendencia y_t^* y, (bajo un supuesto aditivo), las otras componentes que aparecen de forma residual y conjunta bajo e_t, En este modelo b es la variación media entre periodos y t el tiempo cronológico.

- *Polinomial*.

  La función Polinomial es en realidad una familia de funciones que se diferencian unas de otras en el grado del polinomio con el que se trabaje. La más común de todas es la parabólica. En este caso las variaciones de la serie no son constantes, ni en términos absolutos ni relativos. Para este modelo la serie temporal se expresa como:

$$y = y_t^* + e_t = a + bt + ct^2 + e_t \qquad (4.6)$$

- *Exponencial*

  Esta función surge cuando la serie cambia a razón de una tasa constante. Para este tipo de series, su tendencia viene dada por:

$$y_t^* = AB^t = y^* e^{bt}_0 = e^{a+bt} \qquad (4.7)$$

  donde b es la tasa de variación instantánea de la serie temporal que es constante, y_0^* es el valor de la tendencia en punto $t = 0$, t es, de nuevo, el tiempo cronológico y e es la base de los logaritmos naturales.

El ajuste por mínimos cuadrados de cualquiera de estos modelos no plantea ningún problema especial, pues los mismos son lineales o fácilmente linealizables y, en estos casos, la aplicación del método de ajuste de los mínimos cuadrados es inmediato.

Además de las tres funciones señaladas antes también hay otras que estarían indicadas para aquellos casos en los que la serie presenta un techo en su crecimiento, es decir, que no es admisible un crecimiento indefinido. Ese tipo de variables se pueden modelizar con alguna de las siguientes funciones en forma de *S*:

- *Logarítmica recíproca*:

$$y_t^* = e^{a-\frac{b}{t}} \qquad (4.8)$$

Si en esta función se toman logaritmos se tiene que:

$$\ln y_t^* = a - b\left(\frac{1}{t}\right) \qquad (4.9)$$

que es una función lineal a la que es fácil aplicar mínimos cuadrados. Esta función se caracteriza porque su tasa de crecimiento frente a cambios unitarios en *t* es inversamente proporcional al cuadrado de *t*:

$$\frac{1}{y_t^*}\frac{dy^*}{dt}_t = \frac{b}{t^2} \qquad (4.10)$$

lo que conlleva que la tasa de crecimiento no es constante.

La gráfica de esta función es del tipo representado en la Figura 4.

- *Logística*:

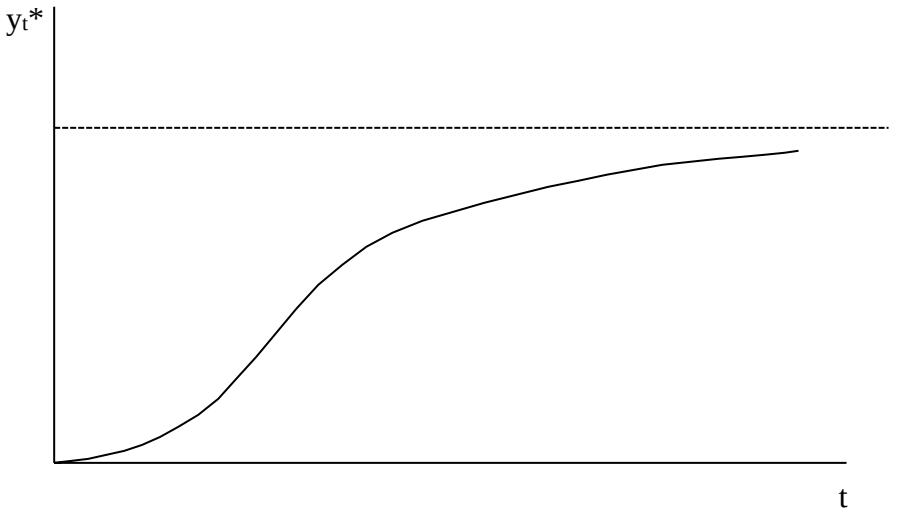
$$y_y^* = \frac{c}{1+ae^{-bt}} \qquad (4.11)$$

El perfil de esta curva es similar al de la anterior y que se ha representado en la Figura 5. Las diferencias estriban en que los puntos de inflexión de ambas son

distintos, pues mientras que en la primera el mismo es $y_t^* = 0{,}135e^a$, en la segunda ese punto se da en $y_t^* = c/2$. Es decir, en la primera el punto de inflexión ocurre antes que en la segunda y, por lo tanto, la caída en la tasa de crecimiento tiene lugar antes.

Al igual que el modelo logarítmico recíproco, la función logística presenta también una tasa de crecimiento variable. El gran problema de este modelo es que no es fácil linealizarlo, por lo que su ajuste es difícil con las técnicas que se manejan en este manual.

Figura 5. Función en forma de S.



Ni que decir tiene que el repertorio de funciones para modelizar la tendencia de una serie no se agota con las expuestas aquí, pero si que puede decirse de ellas que son las más frecuentes.

La obtención de la tendencia mediante el método analítico presenta dos ventajas frente a los otros procedimientos descritos antes. La primera es que se tiene una medida de la bondad del ajuste, lo que permite determinar hasta qué punto la función seleccionada recoge de manera correcta la tendencia de la serie. Ahora la adecuación de la tendencia obtenida no se deja a la mera inspección gráfica, como ocurría con los dos procedimientos anteriores. Por otro lado, al obtener una función explícita, que relaciona los datos observados con el tiempo, se cubre mejor el segundo de los objetivos del análisis de series temporales, que como se señaló era la predicción.

Solo queda una cuestión de orden práctico que no se ha abordado hasta ahora. Es la forma de tratar a las series cuando los datos tienen periodicidad inferior al año y presentan estacionalidad. En estos casos, si el ajuste se realiza directamente con los datos observados (mensuales, trimestrales, etc.), la bondad del mismo será inferior a la que se obtendría si se trabajara con valores anuales (medias anuales), pues en este caso la serie presenta menos fluctuaciones y los errores del ajuste serán más pequeños. Pero ese criterio basado en la bondad del ajuste no es del todo trascendente en este contexto, pues la estacionalidad no tiene porque interferir en la tendencia, por lo que esta será la misma con independencia de que se obtenga con datos anuales o con periodicidad inferior al año.

Ejemplo 2. *Obtener la tendencia correspondiente a la serie de paro registrado dada en la Tabla 1 mediante el método analítico.*

A la vista de la representación gráfica de esta serie en la Figura 1, parece poco razonable ajustar una función lineal, pues el comportamiento del paro registrado a lo largo de esos seis años no puede decirse que fuera lineal. Quizás fura más apropiado ensayar un modelo parabólico, el cual podrá recoger la manifiesta no linealidad de los últimos años.

Inicialmente se trabajará con los datos mensuales. Otra decisión que habrá que tomarse es fijar el origen del tiempo cronológico. Como se sabe, el origen del tiempo se ha fijado siempre de forma arbitraria (el mundo occidental, el mundo árabe, el mundo hebreo, etc, tienen un calendario con un origen diferente). Esa libertad de elección permite que, desde el punto de vista estadístico, ese origen se fije en el punto que convenga más en cada momento, siempre que se indique conde se ha hecho t=0. En consecuencia, para este ejemplo, se situará ese origen en el primer mes. Es decir, se realizará un cambio de origen en la variable independiente, y este cambio ya se sabe como afecta a los distintos coeficientes del análisis de regresión. Así pues, t=0 se corresponde con el mes de enero de 1997. A partir de ese primer valor, la variable t experimentará variaciones unitarias, es decir, t=1 para febrero de 1997, t=2 para marzo de 1997, y así sucesivamente.

Una vez que se ha establecido ese cambio se procederá al ajuste del modelo:

$$y_t = y_t^* + e_t = a + bt + ct^2 + e_t$$

Para este modelo no lineal, el sistema de ecuaciones normales al que se llega, tras la aplicación de mínimos cuadrados, es:

$$\sum_t y_t = Na + b\sum_t t$$

$$\sum_t y_t t = a\sum_t t + b\sum_t t^2$$

$$\sum_t y_t t^2 = a\sum_t t^2 + b\sum_t t^3 + c\sum_t t^4$$

La solución de este sistema de ecuaciones con los datos de la tabla 1 nos lleva a los siguientes resultados:

$$y_t = y_t^* + e_t = 2287203 - 29773{,}4t + 294{,}065t^2 + e_t$$

A su vez la bondad del ajuste medida por el coeficiente de determinación es:

$$R^2 = 0{,}933$$

Todos estos resultados, así como la información previa, se pueden expresar de forma resumida en los siguientes términos:

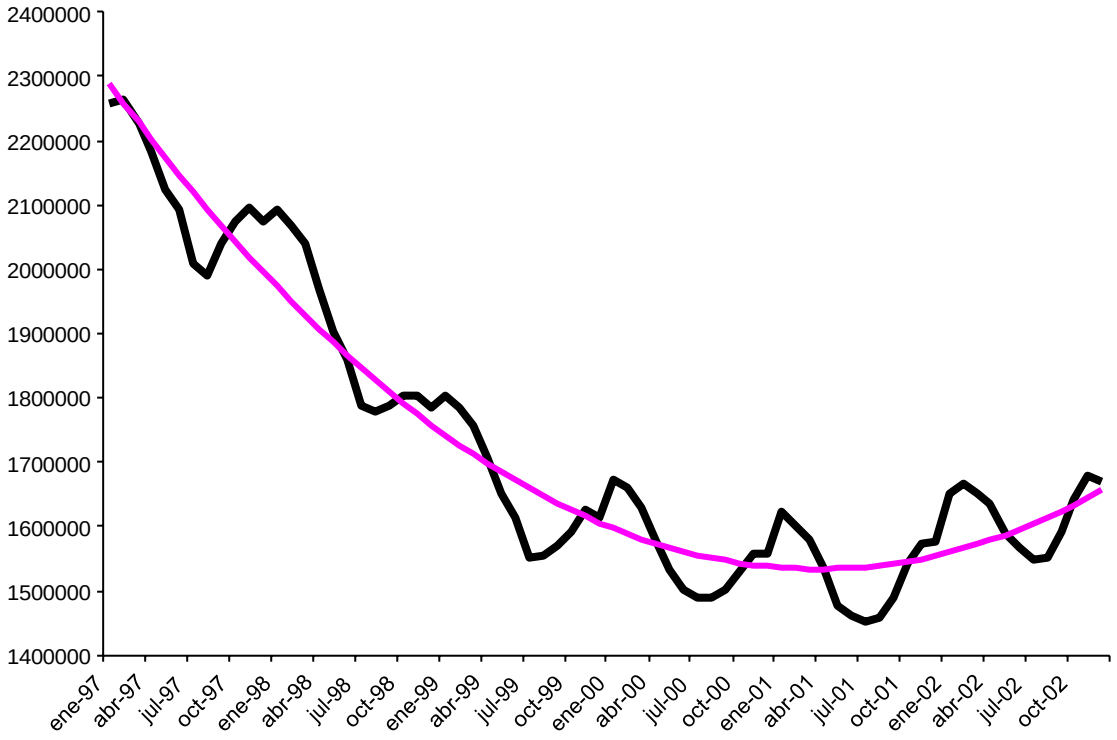
$Y_t^* = 2287203 - 29773{,}4t + 294{,}065t^2$ $\qquad R^2 = 0{,}933$

(t = 0 en enero de 1997; unidad temporal , 1 mes)

En la Figura 6 se han representado la serie original del paro registrado así como la tendencia ajustada[3].

[3] Hay que señalar que aunque los resultados del ajuste son bastante buenos a tenor del valor del coeficiente de determinación, sin embargo, desde un punto de vista formal, es aconsejable trabajar con datos anuales, pues bajo el enfoque analítico de obtención de la tendencia, los residuos del ajuste por mínimos cuadrados recogen tres de las cuatro componentes (todas menos la tendencia), lo que hace que esos residuos no tengan el comportamiento aleatorio deseable en los mismos, pues en ellos queda recogida al menos una componente sistemática, la estacionalidad, siempre que la serie tenga esa componente. Pero en nuestro ejemplo no es del todo aconsejable trabajar con datos anuales pues la información disponible es escasa (solo seis años). Sería preferible eliminar en primer lugar la componente estacional, que es la que podría causar problemas, y, posteriormente, obtener las tendencia de la serie desestacionalizada.

Firgura 6. Paro registrado y tendencia parabólica ajustada por mínimos cuadrados



Ejemplo 3. *En la Tabla 4 se recoge la evolución, desde 1971, del Consumo Final para toda España, expresado en pesetas constantes de ese año. A partir de esos datos obtenga la tendencia de esa serie.*

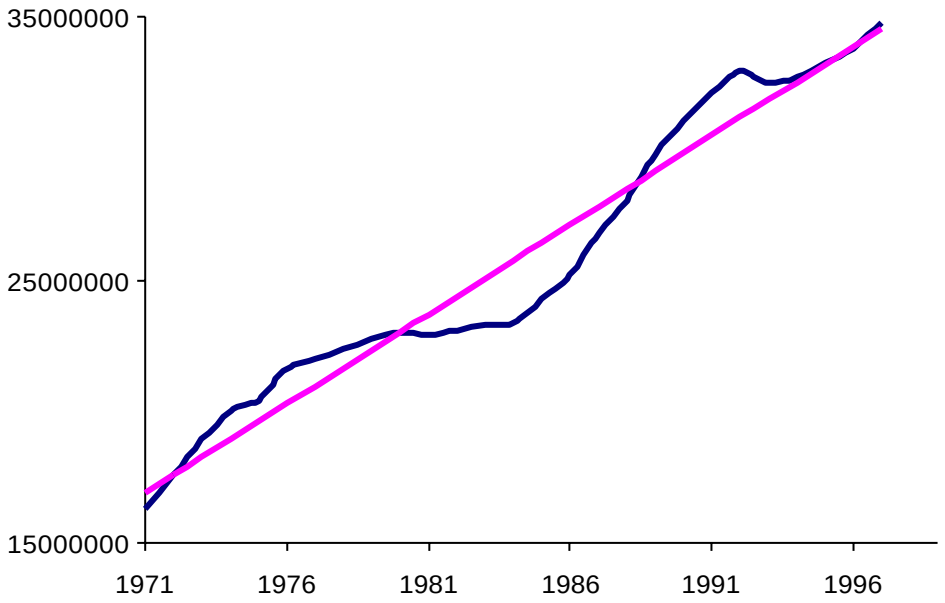
Tabla 4. Consumo Final (10^6 pesetas de 1971)

1971	*16309153*	*1980*	*23034875*	*1989*	*29779333*
1972	*17593135*	*1981*	*22919130*	*1990*	*31036353*
1973	*18932295*	*1982*	*23099009*	*1991*	*32100077*
1974	*20003509*	*1983*	*23308968*	*1992*	*32929885*
1975	*20457068*	*1984*	*23370031*	*1993*	*32514354*
1976	*21640271*	*1985*	*24279204*	*1994*	*32723353*
1977	*22039094*	*1986*	*25177951*	*1995*	*33258278*
1978	*22382034*	*1987*	*26781757*	*1996*	*33843015*
1979	*22771211*	*1988*	*28051856*	*1997*	*34770733*

Fuente: Web INE.

Como paso previo a la obtención de la tendencia es aconsejable visualizar la representación gráfica de esta serie, pues ello facilita la elección del modelo que mejor se ajuste a los datos de la Tabla 4. Por tal razón en la Figura 7 se ha representado la serie del consumo final. A tenor de esta representación puede pensarse en un modelo lineal para la tendencia de esta serie. Pero aunque el modelo propuesto sea acertado en este caso, de lo que no cabe duda es que los valores anuales del consumo no están perfectamente alineados. Presentas desviaciones que se alejan de la línea de tendencia. Esas desviaciones pueden deberse tanto a la componente cíclica como las variaciones residuales. En cambio, por tratarse de una serie con datos anuales, no se verá afectada por la componente estacional.

Figura 7. Evolución del Consumo Final y tendencia lineal ajustada



En el proceso de ajuste de la tendencia a una línea recta vamos a fijar el origen del tiempo en 1971 y trabajar en términos de miles de millones de pesetas. Esto nos lleva a que:

$$\sum t = 351 \qquad \sum y_t = 695105{,}9 \qquad \sum t^2 = 351 \qquad \sum t y_t = 10149170{,}3$$
$$\sum y_t^2 = 18685901587{,}2$$

A partir de esta información se tiene que:

$$y_t^* = 16912{,}97 + 679{,}36t \qquad R^2 = 0{,}956$$

(t = 0 en 1971; unidad temporal , 1 año)

Ejemplo 4. *Según la Encuesta General de Medios, el número de usuarios de internet, desde 1996 a 2001, ha evolucionado como muestra la Tabla 5.*

Tabla 5. Usuarios de internet en España. (Miles)

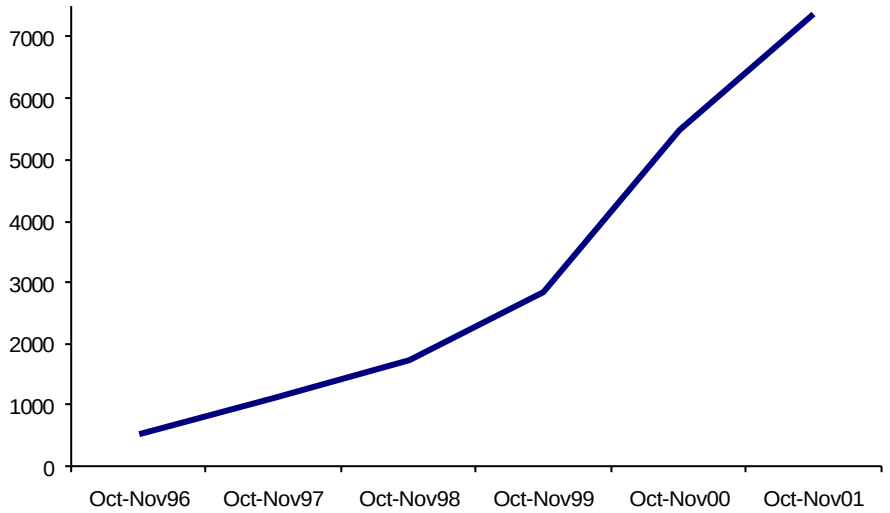
Oct-Nov96	*526*
Oct-Nov97	*1110*
Oct-Nov98	*1733*
Oct-Nov99	*2830*
Oct-Nov00	*5486*
Oct-Nov01	*7388*

Fuente: EGM

Obtener la tenencia de esta serie.

Como en los ejemplos anteriores, lo primero que debe hacerse es obtener la representación gráfica de la serie. La misma es la que aparece en la Figura 8.

Figura 8. Número de usuarios de internet en España (Miles)



A la vista de la Figura 8 no puede afirmarse que el crecimiento medio anual del número de usuarios de internet a lo largo del periodo considerado haya sido constante, por lo que no sería muy acertado modelizar la tendencia mediante una línea recta. Esta figura parece mostrar más un crecimiento exponencial. No obstante se ajustarán los dos modelos y en función de los resultados finales se optará por el mejor para representar la tendencia de esa serie.

Los cálculos necesarios para el ajuste de los dos modelos propuestos son los que aparecen en la tabla siguiente.

t	y_t	lny_t	$t2$	y_t^2	ty_t	$(lny_t)^2$	$tlny_t$
0	526	6,27	0	276676	0	39,25	0
1	1110	7,01	1	1232100	1110	49,17	7,01
2	1733	7,46	4	3003289	3466	55,62	14,92
3	2830	7,95	9	8008900	8490	63,17	23,84
4	5486	8,61	16	30096196	21944	74,13	34,44
5	7388	8,91	25	54582544	36940	79,35	44,54
15	**19073**	**46,20**	**55**	**97199705**	**71950**	**360,69**	**124,75**

A partir de estos datos, los resultados del ajuste lineal son los siguientes:

$$y_t^* = -287,95 + 1386,71t \qquad R^2 = 0,92$$

(t = 0 en 1996; unidad temporal , 1 año)

Aunque el ajuste no es malo, sin embargo nos dice que el crecimiento medio anual ha sido de casi 1400 miles de usuarios, cuando los datos de la Tabla 5 muestran que ese crecimiento medio constante no es muy verosímil. Por esa razón se va aprobar el ajuste con el modelo exponencial especificado en (4.7). La versión linealizada del mismo viene dada por:

$$\ln\ y_t^* = a + bt$$

El ajuste por mínimos cuadrados de este modelo es el siguiente:

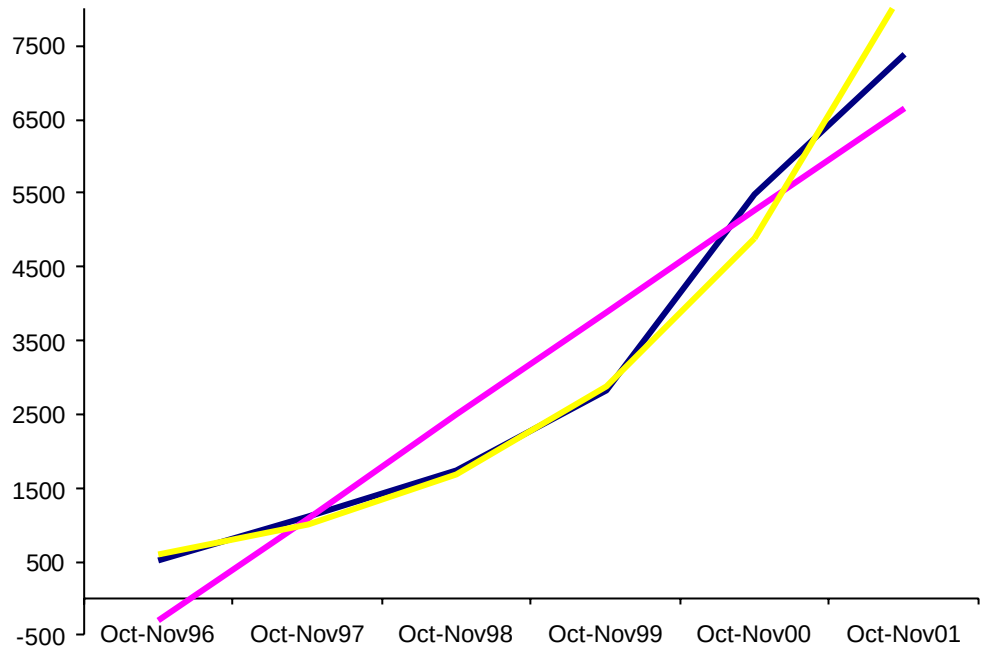
$$\ln y_t^* = 6,38 + 0,528t \qquad R^2 = 0,9896$$

(t = 0 en 1996; unidad temporal , 1 año)

Como puede observarse, estos resultados son mejores que los anteriores, pues la bondad del ajuste es superior y, además, la Figura 9 corrobora esta conclusión.

En cualquier caso hay que señalar que el periodo temporal con el que se ha trabajado es pequeño, por lo que no se puede pensar que esa tendencia observada de 1996 a 2001 se vaya a mantener de manera indefinida. Para este tipo series, la tendencia a largo plazo que mejor se ajusta es la de tipo logístico, donde se combina un crecimiento inicial de tipo exponencial con otro posterior más lento y que tiende a estancarse.

Figura 9. Numero de usuarios de internet en España.
Series original y tendencias lineal y exponencial



4.3.4 Alisado exponencial.

Estos procedimientos de obtención de la tendencia son muy parecidos a los de medias móviles. En ambos casos la tendencia es el resultado de promediar los valores de la serie, bien de forma ponderada o sin ponderar. En el caso de las medias móviles, los promedios son sin ponderar, salvo cuando las medias se obtienen para un número par de valores, pues entonces hay que promediar dos veces, siendo el resultado final una

media ponderada, como ya se ha visto. A estas técnicas se les podría denominar como de alisado proporcional.

Sin embargo, en los procedimientos de alisado exponencial, siempre se utilizan ponderaciones y, además, los valores suavizados o alisados que se obtienen son una combinación lineal de todas las observaciones de la serie, pero con la particularidad de que la ponderación decrece conforme nos alejamos del origen. Estos procedimientos están especialmente diseñados para la predicción.

A continuación se expondrá la idea básica del alisado exponencial, el cual está especialmente indicado para el tratamiento de series no estacionales y con una tendencia no definida, en el sentido de que ni es creciente ni decreciente.

Este procedimiento para el suavizado de una serie se basa en suponer que el nivel o valor medio de la serie en el instante *t* (y_t^*) se puede obtener de la forma siguiente:

$$
\begin{aligned}
& y_t^* = \alpha y_t + (1-\alpha) y_{t-1}^* = \\
& = \alpha y_t + (1-\alpha)\left[\alpha y_{t-1} + (1-\alpha) y_{t-2}^*\right] = \\
& = \alpha y_t + \alpha(1-\alpha) y_{t-1} + (1-\alpha)^2 y_{t-2}^* = \\
& = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \left[\alpha(1-\alpha)^2 y_{t-2} + (1-\alpha) y_{t-3}^*\right] = \\
& = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + (1-\alpha)^3 y_{t-3}^* = \\
& = \ldots\ldots\ldots\ldots = \\
& = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots\ldots\ldots + \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t y_0^* = \\
& = y_{t-1}^* + \alpha\left(y_t - y_{t-1}^*\right) \qquad (0 < \alpha < 1) \qquad (4.12)
\end{aligned}
$$

Según expresión, el nivel medio en un instante *t* cualquiera es una combinación lineal (media ponderada) del nivel del periodo anterior y del valor observado en ese periodo, pero que a su vez, ese nivel es, también, una media ponderada de todas las observaciones de la serie (la suma de todas las ponderaciones vale la unidad). Todas las ponderaciones son una función de α, a la que se le conoce como constante de suavizado. La elección del valor de esta constante dependerá de que la serie sea más o menos estable. Así, si la misma es muy estable el valor de α deberá estar próximo a la unidad, mientras que si la serie presenta gran volatilidad, entonces es aconsejable que α sea pequeño para evitar darle demasiado peso al último dato observado. En cualquier caso, la selección del valor de α implica introducir una cierta carga de subjetividad en el análisis de la serie, lo que no deja de ser un serio inconveniente.

La única cuestión que queda pendiente para que este procedimiento de alisado sea aplicable es la forma de obtener y_0*. Este valor se fija como: y_0*= y_0.

Ejemplo 5. *Obtener la serie suavizada del número de empresas que presentaron suspensión de pagos en España mediante un alisado exponencial simple.*

En la Tabla 6 se recoge la serie original y tres alisadas según distintos valores de α. Las mismas se han representado en la Figura 10. Se ha optado por trabajar con más de un valor para la constante de suavizado. Ello permite apreciar el efecto que tiene esa constante sobre la serie alisada. Así, cuando α=0,9, entonces la serie original y la suavizada son prácticamente la misma. En este caso, la serie se suaviza muy poco. Ese valor de α estaría indicado cuando la serie cambia muy poco de un periodo de tiempo a otro. Cuando α=0,1 la serie resultante si que elimina todas las crestas que presentaban los datos originales. Pero este suavizado puede resultar excesivo en algunos casos, como es el de nuestro ejemplo, pues lo que se persigue con el alisado exponencial de una serie es eliminar de la misma solo la componente residual, mientras que en nuestro ejemplo, con α=0,1, se ha eliminado algo más que eso. Casi se ha eliminado también la componente cíclica, la única que aparece de forma clara en la serie original. La tercera es una solución intermedia entre las dos anteriores (α=0,5).

Tabla 6. ***Evolución del número de suspensiones de pagos de empresas en España. Series original y alisadas.***

Años	*Suspensión de pagos*	*Serie alisada (α=0,9)*	*Serie alisada (α=0,1)*	*Serie alisada (α=0,5)*
1983	*841*	*841*	*841*	*841*
1984	*814*	*817*	*838*	*828*
1985	*459*	*495*	*800*	*643*
1986	*231*	*257*	*743*	*437*
1987	*188*	*195*	*688*	*313*
1988	*154*	*158*	*635*	*233*
1989	*167*	*166*	*588*	*200*
1990	*351*	*333*	*564*	*276*
1991	*798*	*751*	*587*	*537*
1992	*1135*	*1097*	*642*	*836*
1993	*1446*	*1411*	*723*	*1141*
1994	*969*	*1013*	*747*	*1055*
1995	*650*	*686*	*738*	*852*
1996	*649*	*653*	*729*	*751*

1997	*479*	*496*	*704*	*615*
1998	*348*	*363*	*668*	*481*
1999	*290*	*297*	*630*	*386*
2000	*223*	*230*	*590*	*304*

Fuente: Servidor web del INE y elaboración propia.

A continuación se detalla la forma en que se han obtenido esas series. Como puede apreciarse, el primer valor es idéntico para las tres e igual al de la serie original. Es decir:

$$y_{83}^* = y_{83} = 841$$

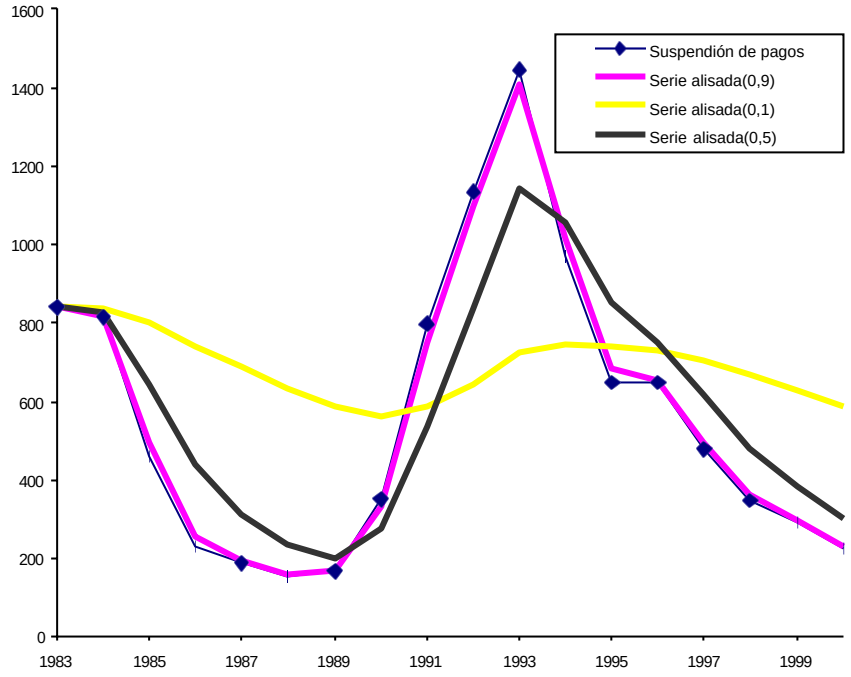
A partir de ese valor inicial, los demás se han calculado según se indicó más arriba:

$$y_{84}^* = (0,9)(y_{84}) + (1-0,9)(\ y_{83}^*) = (0,9)(814) + (0,1)(841) = 817$$

$$y_{85}^* = (0,9)(y_{85}) + (1-0,9)(\ y_{84}^*) = (0,9)(459) + (0,1)(817) = 495$$

y así para los demás años y los otros valores de α.

Figura 10. Evolución del número de suspensiones de pagos en España. Series original y alisadas exponencialmente



Cuando la serie que se pretende suavizar tiene una tendencia definida y es estacional, entonces el método de alisado exponencial simple que se acaba de exponer se sustituye por otros procedimientos que, basándose en él, permiten esas nuevas situaciones. Entre ellos está el de Holt-Winters.

4.4 Componente estacional.

Previamente ya se ha definido la componente estacional como aquellos movimientos de la serie que se repiten de forma periódica, siendo la periodicidad inferior al año. Estos movimientos de la serie, que se repiten de forma sistemática, dificultan la posibilidad de hacer comparaciones entre los valores sucesivos de una serie, pues el nivel medio de la misma se ve alterado por la estacionalidad. Para aclarar esta idea haremos uso de la variable paro registrado que se aparece en la Tabla 1. Para un año cualquiera de esta serie se puede observar como a partir de septiembre la misma empieza a tomar valores que son sistemáticamente superiores a los de meses anteriores. Este comportamiento, si se analizara para un solo año, podría hacer pensar que al serie ha cambiado de tendencia y que los niveles medios de la misma han empezado a crecer. Sin embargo ya se ha podido comprobar que ese no es el caso, pues la tendencia mostraba un perfil de continuo decrecimiento, que solo cambia a partir de la primavera del año 2001. Ese comportamiento anómalo de la tendencia, dentro de cada año, es el efecto de la estacionalidad, que hace que, para determinados meses (u otros periodos de tiempo inferiores al año), se observen movimientos ajenos a la tendencia (causados por motivos económicos, climatológicos, etc) que impiden una correcta comparación de los valores de esa serie en esos meses.

Para evitar esas distorsiones en los valores medios se recurre a lo que se conoce como desestacionalización de la serie o corrección estacional. Para realizar esta operación es necesario aislar en primer lugar la componente estacional, lo que posibilitará su posterior eliminación.

En los distintos métodos elementales que pueden utilizarse para obtener la componente estacional, siempre hay un paso previo que consiste en eliminar la tendencia, obtenida mediante alguno de los procedimientos señalados con anterioridad, medias móviles o regresión, principalmente.

Se centrará la atención en el que hace uso de la tendencia obtenida por media móviles. A este procedimiento se le conoce como razón a las medias móviles. La exposición de este procedimiento se realizará con el apoyo de un ejemplo. En concreto se usaran los datos de la Tabla 1, es decir, la serie de paro registrado. El primer paso a dar es suavizar la serie para de esa forma obtener la tendencia o, más bien, la componente tendencia-ciclo. Los resultados se recogen en la Tabla 3. A continuación, y admitiendo que la serie sigue un esquema multiplicativo (si fuera aditivo se procedería de una forma similar), se procede a dividir la serie original entre el resultado que se obtiene mediante aplicación de medias móviles. El resultado de ese cociente:

$$\frac{y_t}{Media \quad móvil} = \frac{TxVExCxR}{TxC} = VExR \qquad (4.13)$$

sería una componente mixta que recoge la estacionalidad y las variaciones residuales y se le conoce como razón a las medias móviles.

Tabla 7. Razón a las medias móviles. (VExR)

	1997	1998	1999	2000	2001	2002
Enero	,,	104,14	103,38	105,27	105,84	104,90
Febrero	,,	103,91	103,35	104,95	104,61	105,28
Marzo	,,	103,47	102,89	103,34	103,39	103,68
Abril	,,	100,97	101,07	100,54	100,56	102,33
Mayo	,,	98,77	98,54	97,83	96,76	98,84
Junio	,,	97,84	97,21	96,17	95,52	97,00
Julio	95,14	95,12	94,22	95,71	94,79	,,
Agosto	94,86	95,86	95,06	95,92	95,03	,,
Septiembre	98,04	97,72	96,63	97,10	96,60	,,
Octubre	100,42	99,78	98,62	99,21	99,47	,,
Noviembre	102,34	101,02	101,24	101,21	101,02	,,
Diciembre	102,39	101,15	101,23	101,43	100,56	,,

Como se ha señalado, el resultado (4.13) no es solo la componente estacional. También aparecen en el mismo las variaciones residuales. Si estas últimas no fueran muy importantes, entonces el contenido de la Tabla 7 nos daría directamente la estacionalidad de la serie estudiada en este ejemplo. A esos cocientes se les conoce como índices específicos de variación estacional. Tal es el caso del ejemplo que se viene desarrollando. En este ejemplo, además, puede observarse como el esquema

de estacionalidad cambia poco de año en año. Se trata de una estacionalidad estable, por lo que es preferible resumir todos esos índices específicos en uno general que recoja la estacionalidad de la serie. La manera más fácil de llegar al mismo es obtener un promedio de los valores de cada mes. Además, al proceder de esta forma lo que se consigue también es eliminar los posibles efectos de R, aunque estos sean poco importantes. El resultado de esta operación se recoge en la Tabla 8.

Tabla 8. **Índices generales de variación estacional (IGVE) del paro registrado en España**

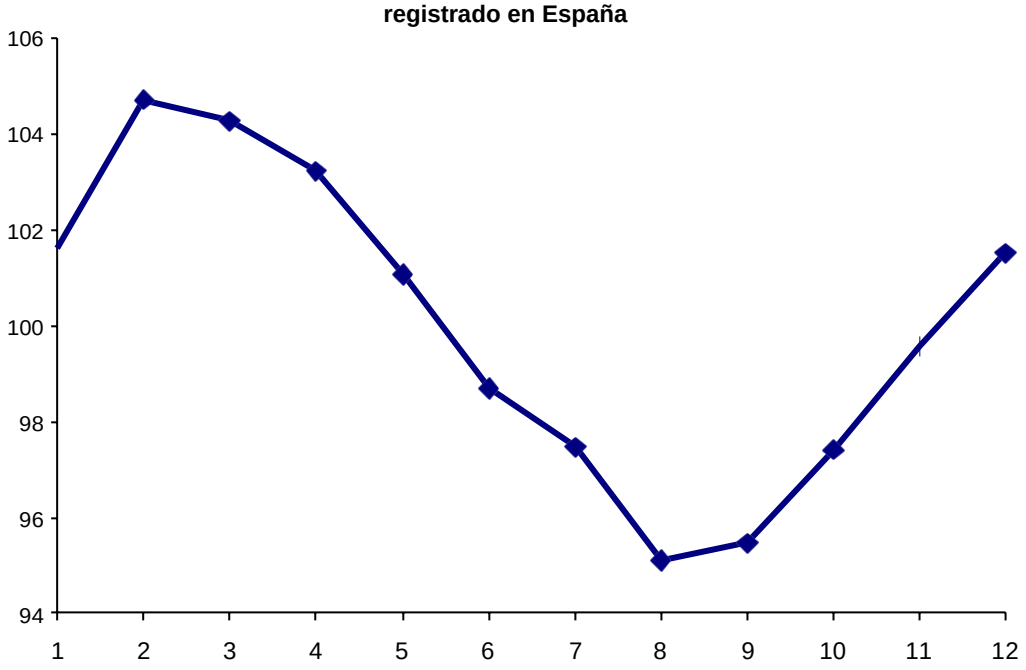
	Media	IGVE
Enero	104,71	104,86
Febrero	104,42	104,57
Marzo	103,35	103,50
Abril	101,09	101,24
Mayo	98,15	98,29
Junio	96,75	96,89
Julio	95,00	95,14
Agosto	95,35	95,49
Septiembre	97,22	97,36
Octubre	99,50	99,65
Noviembre	101,37	101,51
Diciembre	101,35	101,50

En esta tabla aparecen dos columnas, una encabezada con el nombre de media y otra con el de IGVE. La primera es el resultado de obtener la media, para cada mes, de los índices de la Tabla 7. Pero, en teoría, la media de todas esas medias debiera ser cien y, en este ejemplo, por centésimas, no lo es. Esto nos lleva a corregir estos índices generales para forzarlos a que cumplan el requisito de que su media sea cien, como ocurre en la segunda columna. La interpretación de estos índices, y para el caso de la serie de paro registrado, es que, debido a los efectos estacionales de la economía española, durante los meses de mayo a octubre el paro registrado se sitúa por debajo de la media anual, siendo durante julio y agosto cuando menos paro registrado se recoge en las Oficinas del INEM, mientras que, por el contrario, es en los meses de Enero y Febrero cuando la situación de demanda de empleo insatisfecha es mayor. Si los IGVE fueran todos igual a la unidad, entonces habría que concluir que la serie no presenta estacionalidad.

Si la estacionalidad fuera evolutiva habría que quedarse con los índices específicos. Si, además, las variaciones residuales ejercieran un efecto significativo en las razones a las medias móviles, entonces habría que eliminar esa componente. En estos casos,

como no tiene sentido calcular la media para cada mes, lo que suele hacerse es dividir y_t en tantas series independientes como periodos inferiores al años se estén considerando. Así si se trabaja con datos mensuales, el número de estas nuevas series sería doce, siendo la longitud de cada una de ellas el número de años de y_t. A continuación se suaviza cada una de esas series mediante medias móviles, con lo que se consigue eliminar la componente residual. El periodo de estas medias será de 3 a 5 años, todo ello en función de la longitud de la serie original, pues cuanto mayor es el periodo más garantía hay de haber eliminado R, pero también se pierde más información. Estos datos perdidos hay que extrapolarlos de alguna forma si lo que se persigue es desestacionalizar la serie completa.

Figura 11. Indices generales de variación estacional (IGVE) del paro registrado en España

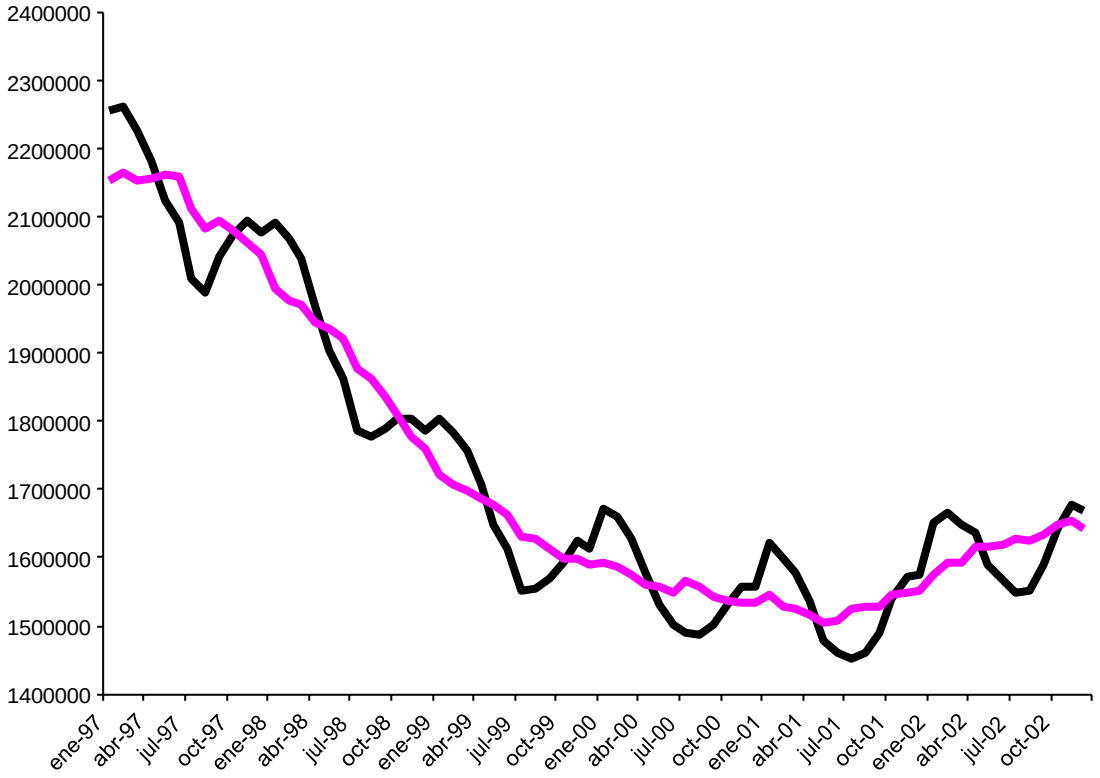


Una vez que se han obtenido los IGVE, lo que ayuda, junto con el análisis de la tendencia, al estudio descriptivo de la serie, el siguiente paso sería eliminar la estacionalidad de nuestra serie, es decir, habría que desestacionalizarla. Esto se consigue de la forma siguiente:

$$\frac{y_t}{IGVE} = \frac{TxVExCxR}{VE} = TxCxR \qquad (4.14)$$

Para el ejemplo del paro registrado la serie desestacionalizada aparece en la Tabla 8 a la vez que en la Figura 12.

Figura 12. Serie desestacionalizada del paro registrado.



Como puede observarse en la Figura 12, la serie desestacionalizada tiene un perfil muy parecido al de la tendencia obtenida por medias móviles, con la única diferencia que ésta última era más suave. La razón de esa diferencia radica en que en la desestacionalizada contiene la tendencia, los ciclos y las variaciones residuales, mientras que en la serie obtenida por medias móviles solo se recogen la tendencia y los ciclos.

Una vez que a la serie se le ha eliminado la componente estacional, los valores de cualquier mes si son comparables en lo que a niveles medios se refiere, pues ahora cada mes ha perdido la especificidad que el confería la estacionalidad de la serie. Se podría afirmar que los meses se han homogeneizado o estandarizado para que sean comparables entre si.

Tabla 8. **Serie de paro registrado en España desestacionalizada**.

	1997	1998	1999	2000	2001	2002
Enero	2151,9	1994,4	1720,6	1593,1	1545,6	1575,2
Febrero	2163,8	1977,4	1705,9	1587,2	1529,0	1593,2
Marzo	2152,1	1970,1	1697,7	1573,4	1525,0	1593,2
Abril	2154,9	1943,9	1687,1	1559,5	1516,3	1616,2
Mayo	2160,7	1935,2	1677,8	1557,8	1503,8	1616,6
Junio	2159,1	1920,4	1664,3	1548,3	1507,5	1617,7
Julio	2111,9	1877,3	1630,2	1564,9	1525,6	1627,6
Agosto	2083,0	1861,1	1627,9	1557,9	1528,0	1625,4
Septiembre	2095,4	1836,9	1612,6	1542,2	1528,9	1633,4
Octubre	2080,2	1810,1	1597,3	1535,6	1545,5	1647,5
Noviembre	2062,6	1777,6	1599,4	1533,6	1549,4	1652,9
Diciembre	2045,0	1759,3	1589,9	1533,4	1551,5	1643,4

Para finalizar este apartado hay que señalar que el procedimiento desarrollado para aislar la componente estacional no es el único. En lugar de trabajar con la tendencia obtenida por medias móviles se podría haber trabajado con la tendencia deducida mediante mínimos cuadrados. Pero en este caso lo más probable es que en el ajuste se recoja solo la tendencia, por lo que al dividir la serie original entre esta, el resultado contenga, además de las variaciones estacionales, los ciclos y las variaciones residuales. Pero como la periodicidad de los ciclos es poco regular, su eliminación puede influir negativamente en la obtención de la componente estacional. Por esta razón puede afirmarse que es preferible trabajar con la tendencia deducida por media móviles. De hecho, en esta idea se basa uno de los métodos de desestacionalización más ampliamente utilizado, como es el *X-9* y su posterior desarrollo el *X-11*, elaborado por el Bureau of the Census de Estados Unidos y que aparece en la mayoría de los paquetes estadísticos.

4.5 Variaciones cíclicas y residuales.

La obtención de la tercera componente sistemática de una serie temporal es, sin lugar a dudas, la que entraña más problemas. Esto es así por tres razones como mínimo. En primer lugar no siempre existe. En segundo lugar hace falta mucha información, series muy largas, para que pueda detectarse la presencia de esta componente. En tercer lugar, es la menos sistemática de las tres, pues los ciclos, cuando existen, no siempre tienen la misma longitud y, además, se puede dar el caso de que se superpongan más de un ciclo de distintas longitudes de onda. Todo ello hace que sea

muy frecuente el tratamiento de las series en el que se prescinde del estudio separado de los ciclos y, en su lugar, se trabaja con la componente mixta ciclo-tendencia, a la que también se le conoce como extraestacional.

Pero para no dejar olvidada esta componente vamos a seguir con el enfoque multiplicativo usado hasta el momento e indicar los pasos a seguir para su obtención. Algunos de esos pasos ya se han dado en los apartados anteriores, por lo que ahora solo se hará alusión a los mismos.

En primer lugar habría que obtener la componente estacional para desestacionalizar la serie. Ello nos llevaría a:

$$\frac{y_t}{IGVE} = \frac{TxVExCxR}{VE} = TxCxR \qquad (4.15)$$

En segundo lugar hay que obtener la tendencia para eliminarla de la serie:

$$\frac{TxCxR}{T} = CxR \qquad (4.16)$$

Por último habría que suavizar, mediante medias móviles o por cualquier otro procedimiento de alisado, esa serie resultante, lo que daría, finalmente, la componente cíclica buscada.

La última componente son las variaciones residuales. La misma tiene poco interés y se obtendría como cociente entre (4.16) y las variaciones cíclicas. La utilidad de esta última componente se basa en poder verificar si satisface ciertos supuestos o hipótesis, como el que sea realmente aleatoria. Pero esto no está al alcance del material que se está manejando en este manual.

4.6 Medición de las variaciones de una serie.

En los apartados anteriores se ha descrito un esquema metodológico que permite analizar una serie temporal a partir del estudio de un conjunto de componentes

definidas previamente. Que duda cabe que esta es una forma de abordar el problema. Pero desde luego que no es la única. Otra posible forma de estudiar el comportamiento de una serie es apoyándose en otras, es decir, tratando de explicar sus variaciones como consecuencia de las variaciones de otra u otras series. En este sentido, y siguiendo con el ejemplo del paro registrado que se ha utilizado en el apartado anterior, se podría haber planteado que el paro es el resultado de la producción, de los salarios y de la oferta de mano de obra medida por la población activa. Este tipo de planteamiento nos llevaría a buscar una función que ligue esas variables para después cuantificarla mediante el análisis de la regresión. Estaríamos frente a una forma distinta de abordar el estudio de las variaciones de una serie. Pero tampoco es procedimiento agota todas las posibilidades de análisis.

En este epígrafe se presentará otra forma de analizar el comportamiento de las variaciones de una serie temporal. Ahora lo que se buscará será simplemente cuantificar los cambios que experimenta la serie al pasar de un periodo de tiempo a otro. Esos periodos pueden ser consecutivos (mes e a mes, trimestre a trimestre, año a año, etc.) o puede que estén separados entre si (un determinado mes-trimestre de un año con respecto a ese mismo mes-trimestre del año anterior). Si la serie temporal la seguimos representando por y_t, entonces la forma más simple de cuantificar la variación de la misma es:

$$\Delta y_t = y_t - y_{t-1} \qquad (4.17)$$

Esta relación lo que nos dice, mediante el signo de la misma, es si la serie, para ese periodo considerado, está creciendo o decreciendo, según que aquel sea positivo o negativo, respectivamente. Pero aparte de esa información poco más nos dice, pues esa diferencia o variación observada en la serie viene expresada en las mismas unidades de medida que la propia serie, por lo que bastaría con realizar un cambio de escala para que podamos modificar la magnitud de esa variación de manera arbitraria. Pero las limitaciones de esta forma de medir las variaciones de una serie no se limitan a la ya mencionada. Puede darse también el caso de cambios de igual magnitud para dos variables distintas pero que no sean comparables si los niveles de esas series son muy distintos o si las mismas vienen expresadas en unidades medida distintas. Así por ejemplo no es comparable el crecimiento de diez euros en la renta de una familia que tiene un nivel de ingresos de mil euros que ese mismo crecimiento para otra familia con unos ingresos de diez mil euros.

Antes de pasar a estudiar las posibles maneras de solucionar las limitaciones que presenta esta forma tan simple de medir las variaciones de una serie, puede resultar de interés detenerse un momento e indagar un poco en el significado de esa diferencia. La misma, por si sola, es otra serie que para el caso de datos anuales y tendencia lineal recogería las componentes cíclica y residual. Se trataría de una serie filtrada de tendencia. Pero si la periodicidad fuera inferior a la anual, por ejemplo mensual, entonces, y como ya se ha señalado, las diferencias podría tomarse respecto al mes anterior o al mismo mes del año anterior. En el segundo caso el resultado es otra serie filtrada de estacionalidad y tendencia y que recoge, de nuevo, las componentes cíclica y residual. Esto es lo que se ha hecho con la serie del paro registrado y cuyo resultado aparece en la Figura 13[4].

Esta serie resultante muestra que desde medidos de 1999 empezó a tener lugar no un cambio de tendencia, como se ha indicado antes, sino más bien una fase distinta del ciclo, en la que el decrecimiento de la misma es cada vez menor. Ese decrecimiento se fue anulando de manera que para octubre de 2001 las variaciones interanuales eran ya positivas. Vemos como esta forma de tratar la serie puede complementar el instrumental analítico presentado en los epígrafes anteriores.

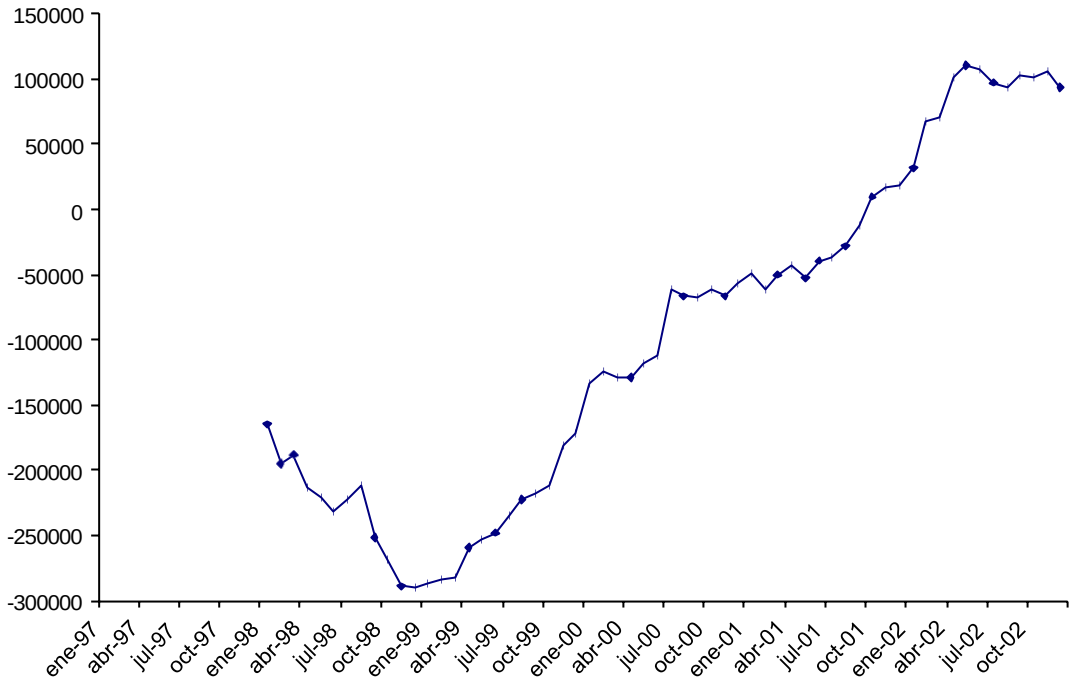
Pero la cuantificación de las variaciones de una serie en términos absolutos ya hemos visto que presenta ciertas limitaciones que es conveniente corregir. La forma más simple de eliminar esos problemas asociados a las unidades de medida de la serie es trabajar en términos relativos, es decir, medir las variaciones de forma adimensional. Esto se consigue cuando se trabaja con tasas de variación. Estas se obtienen comparando la variación intertemporal de la serie con respecto al valor inicial de la misma. De acuerdo con esta definición, las tasas vendrían dadas por:

$$T(h,n) = T_h^n = \frac{\Delta y_t}{y_{t-h}} = \frac{y_t - y_{t-h}}{y_{t-h}} = \frac{y_t}{y_{t-h}} - 1 \qquad (4.18)$$

[4] Hay que indicar que esta forma de eliminar la componente tendencia de una serie solo da buenos resultados si la misma es lineal, pues si las variaciones fueran de tipo exponencial, entonces, la diferenciación sucesiva de la serie no garantiza que se elimine completamente la tendencia.

donde *h* indica el número de periodos que hay entre los valores comparados y *n* es el número de observaciones que se utilizan para el cálculo de la tasa. Las tasas se pueden expresar en tantos por uno, aunque lo más habitual es que se multipliquen por cien, o cualquier otra potencia de diez, en cuyo caso se hablaría de porcentajes o lo que corresponda. Tanto por adimensionalidad como por hecho de que se está comparando variaciones absolutas con niveles de la variable, hace que esas tasas sean comparables con las obtenidas para otro tipo de variables o con la misma variable para periodos de tiempo muy alejados entre si.

Figura 13. Serie filtrada de tendencia y estacionalidad



A partir de (4.18) se deduce que:

$$y_t = \left(T\left(h,n\right)+1\right)y_{t-h} \qquad \left(4.19\right)$$

En función de los valores de *h* y *n*, las tasas más habituales que suelen calcularse son las siguientes:

a)

$$T_1^1 = \left| \left(\frac{y_t}{y_{t-1}} - 1 \right) \right| x100 \qquad (4.20)$$

Esta expresión se utilizaría para calcular las tasas de variación para series con datos anuales. También podría utilizarse para series cuya periodicidad sea inferior al año (mensual, trimestral, etc.). Pero en estos casos, si la serie presenta estacionalidad, no es aconsejable hacer uso de la misma para calcular las tasas de variación, pues ese factor distorsiona el valor de las mismas.

b)

$$T_{12}^1 = \left| \left(\frac{y_t}{y_{t-12}} - 1 \right) \right| x100 \qquad (4.21)$$

Para datos mensuales, esto es una tasa de variación anual en un mes concreto y que se puede obtener en meses sucesivos (tasa interanual). La misma puede utilizarse para datos con y sin estacionalidad, pues los periodos que se comparan son homogéneos, el mismo mes de cada año. Es aplicable tanto a variables flujos (producción, ventas, etc.) como stock (paro registrado, empleo de una empresa, etc).

c)

$$T_6^1 = \left| \left(\frac{y_t}{y_{t-6}} - 1 \right) \right| x100 \qquad (4.22)$$

Si los datos fueran mensuales se trataría de tasas semestrales en un mes concreto y que se pueden obtener en meses sucesivos. No debe utilizarse cuando la serie presente estacionalidad, pues los meses que se comparan no son homogéneos (diciembre-junio, noviembre-mayo, etc). Como la anterior, es aplicable tanto a variables flujo como de naturaleza stock.

d)

$$T_4^1 = \left| \left(\frac{y_t}{y_{t-4}} - 1 \right) \right| x100 \qquad (4.23)$$

Para datos mensuales da tasas de variación cuatrimestral en un mes concreto y se pueden obtener en meses sucesivos. A este tipo de tasas le son aplicables los comentarios hechos a las semestrales en lo referido a la estacionalidad. Si la serie fuera trimestral, esta expresión nos daría las tasas de variación anual en un trimestre concreto y además ahora si que sería aplicable cuando hay estacionalidad (primer trimestre de t-primer trimestre de t-1, etc.).

e)

$$T_3^1 = \left| \left(\frac{y_t}{y_{t-3}} - 1 \right) \right| x100 \qquad (4.24)$$

Para datos mensuales da tasas de variación trimestral en un mes concreto y se pueden obtener en meses sucesivos. A este tipo de tasas le son aplicables los comentarios hechos a las semestrales y cuatrimestrales en lo referido a la estacionalidad. Si la serie fuera cuatrimestral, esta expresión nos daría las tasas de variación anual en un cuatrimestre concreto y además ahora si que sería aplicable cuando hay estacionalidad (primer cuatrimestre de t-primer cuatrimestre de t-1, etc.)

f)

$$T_{12}^{12} = \left| \left(\frac{y_t + y_{t-1} + ... + y_{t-11}}{y_{t-12} + y_{t-13} + ... + y_{t-23}} - 1 \right) \right| x100 \qquad (4.25)$$

Cuando los datos son mensuales, ahora se estaría obteniendo tasas de variación anual. Esta expresión solo es aplicable a variables de tipo flujo con o sin estacionalidad.

g)

$$T_3^3 = \left(\frac{y_t + y_{t-1} + y_{t-2}}{y_{t-3} + y_{t-4} + y_{t-5}} - 1 \right) x100 \qquad (4.26)$$

Si los datos son mensuales, con esta expresión se obtienen tasas de variación trimestral para variables de tipo flujo. No es muy aconsejable utilizarla si la serie presenta estacionalidad.

h)

$$T_1^{12} = \left(\frac{y_t + y_{t-1} + ... + y_{t-11}}{y_{t-1} + y_{t-2} + ... + y_{t-12}} - 1 \right) x100 \qquad (4.27)$$

Esta expresión, con datos mensuales, lo que da son tasas mensuales basadas en medias móviles anuales.

i)

$$T_1^3 = \left(\frac{y_t + y_{t-1} + y_{t-2}}{y_{t-1} + y_{t-2} + y_{t-3}} - 1 \right) x100 \qquad (4.28)$$

Esta expresión, con datos mensuales, lo que da son tasas mensuales basadas en medias móviles trimestrales.

Para finalizar este epígrafe hay que hacer referencia a dos cuestiones que tienen un gran interés desde el punto de vista práctico. Se trata, en primer lugar, de la conversión (elevación) de tasas con periodicidad inferior al año a tasas anuales y, en segundo lugar, de la obtención de las tasas medias mensuales, o de otro periodo, en un año.

Por lo que respecta al primer punto, si admitimos inicialmente que se trabaja con tasas mensuales, entonces las tasa mensual elevada a anual sería aquella que representara el mismo crecimiento absoluto que la anual. Bajo este supuesto se tiene que, cuando se trabaja con tasas mensuales (que representaremos por T_i , i =1,2,...,12), el valor de la serie en el último mes vendrá dado por:

$$y_{12} = y_0 (1+T_1)(1+T_2)....(1+T_{12}) \qquad (4.29)$$

pues:

$$y_1 = y_0 (1+T_1)$$
$$y_2 = y_1 (1+T_2)$$
$$.$$
$$.$$
$$y_{12} = y_{11} (1+T_{12})$$

Por otro lado, según la tasa anual (T_a) se tiene que:

$$y_{12} = y_0 (1+T_a) \qquad (4.30)$$

Si ahora se iguala (4.29) y (4.30) el resultado es que:

$$T_a = (1+T_1)(1+T_2)....(1+T_{12}) - 1 \qquad (4.31)$$

Si además se supone que las tasas mensuales son constantes e iguales a T_m, entonces:

$$T_a = (1+T_m)^{12} - 1 \qquad (4.32)$$

En general esa tasa anualizada vendría dada por:

$$T_a = \left(1 + T_{12/h} \right)^{12/h} - 1 \qquad (4.33)$$

La segunda cuestión planteada era la de obtener una tasa media mensual a partir de las doce tasas. Esa tasa, que la representaremos por TM, se aquella que satisfaga la relación:

$$(1+TM)^{12} = \left[(1+T_1)(1+T_2).....(1+T_{12}) \right] \qquad (4.34)$$

De donde se tiene que:

$$TM = \sqrt[12]{\left[(1+T_1)(1+T_2)\,..\,(1+T_{12})\right]}-1 \qquad (4.35)$$

Es decir, la tasa media mensual no es la media aritmética de las tasas mensuales, sino la media geométrica de los factores de variación unitaria de cada mes ($1+T_i$) menos la unidad.

Ejemplo 6*. La fabricación mensual de vehículos (turismos) en España durante los años 1998, 1999 y 2000 fue la que se recoge en la tabla siguiente:*

	1998	*1999*	*2000*
Enero	*182630*	*191620*	*186070*
Febrero	*191620*	*214600*	*225570*
Marzo	*213540*	*230990*	*247000*
Abril	*196140*	*228600*	*195030*
Mayo	*205900*	*221610*	*252360*
Junio	*221080*	*220660*	*226210*
Julio	*200220*	*211590*	*217540*
Agosto	*60190*	*40070*	*60030*
Septiembre	*207430*	*212930*	*202780*
Octubre	*217990*	*198690*	*214570*
Noviembre	*210580*	*161350*	*238970*
Diciembre	*164530*	*211950*	*156810*
Total	*2271850*	*2344660*	*2422940*

Fuente: Página web del INE

Analizar la variación de esta serie en términos de:

a) Tasas mensuales de variación
b) Tasas anuales de variación
c) Tasas anuales mes a mes

a) Para obtener las tasas mensuales de variación habría que aplicar (4.20), lo que da como resultado:

	1998	*1999*	*2000*
Enero	*,,*	*16,47*	*-12,21*
Febrero	*4,92*	*11,99*	*21,23*
Marzo	*11,44*	*7,64*	*9,50*
Abril	*-8,15*	*-1,03*	*-21,04*
Mayo	*4,98*	*-3,06*	*29,40*
Junio	*7,37*	*-0,43*	*-10,36*
Julio	*-9,44*	*-4,11*	*-3,83*
Agosto	*-69,94*	*-81,06*	*-72,41*

Septiembre	*244,63*	*431,40*	*237,80*
Octubre	*5,09*	*-6,69*	*5,81*
Noviembre	*-3,40*	*-18,79*	*11,37*
Diciembre	*-21,87*	*31,36*	*-34,38*

Donde la tasa para febrero de 1998 es el resultado de:

$$4{,}92 = \left(\frac{191620}{182630} - 1 \right) x100$$

y así para los demás meses.

Como puede apreciarse, estas tasas se ven afectadas por todas las componentes de la serie. Así, los valores de agosto y septiembre son un claro exponente de la estacionalidad de al serie, mientras que la componente errática o residual se manifiesta en distintos meses de esta serie.

b) En este caso, como la serie es un flujo, para obtener las tasas de variación anual lo primero que debe hacerse es sumar la producción de los doce meses, lo que nos daría la producción anual para cada uno de los tres años considerados y, a partir de esos datos, se obtendrían las tasas correspondientes. La mismas son el resultado de aplicar, de nuevo, la expresión (4.20):

$$3{,}20 = \left(\frac{2344660}{2271850} - 1 \right) x100 \qquad 3{,}34 = \left(\frac{2422940}{2344660} - 1 \right) x100$$

Es decir, que la producción de 1999 fue superior a la de 1998 en un 3,20% y la de 2000 superó a la de 1999 en un 3,34%.

A partir de las tasas de variación mensual del apartado anterior, también se puede calcular la tasa mensual elevada a anual. Para ello habría que obtener la tasa media mensual para cada año aplicando (4.35):

$$TM = \sqrt[12]{[(1+T_1)(1+T_2)....(1+T_{12})]} - 1 = \sqrt[12]{\left[\left(\frac{Enero_t}{Diciembre_{t-1}}\right)\left(\frac{Febrero_t}{Enero_t}\right).....\left(\frac{Diciembre_t}{Noviembre_t}\right)\right]} - 1 =$$

$$= \sqrt[12]{\left[\left(\frac{Diciembre_t}{Diciembre_{t-1}}\right)\right]} - 1$$

$$\text{pues } (1+T) = \left[1 + \left(\frac{y_t}{y_{t-1}} - 1\right)\right] = \left(\frac{y_t}{y_{t-1}}\right)$$

Todo esto lleva a que esa tasa, para 1999 sea:

$$T_m(1999) = \sqrt[12]{\left[\left(\frac{Diciembre_t}{Diciembre_{t-1}}\right)\right]} - 1 = \sqrt[12]{\left[\left(\frac{211950}{164530}\right)\right]} - 1 = 0{,}0213$$

y para 2000

$$T_m(2000) = \sqrt[12]{\left[\left(\frac{Diciembre_t}{Diciembre_{t-1}}\right)\right]} - 1 = \sqrt[12]{\left[\left(\frac{156810}{211950}\right)\right]} - 1 = -0{,}0248$$

Una vez que han obtenido esas tasas medias mensuales, para elevarlas a anuales hay hacer uso de (4.32). El resultado de ello es:

$$T_a(1999) = (1+0.0213)^{12} - 1 = 0{,}2882$$

$$T_a(2000) = (1-0.0248)^{12} - 1 = -0{,}26$$

Estos resultados, si se comparan con los obtenidos antes en este mismo apartado, merecen ser comentados. En primer lugar hay que señalar que las tasas medias mensuales elevadas a anuales no son otra cosa que las interanuales que se dan en el siguiente apartado. En segundo lugar, estas tasas se obtienen a partir de la información que suministran solo dos meses, sin tener en cuenta, para nada, lo ocurrido en los demás, de forma que si en esos periodos la componente errática (por ejemplo, una huelga) es decisiva, entonces condiciona la tasa de todo el año, cuando eso no tiene razón de ser.

Esto lleva a que este procedimiento de obtener las tasas anuales deba usarse con cautela y, desde luego, no es aconsejable cuando se trabaja con variables flujo. Estarían más indicadas en los casos de magnitudes de tipo stock, donde un mes cualquiera puede ser "representativo" de los demás, pues como se sabe, el valor del año no es la suma de los valores de todos los meses. También se pueden utilizar como un método de proyectar la variación de todo un año, cuando solo se conoce lo ocurrido en los primeros meses y se asume que para los demás la tendencia en la variación será similar a la observada en esos periodos iniciales.

Para finalizar, hay que indicar que las tasas próximas al 3%, obtenidas en primer lugar, reflejan de forma más fidedigna lo ocurrido en esos dos años que lo que sugieren las mensuales elevadas a anuales.

c) *Las tasas anuales mes a mes (tasas interanuales) se obtienen por la aplicación de (4.21) dando como resultado:*

	1998	*1999*	*2000*
Enero	,,	*4,92*	*-2,90*
Febrero	,,	*11,99*	*5,11*
Marzo	,,	*8,17*	*6,93*
Abril	,,	*16,55*	*-14,69*
Mayo	,,	*7,63*	*13,88*
Junio	,,	*-0,19*	*2,52*
Julio	,,	*5,68*	*2,81*
Agosto	,,	*-33,43*	*49,81*
Septiembre	,,	*2,65*	*-4,77*
Octubre	,,	*-8,85*	*7,99*
Noviembre	,,	*-23,38*	*48,11*
Diciembre	,,	*28,82*	*-26,02*

En este caso se pierden los datos de 1998 al no disponerse de los correspondientes a 1997, pues las tasas se obtienen comparando el valor de un mes para un año con el del mismo mes el año anterior. Es decir, siempre se perderá un año.

Estas tasas no reflejan la estacionalidad de al serie, siempre que la misma sea estable. En cambio, si que recogen las otras tres componentes. Es como si se trabajara con una serie desestacionalizada.

CAPÍTULO 5.- NÚMEROS ÍNDICES.

5.1 Introducción.

Ya hemos visto que una de las principales preocupaciones de la Estadística es el análisis de variables, tanto consideradas individualmente como en conjunto. Para realizar tal tipo de análisis estadístico se han definido distintos instrumentos que han facilitado, no solo el análisis individualizado de cada variable, sino que algunos de ellos adquirían mayor entidad cuando se utilizaban para comparar variables.

Este problema de la comparación es de gran importancia en estadística. Las comparaciones entre variables o entre los valores de una sola variable pueden realizarse de distintas formas. Las más simples son las que se llevan a cabo por diferencia o aquellas que se realizan por cociente. Estas segundas tiene la ventaja frente a las primeras que eliminan el problema de las unidades de medida, que como hemos podido comprobar a lo largo de las lecciones anteriores es un verdadero problema. En cambio el segundo procedimiento, aunque no adolece de ese problema, no deja de estar afectado por otros, como el de elegir la unidad de referencia para realizar las comparaciones.

Este problema de la comparación estadística se resuelve en buena manera mediante el uso de números índices. En general diremos que un número índice es aquella medida estadística que permite estudiar las fluctuaciones o variaciones de una sola magnitud o de más de una en relación al tiempo o al espacio. Los índices más habituales son los que realizan las comparaciones en el tiempo, por lo que, como veremos más adelante, los números índices son en realidad series temporales.

Como puede verse, este nuevo concepto que acaba de introducirse, es muy parecido al de tasa de variación que se estudió en el capitulo anterior.

5.2 Índices simples.

Si la comparación se realiza para los valores de una sola magnitud, hablaremos de índices simples. En cambio, cuando se trabaja con más de una magnitud a la vez, hablaremos de índices complejos. En cualquiera de los dos casos vamos a comparar siempre dos situaciones, una de las cuales se considera de referencia. A la situación inicial, cuando las comparaciones son temporales, se le conoce como periodo base o referencia, frente al periodo corriente o actual con el que se realiza la comparación.

En la construcción de un número índice se le asigna al periodo de referencia el valor 100. Esto implica que los números índices no son otra cosa que porcentajes. Se trata de los porcentajes de cada valor de la magnitud con respecto al valor de referencia o base. Al ser los número índices porcentajes definidos sobre los propias valores de la variable hace que sean adimensionales, lo que permite la comparación de las variaciones de distintas variables que pueden venir expresadas en unidades diferentes.

Formalmente, un índice simple, para una variable concreta, se define de la forma siguiente:

$$I_0^t = I_0^t(i) = \frac{y_{it}}{y_{i0}} x100 \qquad (6.1)$$

Donde y_{it} y y_{i0} son dos valores concretos de una magnitud o variable Y_i. El primero de los valores corresponde al momento actual (t) y el segundo al momento base o de referencia ($t=0$). Una vez que se han elaborado lo números índices, según se recoge en (6.1), es fácil determinar la variación, en términos porcentuales, que ha sufrido la variable Y_i al pasar del periodo de referencia al actual.

***Ejemplo 1**. Obtenga los índices simples para el paro estimado en España y Andalucía.*

En la Tabla 1 se dan los datos para el periodo que va de 1981 a 2001 correspondientes al paro estimado por el INE a través de la Encuesta de Población Activa (EPA). Estas dos variables constituyen dos series temporales para las que pretende analizar su evolución en ese conjunto de años. Una forma de realizar ese estudio es recurriendo a la

construcción de números índices, pues de la simple observación de las mismas se saca poca información y, en consecuencia, es difícil obtener conclusiones. Lo más que se puede decir es que una es sistemáticamente superior a la otra, lo cual es una trivialidad, pues en los datos de España están incluidos los de Andalucía. Por otro lado, durante esos años, ni siquiera mantienen una tendencia definida.

Todo ello ha llevado a elaborar los índices de la Tabla 1 que reflejan la evolución del paro en Andalucía y España y permiten, no solo determinar cual ha sido el ritmo de variación de esta magnitud en cualquiera de estas dos áreas, sino que además posibilitan que se puedan realizar comparaciones entre ambas, pues si se hubiera trabajado en términos absolutos, entonces, no sería posible comparar las dos variables, pese a que ambas están expresadas en las mismas unidades de medida, pues las mismas hacen referencia a áreas geográficas muy diferentes.

Estos índices se han obtenido aplicando, de manera reiterada, la expresión (6.1) en la forma siguiente:

$$\begin{array}{cc} \textit{Andalucía} & \textit{España} \\ I_{1981}^{1981} = \frac{388,6}{388,6} x100 = 100,0 & I_{1981}^{1981} = \frac{1853,7}{1853,7} x100 = 100,0 \\ I_{1981}^{1982} = \frac{403,3}{388,6} x100 = 103,8 & I_{1981}^{1982} = \frac{2120,5}{1853,7} x100 = 114,4 \\ I_{1981}^{1983} = \frac{453,2}{388,6} x100 = 116,6 & I_{1981}^{1983} = \frac{2340,5}{1853,7} x100 = 126,3 \\ \vdots & \\ I_{1981}^{2001} = \frac{874,5}{388,6} x100 = 165,3 & I_{1981}^{2001} = \frac{2213,3}{1853,7} x100 = 119,39 \end{array}$$

Estos índices se han calculado tomando una base fija. Esto tiene el inconveniente de que si el periodo de referencia tomado como base es un valor anómalo, esta incidencia repercutirá de forma negativa en todos los valores del índice calculado. Por lo que es de suma importancia que el val or que setome como referencia sea “normal”.

Tabla 1. Paro estimado (índices simples)

	Valores Observados (miles de personas)		Números índices simples (Base 1981=100)	
Años	Andalucía	España	Andalucía	España
1981	388,6	1.853,7	100,00	100,00
1982	403,3	2.120,5	103,79	114,39
1983	453,2	2.340,5	116,62	126,26
1984	592,1	2.728,2	152,37	147,18
1985	619,7	2.938,5	159,48	158,52
1986	652,8	2.933,0	168,00	158,22
1987	705,1	2.937,7	181,46	158,48
1988	688,6	2.847,9	177,22	153,63
1989	652,9	2.560,8	168,02	138,15
1990	626,1	2.441,2	161,14	131,69
1991	638,1	2.463,7	164,21	132,91
1992	706,2	2.788,5	181,75	150,43
1993	836,6	3.481,3	215,30	187,80
1994	894,2	3.738,1	230,11	201,66
1995	888,4	3.583,5	228,62	193,32
1996	875,3	3.540,1	225,26	190,97
1997	874,6	3.356,5	225,07	181,07
1998	818,5	3.060,3	210,63	165,09
1999	759,5	2.605,5	195,44	140,56
2000	703,1	2.370,4	180,95	127,87
2001	642,5	2.213,1	165,34	119,39

Fuente: EPA. INE. Elaboración propia.

Una forma de evitar este problema de selección del periodo base es hacer que el mismo sea variable. En tal caso llegamos a lo que se conoce como índices en cadena . En este caso, esta modalidad de números índices permite obtener las variaciones porcentuales de una magnitud en un periodo con respecto, siempre, al anterior. Un ejemplo de este tipo de índices viene recogido en la tabla 2 y los mismos se han obtenido de la manera siguiente:

Andalucía *España*

$I_{1980}^{1981} = ..$ $I_{1980}^{1981} = ..$

$$I_{1981}^{1982} = \frac{403,3}{388,6} x100 = 103,8 \quad I_{1981}^{1982} = \frac{2120,5}{1853,7} x100 = 114,4$$

$$I_{1982}^{1983} = \frac{453,2}{403,3} x100 = 112,4 \quad I_{1982}^{1983} = \frac{2340,5}{2120,5} x100 = 110,4$$

. . . . .

$$I_{2000}^{2001} = \frac{642,5}{703,1} x100 = 91,4 \quad I_{2000}^{2001} = \frac{2213,3}{2370,4} x100 = 93,4$$

Tabla 2. Paro estimado (índices en cadena)

	Valores observados (miles de personas)		Índices en cadena	
Años	Andalucía	España	Andalucía	España
1981	388,6	1.853,7	,,	,,
1982	403,3	2.120,5	103,79	114,39
1983	453,2	2.340,5	112,36	110,38
1984	592,1	2.728,2	130,66	116,57
1985	619,7	2.938,5	104,67	107,71
1986	652,8	2.933,0	105,34	99,81
1987	705,1	2.937,7	108,02	100,16
1988	688,6	2.847,9	97,66	96,94
1989	652,9	2.560,8	94,81	89,92
1990	626,1	2.441,2	95,90	95,33
1991	638,1	2.463,7	101,91	100,92
1992	706,2	2.788,5	110,68	113,18
1993	836,6	3.481,3	118,46	124,84
1994	894,2	3.738,1	106,88	107,38
1995	888,4	3.583,5	99,35	95,86
1996	875,3	3.540,1	98,53	98,79
1997	874,6	3.356,5	99,91	94,81
1998	818,5	3.060,3	93,59	91,18
1999	759,5	2.605,5	92,79	85,14
2000	703,1	2.370,4	92,58	90,98
2001	642,5	2.213,1	91,37	93,36

Fuente: EPA. INE. Elaboración propia.

Los datos de la Tabla 2 muestran dos periodos de fuerte crecimiento del paro, los primeros años de la década de los ochenta y los de los noventa, siendo durante 1984

para Andalucía y 1993 para España cuando la tasa de crecimiento fue más elevada, un 30,7% y un 24,8 respectivamente. Por el contrario, durante el trienio 1988-90 y, especialmente, a partir de 1995 se observa como el volumen de parados, tanto en Andalucía como en España, decreció de forma continuada. A las mismas conclusiones se habría llegado si hubiéramos trabajado con los datos de la Tabla 1, aunque en ese caso la variación interanual no es tan evidente.

Pese a que el Ejemplo 1 haga referencia a una magnitud medida en términos de personas, sin embargo los números índices más habituales utilizados en Economía son los que hacen referencia a precios (medidos en unidades monetarias por unidad física), cantidades (medios en unidades físicas) y valor (medidos en unidades monetarias).

De acuerdo con la definición general de número índice dada con anterioridad, estas tres modalidades de índices se expresan en la forma siguiente:

1º. Indice de precios. Se define, para un bien *i*, como el cociente entre el precio de ese bien en el periodo *t* (p_{it}) y el precio de dicho bien en el perdido base (p_{i0}):

$$p_0^t(i) = \frac{p_{it}}{p_{i0}} x100 \qquad (6.2)$$

2º. Indice de cantidad. Se define, para un bien *i*, como el cociente entre la cantidad de ese bien en el periodo *t* (q_{it}) y la cantidad de dicho bien en el perdido base (q_{i0}):

$$q_0^t(i) = \frac{q_{it}}{q_{i0}} x100 \qquad (6.3)$$

3º. Indice de valor. Si se define el valor de un bien *i* en un periodo cualquiera como el producto del precio de ese bien por la cantidad del mismo (producida, vendida o comprada), entonces el índice de valor será el cociente entre el valor de ese bien ($p_{it}\ q_{it}$) en el periodo actual *t* y el valor del mismo en el periodo base ($p_{i0} q_{i0}$):

$$V_0^t(i) = \frac{V_{it}}{V_{i0}} x100 = \frac{p_{it}q_{it}}{p_{i0}q_{i0}} x100 = \left[\left(\frac{p_{it}}{p_{i0}}\right)\left(\frac{q_{it}}{q_{i0}}\right)\right] x100 \qquad (6.3)$$

De donde vemos que el índice de valor es el producto de los índices de precios y cantidades.

Todos estos índices, como ya hemos venidos señalando, deben expresarse en forma de porcentajes.

5.3. Índices compuestos o complejos no ponderados.

Una vez definidos los índices de precios, cantidades y valor aplicados todos al caso de un solo bien, el siguiente paso que debemos dar es la construcción de índices de esa naturaleza pero que abarquen más de un bien simultáneamente. Ello nos llevará al concepto de índice compuesto o complejo. En general, este índice compuesto no será otra cosa que la agregación de los distintos índices simples elaborados para cada bien por separado. Sin embargo, en otras ocasiones, lo que se agregan no son índices, sino las propias magnitudes (precios o cantidades) observadas.

La agregación puede realizarse según distintos métodos o procedimientos. Ahora bien, el que se elija ha de reunir algunas propiedades, tales como que el resultado sea un número índice sencillo y que en el mismo se reúna gran cantidad de información. En función de cual de esos criterios prevalezca nos llevará a dos categorías de índices compuestos distintas. Los que podríamos definir como índices compuestos no ponderados, en los que prevalece el criterio de la sencillez frente al de la información. El segundo grupo sería el de índices compuestos ponderados, donde se prima especialmente la información frente a la sencillez.

Dentro de la primera categoría, el más sencillo es el que define el índice compuesto como la media aritmética simple de los índices simples. Al mismo se le conoce como Indice de Sauerbeck y viene dado por:

$$P_S = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}}}{N} x100 \qquad (6.4)$$

$$Q_S = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}}}{N} x100 \qquad (6.5)$$

para precios y cantidades, respectivamente.

Frente a este procedimiento de obtener un índice compuesto no ponderado se podría haber utilizado el que se conoce como el de la media agregativa simple, o de Bradstreet-Dutot. Este consiste en sumar, cuando se trata de un índice de precios, los precios de todos los bienes para un periodo y obtener la media de esos precios. Con la serie resultante se obtendría un índice simple que es, de hecho, compuesto, pues en el mismo se han reunido los precios de más de un bien. Este procedimiento tiene el inconveniente, frente al anterior, de que suma inicialmente magnitudes que puede que no sean homogéneas, lo que lleva a que el índice resultante pierda significado.

Estos índices vienen dados por:

$$P_{BD} = \frac{\sum_{i=1}^{N} \frac{p_{it}}{N}}{\sum_{i=1}^{N} \frac{p_{i0}}{N}} x100 = \frac{\sum_{i=1}^{N} p_{it}}{\sum_{i=1}^{N} p_{i0}} x100 \qquad (6.6)$$

$$Q_{BD} = \frac{\sum_{i=1}^{N} \frac{q_{it}}{N}}{\sum_{i=1}^{N} \frac{q_{i0}}{N}} x100 = \frac{\sum_{i=1}^{N} q_{it}}{\sum_{i=1}^{N} q_{i0}} x100 \qquad (6.7)$$

Ejemplo 2. Los registros de una empresa dedicada a la producción de acero, relativos a sus principales inputs productivos, son los que se recogen en la Tabla 3.

Tabla 3.

	Hierro		Carbón		Electricidad	
	Precio (ptas./k.)	Cantidad (Tm.)	Precio (ptas./k.)	Cantidad (Tm.)	Precio (ptas./kwh)	Cantidad (kwh.)
1995	80	300	25	500	10	300000
1996	84	285	26	485	10,5	295000
1997	87	315	26	550	11	305000
1998	89	320	28	600	12	320000

A partir de esa información obtenga los índices de precios y de cantidades compuestos.

Para obtener los índices por el procedimiento de la media aritmética simple es necesario calcular previamente lo índices simples. Estos son los que se recogen en la Tabla 4.

Tabla 4.

	Indices Simples de Precios (Base 1995=100)			Indices Simples de Cantidad (Base 1995=100)		
	Hierro	Carbón	Electricidad	Hierro	Carbón	Electricidad
1995	100,0	100,0	100,0	100,0	100,0	100, 0
1996	105,0	104,0	105,0	95,0	97,0	98,3
1997	108,8	104,0	110,0	105,0	110,0	101,7
1998	111,3	112,0	120,0	106,7	120,0	106,7

A partir de ellos se obtienen los índices compuestos, de precios y cantidades, por el método de la media aritmética simple aplicando las expresiones (6.4) y (6.5) y que son los que aparecen en la Tabla 5.

Tabla 5. Indices compuestos no ponderados. (Media aritmética simple)

(Base 1995=100)

	P_S	Q_S
1995	100,0	100,0
1996	104,7	96,8
1997	107,6	105,6
1998	114,4	111,1

Estos índices se han obtenido de la forma siguiente:

$$P_{S\,95}^{\,95} = \frac{100,0 + 100,0 + 100,0}{3} = 100,0 \qquad P_{S\,95}^{\,96} = \frac{105,0 + 104,0 + 105,0}{3} = 104,7$$

$$P_{S\,95}^{\,97} = \frac{108,8 + 104,0 + 110,0}{3} = 107,6 \qquad P_{S\,95}^{\,98} = \frac{111,3 + 112,0 + 120,0}{3} = 114,4$$

$$Q_{S\,95}^{\,95} = \frac{100,0 + 100,0 + 100,0}{3} = 100,0 \qquad Q_{S\,95}^{\,96} = \frac{95,0 + 97,0 + 98,3}{3} = 96,8$$

$$Q_{S\,95}^{\,97} = \frac{105,0 + 110,0 + 101,7}{3} = 105,6 \qquad Q_{S\,95}^{\,98} = \frac{106,7 + 120,0 + 106,7}{3} = 111,1$$

A su vez, estos índices de precios y cantidades obtenidos por el procedimiento de la media agregativa simple son los que se recogen en la Tabla 6 y que se han obtenido como se indica a continuación:

$$P_{BD95}^{\,95} = \frac{80 + 25 + 10}{80 + 25 + 10} x100 = 100,0 \qquad P_{BD95}^{\,96} = \frac{84 + 26 + 10,5}{80 + 25 + 10} x100 = 104,8$$

$$P_{BD95}^{\,97} = \frac{87 + 26 + 11}{80 + 25 + 10} x100 = 107,8 \qquad P_{BD95}^{\,98} = \frac{89 + 28 + 12}{80 + 25 + 10} x100 = 112,2$$

$$Q_{BD\,95}^{\,95} = \frac{300 + 500 + 300000}{300 + 500 + 300000} x100 = 100,0 \qquad Q_{BD95}^{\,96} = \frac{285 + 485 + 295000}{300 + 500 + 300000} x100 = 98,3$$

$$Q_{BD95}^{\,97} = \frac{315 + 550 + 305000}{300 + 500 + 300000} x100 = 101,7 \qquad Q_{BD\,95}^{\,98} = \frac{320 + 600 + 320000}{300 + 500 + 300000} x100 = 106,7$$

Tabla 6. Indices compuestos no ponderado. (Media agregativa simple)
(Base 1995=100)

	P_{BD}	Q_{BD}
1995	100,0	100,0
1996	104,8	98,3
1997	107,8	101,7
1998	112,2	106,7

Como puede observarse, aunque las diferencias sean pequeñas, los distintos índices cambian de valores según el procedimiento utilizado para agregar la información primaria.

Ninguno de estos dos procedimientos tiene en cuenta el peso relativo de cada uno de los inputs a la hora de obtener el índice. Es decir, se calculan sin ponderar los distintos bienes o productos que se están considerando. Además, el método de la media agregativa simple presenta un inconveniente añadido, pues agrega magnitudes que pueden ser muy heterogéneas, como en el ejemplo que estamos tratando.

Los procedimientos señalados en el párrafo anterior se basan en el uso de la media aritmética. En realidad esos índices compuestos se pueden elaborar a partir del promedio que se considere más oportuno, lo que nos da una idea de los distintos procedimientos que se pueden utilizar para construir un índice complejo o compuesto.

5.4. Índices compuestos o complejos ponderados.

A continuación hablaremos de los algunos métodos para obtener índices compuestos ponderados. A diferencia de los métodos anteriores, en este caso se trata de promediar la información inicial haciendo uso de ciertas ponderaciones. Estas deben reflejar la importancia de los precios y las cantidades de cada uno de los bienes que entran en la definición del índice compuesto. Para ello sería buena idea retomar el concepto de valor que se dio en 5.2

Como señalamos antes, el valor de un bien se define como el producto del precio del mismo por su cantidad, de forma que si tenemos una serie de precios p_{it} y de cantidades q_{it} para un bien i determinado, entonces la serie de valores para ese bien será:

$$v_{i0} = p_{i0} q_{i0}$$
$$v_{i1} = p_{i1} q_{i1}$$
$$\vdots$$
$$v_{it} = p_{it} q_{it}$$
$$\vdots$$

Esta serie temporal depende de dos variables, el precio y la cantidad. Basta con que cambie una de ellas para que el valor cambie también. Así, si a lo largo del tiempo las cantidades permanecen fijas, las variaciones en el valor de ese bien se deberán solo y exclusivamente a las variaciones experimentadas en el precio. Igual podríamos argumentar si el precio permanece fijo. Con este tipo de argumentación o planteamiento se llegaría a las siguientesseries:

q fijo	p fijo	p y q variables
$p_{i0}q_{i0}$	$p_{i0}q_{i0}$	$p_{i0}q_{i0}$
$p_{i1}q_{i0}$	$p_{i0}q_{i1}$	$p_{i1}q_{i1}$
.	.	.
.	.	.
.	.	.
$p_{it}q_{i0}$	$p_{i0}q_{it}$	$p_{it}q_{it}$

En la primera serie las cantidades permanece fijas, en la segunda el precio no varía y en la tercera varían precios y cantidades. Pero las tres series expresan valores. Es decir,

vienen expresadas en las mismas unidades de medidas, por lo que son fácilmente agregables o sumables. Así, si tuviéramos N bienes distintos, entonces la suma de los valores de los mismos sería:

q fijo	*p* fijo	*p* y *q* variables
$\Sigma_i p_{i0}q_{i0}$	$\Sigma_i p_{i0}q_{i0}$	$\Sigma_i p_{i0}q_{i0}$
$\Sigma_i p_{i1}q_{i0}$	$\Sigma_i p_{i0}q_{i1}$	$\Sigma_i p_{i1}q_{i1}$
.	.	.
.	.	.
.	.	.
$\Sigma_i p_{it}q_{i0}$	$\Sigma_i p_{i0}q_{it}$	$\Sigma_i p_{it}q_{it}$

Al igual que antes, ahora, la diferencia entre las tres series radica en la componente que cambia, pues por lo demás, tan valores son las unas como las otras. En el primer caso para un bien y en el segundo para N bienes. A partir de estas tres series se pueden obtener índices “simples” aunque de hecho serán índices complejos. Así, si el año base es *t = 0*, entonces los índices correspondientes al año t vendrán dados por:

$$IP_0^t = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 \qquad (5.8)$$

$$IQ_0^t = \frac{\sum_{i=1}^{N} p_{i0} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 \qquad (5.9)$$

$$IV_0^t = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 \qquad (5.10)$$

De las tres expresiones dadas, la (5.10) es un índice de valor en sentido estricto. La (5.9) es también un índice de valor, pero las variaciones de éste vienen motivadas por las variaciones en cantidades, por lo que el mismo puede interpretarse con un índice de cantidades o cuántico. De forma similar puede argumentarse para el primer caso, pero ahora en términos de precios.

Estos índices de precios, cantidades y de valor son índices complejos, pues tienen en cuenta N bienes. Los mismos se han obtenido sumando o agregando los valores de cada bien, por lo que se le conoce como índices agregativos, aunque ahora esa agregación se ha realizado con ponderaciones. En el caso del índice de precios, las ponderaciones son las cantidades , mientras que para el cuántico, las ponderaciones son los precios. Además la agregación en este caso no entraña ningún problema, pues todas las series vienen expresadas en las mismas unidades de medida.

Esta forma de obtener índices da una salida bastante general al problema del cálculo de índices complejos. Sin embargo hay una cuestión que no está cerrada totalmente. Para el caso de los índices de precios hemos supuesto que las cantidades permanecen fijas, pero nos podemos preguntar cuáles son las que deben permanecer fijas. Se podría tomar como valor el correspondiente al año base, como hemos hecho hasta ahora. Pero esto solo es una de las muchas posibles soluciones, pues ese valor fijo puede ser el de cualquier periodo de los considerados. De igual forma se podría razonar para el índice de cantidades en relación a qué precio se deja fijo. De los distintos valores de *p* y *q* que se pueden tomar como fijos, en la práctica se ha optado por dos soluciones. La primera consiste en tomar como constantes el precio o la cantidad del tiempo elegido como base, es decir, la opción presentada hasta ahora. A los índices que se elaboran de esta forma se les conoce como índices de Laspeyres. La segunda consiste en tomar como constante

el precio o la cantidad correspondient e al tiempo para el cual se va a calcular el índice (o sea, el periodo *t*); a los índices obtenidos de esta forma se les conoce como *índices de Paasche*. De acuerdo con estos criterios se llega a los siguientes índices de precios y cantidades:

$$P_L = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 \qquad (5.11) \qquad Q_L = \frac{\sum_{i=1}^{N} p_{i0} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 \qquad (5.12)$$

$$P_P = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}} x100 \qquad (5.13) \qquad Q_P = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 \qquad (5.14)$$

Una tercera solución consiste en combinar las anteriores mediante una media geométrica de las mismas. A estos índices se les conoce como índices de Fisher y vendrían dados por:

$$P_F = \sqrt{P_L P_P} \qquad (5.15) \qquad Q_F = \sqrt{Q_L Q_P} \qquad (5.16)$$

Estos índices de precios y cantidades se han obtenido como resultado de una agregación de magnitudes ponderadas y como tal puede decirse de los mismos que son índices obtenidos como una media agregativa ponderada. Pero vamos a ver a continuación que también pueden contemplarse como una media aritmética ponderadas de índices simples.

$$P_L = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} p_{i0} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} w_i}{\sum_{i=1}^{N} w_i} x100 = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} x100 \qquad (5.17)$$

$$Q_L = \frac{\sum_{i=1}^{N} p_{i0} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} p_{i0} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} w_i}{\sum_{i=1}^{N} w_i} x100 = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} x100 \qquad (5.18)$$

$$P_P = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}} x100 = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} p_{i0} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}} x100 = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} w_i}{\sum_{i=1}^{N} w_i} x100 = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} x100 \qquad (5.19)$$

$$Q_P = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 = \frac{\sum_{i=1}^{N} \frac{q_{it}}{q_{i0}} w_i}{\sum_{i=1}^{N} w_i} x100 = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} x100 \qquad (5.20)$$

Como puede observarse, el índice de precios es la media aritmética de los índices simples ($I_i = \frac{p_{it}}{p_{i0}}$) ponderada por los valores ($w_i = p_{i0} q_{i0}$, o bien $w_i = p_{i0} q_{it}$), según se trate de un índice de Laspeyres o de Paasche, respectivamente. De forma similar ocurre con los índices de cantidad.

Estas relaciones muestran que puede resultar interesante calcular primero los índices simples o elementales de todos los bienes y luego calcular su media aritmética ponderada, lo permite realizar un estudio por separado y después conjuntamente.

De todos los índices compuestos que se han definido, los de Laspeyres son los que requieren menos información, pues las ponderaciones son siempre fijas, las del periodo base, mientras que para los de Paasche las mismas varían en cada periodo. Pero esa ventaja que presentan los primeros puede llegar a ser un inconveniente, pues, con el transcurso del tiempo, esas ponderaciones iniciales pueden llegar a quedarse obsoletas, lo que obliga a realizar una renovación de las mismas.

Para concluir este apartado, debemos señalar que los índices definidos debieran satisfacer algunas propiedades de entre las que se van indicar solo dos. La de compatibilidad y la de proporcionalidad. La primera consistente en que si un precio por una cantidad da un valor, también debiera ocurrir con los índices. Sin embargo no ocurre siempre, pues es fácil comprobar que:

$$P_L Q_L \neq V; \qquad P_P Q_P \neq V. \quad (5.21)$$

En cambio si se cumple que:

$$P_F Q_F = V; \qquad P_L Q_P = V; \qquad P_P Q_L = V \qquad (5.22)$$

La propiedad de proporcionalidad establece que si en el periodo corriente todos los precios sufren una variación proporcional, el índice debe quedar afectado por esa variación. Esta propiedad la cumplen todos los índices definidos en este capítulo, tanto sin son simples como ponderados.

Ejemplo 3. A partir de los datos de la Tabla 3, obtener los índices de precios de Laspeyres, Paasche y Fisher.

Para calcular los índices de precios de Laspeyres y de Paasche se puede hacer uso de las expresiones (5.11) y (5.13) o bien (5.17) y (5.19). En este ejemplo se utilizaran las dos últimas, pues, como ya se ha indicado, procediendo de esta forma se tienen también los

índices simples para cada uno de los bienes o productos que entran en la construcción del índice complejo.

Empezaremos calculando las ponderaciones necesarias para la construcción de estos índices tomando como año base 1995. Las mismas se recogen en la tabla 7:

Tabla 7. Ponderaciones.

	Hierro (p_0q_t)	Carbón (p_0q_t)	Electricidad (p_0q_t)	Total	Valor (p_tq_t)
1995	24000000	12500000	3000000	39500000	39500000
1996	22800000	12125000	2950000	37875000	39647500
1997	25200000	13750000	3050000	42000000	45060000
1998	25600000	15000000	3200000	43800000	49120000

donde:

$$39500000 = \sum_{i=1}^{3} p_{i95}q_{i95}$$
$$37875000 = \sum_{i=1}^{3} p_{i95}q_{i96}$$
$$42000000 = \sum_{i=1}^{3} p_{i95}q_{i\,97}$$
$$43800000 = \sum_{i=1}^{3} p_{i95}q_{i98}$$

A continuación se obtendrán los índices simples de precios de los tres bienes, que son los que aparecen en la Tabla 4, y que volvemos a reproducir en la Tabla 8:

Tabla 8. Indices simples de precios. (Base 1995=100)

	Indices Simples de Precios		
	Hierro	Carbón	Electricidad
1995	100,0	100,0	100,0
1996	105,0	104,0	105,0
1997	108,8	104,0	110,0
1998	111,3	112,0	120,0

A partir de los índices recogidos en la Tabla 8 y de las ponderaciones dadas en la Tabla 7 se obtienen los índices compuestos de precios de Laspeyres, Paasche y Fisher de la Tabla 9. Como puede observarse las diferencias entre unos y otros son muy pequeñas, pese a que las ponderaciones sean distintas de un caso a otro. Esto se debe, fundamentalmente, a que el horizonte temporal con el que se trabaja es muy corto, solo cuatro años, por lo que la estructura de precios y cantidades no ha cambiado de forma significativa como para alterar los valores de los índices calculados. Las diferencias entre unos y otros se dan cuando se trabaja con series largas, pues en esos casos si que es posible que cambien las relaciones precios cantidades iniciales entre los distintos bienes.

Tabla 9. Indices compuestos de precios (Base 1995=100)

	Laspeyres	Paasche	Fisher
1995	100,00	100,00	100,00
1996	104,68	104,68	104,68
1997	107,34	107,29	107,31
1998	112,15	112,15	112,15

A titulo de ilustración vamos a indicar los pasos seguidos para obtener los índices de 1998 con base en 1995.

$$P_{L\,95}^{98} == \frac{\sum_{i=1}^{N} \frac{p_{i98}}{p_{i95}} p_{i95} q_{i95}}{\sum_{i=1}^{N} p_{i95} q_{i95}} x100 = \frac{(111{,}3)(24000000) + (112{,}0)(12500000) + (120)(3000000)}{39500000} = 112{,}15$$

$$P_{P95}^{98} == \frac{\sum_{i=1}^{N} \frac{p_{i\,98}}{p_{i\,95}} p_{i\,95} q_{i98}}{\sum_{i=1}^{N} p_{i\,95} q_{i98}} x100 = \frac{(111{,}3)(25600000) + (112{,}0)(15000000) + (120)(3200000)}{43800000} = 112{,}15$$

$$P_{F95}^{98} = \sqrt{P_{L\,95}^{98} P_{P\,95}^{98}} = \sqrt{(112.15)(112.15)} = 112.15$$

Si se comparan estos resultados con los de la Tabla 8 se observa como el tercer bien que entra en juego, la electricidad en nuestro caso, tiene poca incidencia, pues sus ponderaciones son más pequeñas que las correspondientes a lo otros.

Una vez construidos los índices de precios habría que calcular los de cantidades y los de valor. El procedimiento a seguir para los de cantidades es similar al utilizado para los índices de precios, por lo que se podrían obtener de esa forma. En su lugar ser calcularán haciendo uso de las relaciones dadas en (5.22). Para ello es necesario obtener en primer lugar los índices de valor. Estos se obtienen como un índice simple de la última columna de la Tabla 7. El resultado de estas operaciones se recogen en la Tabla 10.

Tabla 10. Indices de valor y de cantidades (Base 1995=100)

		Cantidades		
	Valor	Laspeyres	Paasche	Fisher
1995	100,00	100,00	100,00	100,00
1996	100,37	95,89	95,88	95,88
1997	114,08	106,33	106,27	106,30
1998	124,35	110,89	110,88	110,88

5.5 Problemática en la construcción de índices complejos.

En los ejemplos de números índices que se han dado en el epígrafe anterior se han obviado un conjunto de problemas, la mayoría de naturaleza práctica, que en el contexto de un ejemplo sencillo no tenía sentido plantear. Sin embargo, hasta llegar a obtener los datos primarios, que nos permiten aplicar ciertas fórmulas para la construcción de los números índices, hay que resolver algunas cuestiones que pueden condicionar de forma decisiva la calidad e incluso la validez de los resultados. Una de ellas tiene que ver con lo que conoce como cobertura del índice. Por tal se entiende el conjunto de variables seleccionadas para la elaboración del índice.

Como se ha indicado con anterioridad, con un índice complejo lo que se pretende es medir la evolución en el tiempo de una cierta magnitud (precios, cantidades, etc.) para un sector o área de actividad concreta. Pero dentro de ese sector se pueden producir, vender o comprar un elevado número de bienes o servicios distintos y cuyos precios o cantidades difícilmente podrían ser todas observadas. Ante estas circunstancias hay que proceder a seleccionar aquel conjunto que represente adecuadamente al total. Es decir hay que procurar que con ese subconjunto seleccionado se obtenga una buena cobertura.

Una vez que se ha fijado la cobertura del índice se pueden suscitar otro conjunto de cuestiones tales como: a) agrupar esas variables en categorías homogéneas que permitan analizar el sector de una forma gradual; b) fijar un periodo base que no presente anomalías para que éstas no se transmitan a todos los valores del índice; c) determinación de las fórmulas de calculo y sus correspondientes ponderaciones acordes, tanto a la información disponible en el presente como en el futuro, como a la idiosincrasia de la parcela de actividad que se pretende medir o estudiar.

Todas estas serían cuestiones previas a la elaboración del índice. Pero la problemática no termina en ese instante. Con posterioridad pueden surgir otros problemas relacionados todos, de forma más o menos directa, con la antigüedad u obsolescencia del índice. Imaginemos que se trata de un índice de precios de Laspeyres. El mismo, como se sabe, utiliza unas ponderaciones fijas que son las del año base. Pero pasado el tiempo esas ponderaciones puede que no reflejen la realidad actual, lo que nos obliga a cambiarlas y, por tanto a cambiar o renovar la base del índice. En realidad lo que se está realizando es una renovación del propio índice, de forma que se tendrían dos índices con bases distintas y que habría que unir. A esta operación se le conoce como enlace de índices.

Para que estas ideas queden más claras haremos uso de un ejemplo.

Ejemplo 4*. En la Tabla 11 se recogen los Indices de Precios Industriales para España, con base 1974 y 1990, para los meses de diciembre de cada año. A partir de esas series obténgase una serie única, tanto en base 1974 como para 1990.*

Tabla 11. Indices de Precios Industriales en España a diciembre de cada año

	Base 1974	*Base 1990*
1985	*424,30*	
1986	*419,53*	
1987	*429,70*	
1988	*444,49*	
1989	*460,67*	
1990	*471,12*	*102,0*
1991		*102,6*
1992		*104,2*
1993		*107,7*
1994		*113,3*
1995		*118,3*

Fuente: Servidor web del INE.

Para cambiar la base de un índice basta con determinar la relación existente entre los valores del mismo para el único periodo en el que se dispone de información en las dos bases. En nuestro caso ese periodo es diciembre de 1990. Si lo que se pretende es enlazar las series tomado como base 1974, entonces la relación buscada o coeficiente de enlace vendrá dada por :

$$\frac{I_{74}^{90}}{I_{90}^{90}} = \frac{471,12}{102,0} = 4,6188$$

En cambio, si lo que se quiere es enlazar tomando 1990 como base, entonces ese coeficiente será:

$$\frac{I_{90}^{90}}{I_{74}^{90}} = \frac{102,0}{471,12} = 0,2165$$

Una vez que se han calculado estos coeficientes, los mismos se les aplican a las series originales y se obtienen las series enlazadas que parecen en la Tabla 12.

Como puede apreciarse, la mecánica conducente al enlace de series de números índices es bastante simple. Pero hay que señalar que tener una sola serie obtenida por este procedimiento, aunque presenta notables ventajas, tiene también algunas limitaciones que deben señalarse. De todas ellas la más importante es que la serie no es homogénea,

pues la cobertura del índice en las dos bases es distinta y, como ocurre en este caso concreto, las ponderaciones y la metodología utilizada para su elaboración también lo son. Todo ello lleva a que el resultado de esta operación mecánica que se ha realizado haya que usarlo con precaución.

Tabla 12. Series de Indices de Precios Industriales enlazadas.

	Base 1974	Base 1990	Base 1990 (Diciembre 1995=100)
1985	424,30	424,30x0,2165 = 91,9	91,9x0,8453 = 77,7
1986	419,53	419,53x0,2165 = 90,8	90,8x0,8453 = 76,8
1987	429,70	429,70x0,2165 = 93,0	93,0x0,8453 = 78,6
1988	444,49	444,49x0,2165 = 96,2	96,2x0,8453 = 81,3
1989	460,67	460,64x0,2165 = 99,7	99,7x0,8453 = 84,3
1990	102,0x4,6188 = 471,12	102,0	102,0x0,8453 = 86,2
1991	102,6x4,6188 = 473,89	102,6	102,6x0,8453 = 86,7
1992	104,2x4,6188 = 481,28	104,2	104,2x0,8453 = 88,1
1993	107,7x4,6188 = 497,45	107,7	107,7x0,8453 = 91,0
1994	113,3x4,6188 = 523,31	113,3	113,3x0,8453 = 95,8
1995	118,3x4,6188 = 546,41	118,3	118,3x0,8453 = 100,0

Una operación similar al enlace de series es el cambio de base para una serie concreta. Así, y para este ejemplo de los Precios Industriales, podría plantearse que la serie con base 1990 tomara el valor cien en diciembre de 1995. Para ello haría falta buscar un coeficiente que permita realizar esa transformación que es el cambio de base. Ese coeficiente es similar al usado para el enlace de series. En nuestro caso sería:

$$\frac{100}{I_{90}^{95}} = \frac{100,0}{118,3} = 0,8453$$

El resultado del cambio aparece en la última columna de la Tabla 12.

5.6 Índice de Precios de Consumo (IPC) y otros Indices: Definición y aplicaciones.

El instrumento estadístico que se viene exponiendo en este capítulo tiene una aplicabilidad tan amplia que sería casi inabarcable la enumeración y análisis de todos y cada uno de los índices que se elaboran, aunque solo fuera dentro del ámbito de la estadística oficial. Por esa razón nos limitaremos a señalar solo aquellos que por su uso más frecuente son los más conocidos. De entre ellos, y por su repercusión social y económica, el Indice de Precios de Consumo (IPC) es, con diferencia, el más conocido. Otros índices que recientemente han adquirido notable popularidad son los bursátiles. También son de interés los siguientes:

- Índices Implícitos de precios
- Indice de Producción Industrial
- Indices de Precios Industriales

5.6.1 Índice de Precios de Consumo.

El Índice de Precios de Consumo (IPC) que se calcula y publica mensualmente. Tiene como objetivo medir la evolución del nivel de precios de los bienes y servicios de consumo adquiridos por los hogares residentes en España.

Se trata de un indicador muy dinámico y que en su dilatada historia ha cambiado tanto en su definición como en muchos de los aspectos técnicos relacionados con el mismo. Inicialmente se le conocía como Indice de Coste de la Vida, y con esta denominación duró hasta 1976, momento a partir del cual cambió a como se le conoce actualmente.

Pero los cambios más relevantes de este indicador son los que hacen referencia a cuestiones relacionadas con su elaboración. Este indicador es un índice complejo que hace uso de la fórmula de Laspeyres. Para el caso del IPC, el índice de precios de Laspeyres requiere, para su cómputo, información relativa a los precios del conjunto de bienes y servicios que consume la población de referencia (en nuestro caso la residente en hogares) así como información referida a las ponderaciones de esos bienes y servicios. Como se recordará, ese índice viene dado por:

$$IPC(t) = P_L^t = I_t = \left(\sum_{i=1}^{N} w_i I_{it} \right) x100 = \left(\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} w_i \right) x100 \qquad (5.23)$$

donde: w_i es la ponderación del artículo *i-ésimo* y representa la proporción del gasto efectuado en ese artículo respecto al gasto total efectuado por los hogares, p_{it} es el precio de ese artículo en el periodo *t* y p_{i0} el del periodo de referencia o año base.

Pero antes de indagar por los precios de esos bienes y servicios consumidos es necesario saber cuales son los mismos. Es decir, hay que fijar la cobertura del índice. A esa cobertura se le conoce en este caso como la cesta de la compra, en el sentido de que constituye el conjunto de bienes y servicios que compran los residentes en viviendas familiares para su consumo. Ahora bien, la estructura de esta cesta de la compra no es invariable, pues los hábitos de consumo de las personas cambian con el transcurso del tiempo. Estos cambios obligan a dejar fuera ciertos productos que dejan de consumirse y a introducir otros nuevos que aparecen en el mercado. Pero no basta con cambiar unos productos por otros. También hay que calibrar la importancia relativa de cada uno de ellos (sus ponderaciones) y adaptarla a cada momento.

Toda esta problemática se ha resuelto recurriendo a una encuesta que se realizaba de forma periódica cada ocho o diez años en España. Se trataba de la Encuesta Básica de Presupuestos Familiares (EBPF). A través de esta encuesta se determinaba la cesta de la compra (conjunto de bienes y servicios consumidos) de una familia media, tanto para todo el estado, como por comunidades y por provincias y la estructura de gasto en bienes y servicios de las familias a las que se dirige la encuesta (cantidades monetarias gastadas). Así pues, una vez que se tenía definida la cobertura del índice y las ponderaciones del mismo para un año concreto (año de referencia del índice o año base, que coincide con el periodo al que se refiere la encuesta), lo único que falta para su elaboración era conoc er los precios de los bienes y servicios que integran la cesta de la compra. Esos precios se observaban de forma continua todos los meses, lo que permite elaborar ese índice con periodicidad mensual.

5.6.1.1. Nuevo Sistema de Indices de Precios de Consumo (Base 2001).

Como se ha indicado, el periodo de referencia del IPC se corresponde con el año en el que se realiza la encuesta y las ponderaciones, que son fijas, son las que se deducían de

la EBPF. Pero, como también se ha indicado, los hábitos de los consumidores cambian con el tiempo, bien sea porque varían los gustos o las modas, su capacidad de compra, o porque han aparecido nuevos productos en el mercado hacia los que se desvía el gasto. Todo esto lleva a que las ponderaciones y la propia cesta de la compra llega un momento que no reflejan el fenómeno que se quiere medir y el IPC empieza a perder vigor. Ello ha obligado a renovar la EBPF de manera repetida, siendo la última renovación la de 1990-91 (de abril del 90 a marzo del 91), lo que permitió elaborar el IPC con base 1992. Este periodo de referencia se ha mantenido hasta diciembre de 2000. Sin embargo a partir de enero de 2001 el INE ha implantando un Nuevo Sistema de Indices de Precios de Consumo, donde la principal novedad radica en que tanto la cesta de la compra como las ponderaciones del índice se obtendrán a partir de la Encuesta Continua de Presupuestos Familiares (ECPF) que tiene periodicidad trimestral. Esta novedad permitirá una mayor dinamicidad en el índice, pues será posible actualizar las ponderaciones en periodos cortos de tiempo así como adaptar la cesta a la realidad de cada momento. Además el INE pretende que el nuevo sistema sea técnicamente más moderno, de forma que permita la inclusión inmediata de mejoras en la metodología que ofrezcan los distintos foros académicos y de organismos nacionales e internacionales.

Los cambios más relevantes de este nuevo sistema de IPC son los siguientes:

a) Se han actualizado las ponderaciones.

b) La clasificación funcional de los artículos se hace en doce grupos, a diferencia del sistema anterior que solo contemplaba ocho grupos.

c) Cambio en el tratamiento de los artículos de recogida centralizada o artículos de tarifas. Para estos artículos se modificará el cálculo de las ponderaciones de las distintas modalidades que intervienen en el precio final de este tipo de artículos, ponderándose, a partir de enero de 2001, según el gasto en lugar del numero de unidades.

A estos cambios hay que añadir otros de carácter más metodológico que hacen referencia a:

a) Selección de la muestra (selección de: municipios, zonas comerciales y establecimientos y determinación del número de observaciones (unos 180.000 precios mensuales)).

Tabla 13. Grupos del IPC y ponderaciones de los mismos. Base 1992

Base 1992		
Grupo	Denominación	Ponderación
1	Alimentos bebidas y tabaco	293,607
2	Vestido y calzado	114,794
3	Vivienda	122,803
4	Menaje y servicios del hogar	66,840
5	Medicina y conservación de la salud	31,260
6	Transporte y comunicaciones	165,419
7	Esparcimiento, enseñanza y cultura	72,671
8	Otros bienes y servicios	152,606
Total		1000,000

Fuente: Página web INE

Tabla 14. Grupos del IPC y ponderaciones de los mismos. Base 2001

Base 2001		
Grupo	Denominación	Ponderación
1	Alimentos y bebidas no alcohólicas	215,051
2	Bebidas alcohólicas y tabaco	32,182
3	Vestido y calzado	100,384
4	Vivienda	114,613
5	Menaje	63,574
6	Medicina	28,718
7	Transporte	157,331
8	Comunicaciones	25,374
9	Ocio y cultura	65,238
10	Enseñanza	16,878
11	Hoteles, cafés y restaurantes	113,259
12	Otros	67,398
Total		1000,000

Fuente: Página web INE

b) Determinación de la cesta de la compra (selección de artículos (484 articulos) y ponderaciones).

c) Método de cálculo. Con anterioridad a 2001, como ya se ha indicado, el IPC en España era un índice tipo Laspeyres con base fija, al igual que en otros muchos países de la Unión Europea. La ventaja fundamental de un índice de este tipo es que permite la comparabilidad de una misma estructura de artículos y ponderaciones a lo largo del tiempo que esté en vigor el Sistema; sin embargo, tiene un inconveniente y es que la estructura de ponderaciones pierde vigencia a medida que pasa el tiempo y evolucionan las pautas de consumo de los consumidores.

El nuevo Sistema utiliza un la fórmula de Laspeyres encadenado, que consiste en referir los precios del periodo corriente a los precios del año inmediatamente anterior. Además, con una periodicidad que no superará los dos años, se actualizarán las ponderaciones de las parcelas con información proveniente de la ECPF.

Básicamente, el proceso de cálculo es el mismo que el de un Laspeyres: se calculan medias ponderadas de los índices de los artículos que componen cada una de las agregaciones funcionales para las cuales se obtienen índices, y se compraran con los calculados el mes anterior. En este caso las ponderaciones utilizadas no permanecen fijas durante el período de vigencia del sistema.

Por tanto, la formulación es la siguiente:

$$ {}_{t-1}I_G^{mt} = \sum_{i=1}^{N} \left({}_{t-1}w_i\right)\left({}_{t-1}I_i^{mt}\right) \qquad (6.24) $$

donde:

${}_{t-1}I_G^{mt}$: es el índice general en el mes *m* del año *t* referido al mismo mes del año *t-1*.

$_{t-1}w_i$: es la ponderación del componente i referida al año t-1.

$_{t-1}I_i^{mt}$: es el índice del componente i en el mes *m* del año *t* referido al mismo mes del año *t-1*.

5.6.1.1.1 Características destacables del nuevo sistema Indices de Precios de Consumo

a) *Período base.* El período base es aquél para el que la media aritmética de los índices mensuales se hará igual a 100. El año 2001 será el periodo base del nuevo Sistema, esto quiere decir que todos los índices que se calculen estarán referidos a este año.

b) *Período de referencia de la estructura.* Es el período al que están referidas las ponderaciones que sirven de estructura del Sistema; dado que éstas se obtienen de la Encuesta Continua de Presupuestos Familiares (ECPF), el período de referencia del IPC es el período durante el cual se desarrolla esta encuesta.

El actual cambio de Sistema se ha realizado con la información proveniente de la Encuesta Continua de Presupuestos Familiares (ECPF), que proporciona la información básica sobre gastos de las familias en bienes y servicios de consumo. El nuevo sistema de índices de base 2001 utiliza la ECPF que se ha llevado a cabo entre el 2º trimestre de 1999 y el 1º de 2000; no obstante, las ponderaciones se han actualizado al año 2001 de forma que el periodo de referencia de la estructura de ponderaciones y el período base coincidan.

c) *Cambios de calidad.* El tratamiento de los cambios de calidad es uno de los temas que más afectan a cualquier índice de precios.

Un cambio de calidad ocurre cuando cambia alguna de las características de la variedad para la que se recoge el precio y se considera que este cambio implica un cambio en la utilidad que le reporta al consumidor.

Para la correcta medición de la evolución de los precios es preciso estimar en qué medida la variación observada del precio es debida al cambio en la calidad del producto y qué parte de esta variación es achacable al precio, independientemente de su calidad.

Los métodos más utilizados en el IPC son la *consulta a expertos* , que consiste en solicitar a los propios fabricantes o vendedores la información para poder estimar el cambio; *los precios de las opciones* , que analiza los elementos componentes del antiguo producto y del nuevo para establecer el coste de las diferencias entre ambos; y *el precio de solapamiento*, basado en suponer que el valor de la diferencia de calidad entre el producto que desaparece y el nuevo es la diferencia de precio entre ellos en el periodo de solapamiento, es decir, en el periodo que estén en vigencia los precios de ambos.

d) *Inclusión de las ofertas y rebajas.* Uno de los cambios más importantes que se recoge en el nuevo Sistema, base 2001, es la inclusión de los precios rebajados.

El IPC, base 1992, no contempla la recogida de estos precios por lo que su inclusión en el nuevo Sistema ha dado lugar una ruptura en la serie de este indicador que no es posible solucionar con el método de los enlaces legales, utilizado cada vez que se lleva a cabo un cambio de base.

5.6.1.2. Índice de Precios de Consumo Armonizado. Unión Europea y España. (IPCA).

- El *Índice de Precios de Consumo Armonizado. Unión Europea y España. (IPCA)* es un indicador estadístico cuyo objetivo es proporcionar una medida común de la inflación que permita realizar comparaciones internacionales y examinar, así, el cumplimiento que en esta materia exige el Tratado de Maastricht para la entrada en la Unión Monetaria Europea.

- *Proceso de armonización del IPCA*.

  Este proceso consta de dos fases:

  La primera se ha desarrollado durante 1996. Establecía el cálculo de los Índices de Precios de Consumo Transitorio (IPCT) basados en el IPC de cada uno de los países miembros cuyos resultados se han venido publicando mensualmente.

La segunda contempla la construcción de los Índices de Precios de Consumo Armonizados, como resultado de homogeneizar los aspectos metodológicos más importantes de cada uno de los Índices de Precios de Consumo (IPC) para hacerlos comparables.

Durante el período de implantación transitoria, se han ido realizando las modificaciones y ajustes necesarios sobre los IPC nacionales hasta conseguir un índice con unas características esenciales comunes a todos los países. El primer índice de esta fase es el correspondiente a enero de 1997, que es el que se hace público el día 7 de marzo. Estos índices tendrán como período de referencia el año 1996.

La base legal del proceso de armonización de los IPC es el Reglamento del Consejo nº 2494/95 de 23 de octubre de 1995 que establece las directrices para la obtención de índices comparables, así como un calendario de obligado cumplimiento para todos los países de la Unión Europea.

- *Características técnicas del IPCA*

Los aspectos técnicos más significativos de este IPCA son los siguientes:

1. *Cobertura*. El IPCA de cada país cubre las parcelas que superan el uno por mil del total de gasto de la cesta de la compra nacional. En cada Estado miembro ha sido necesario realizar particulares ajustes para conseguir la comparabilidad deseada mediante determinadas inclusiones o exclusiones de partidas de consumo.

En este sentido han quedado excluidas del IPCA los *Servicios médicos* y la *Enseñanza reglada*. Además, la ponderación de algunas parcelas no se incluye totalmente. Tal es el caso de los *Seguros*, para los que sólo se consideran los gastos ligados a las primas netas, los *Automóviles*, de los cuales se elimina los gastos correspondientes a ventas entre consumidores, o los *Medicamentos y productos farmacéuticos*, que sólo incluyen los no subvencionados.

Como resultado de estas exclusiones, la ponderación total eliminada de la estructura del IPC español se sitúa en torno al cinco por ciento.

El IPCA está formado por doce grandes grupos. Para definir estos grupos se ha utilizado la clasificación de consumo COICOP (Classification of Individual Consumption by Purpose).

2. *Período común de referencia***.** El período de referencia para todos los IPCA es el año 1996, es decir, la media de los doce índices mensuales de este año se hace 100.

3. *Fórmula general*. Para obtener el IPCA se utiliza, como en el caso del IPC español, la fórmula de Laspeyres:

$$I = \sum_i I_i w_i \qquad (6.25)$$

donde el índice de cada artículo o agregado elemental, I_i, se obtiene como cociente de las medias aritméticas de sus precios.

Las ponderaciones w_i permanecerán fijas mes a mes, como corresponde a un índice de Laspeyres.

4. *Ponderaciones*. Las ponderaciones de cada componente del IPCA se han actualizado con referencia al año 1996, en función de la evolución de su índice respecto del índice general hasta 1996.

5. *Índice de precios de consumo europeo*. A partir de los IPCA de los quince países miembros EUROSTAT obtiene un Índice de Precios de Consumo de la Unión Europea, como media ponderada de los IPCA de dichos índices.

Si bien, los IPCA proporcionan la mejor base estadística para hacer comparaciones internacionales de inflación, y representan un considerable progreso en la armonización de las metodologías, todavía no se puede hablar de una completa armonización de los índices de precios de consumo. En este sentido se seguirán proponiendo acuerdos técnicos sobre distintos aspectos, entre los que se encuentran la ampliación de la cobertura, la homogeneización de los procedimientos de ponderaciones y el tratamiento metodológico de parcelas concretas.

Tabla 15. Ponderaciones de los doce grupos IPCA.

Base 2001		
Grupo	Denominación	Ponderación
1	Alimentos y bebidas no alcohólicas	27,5
2	Bebidas alcohólicas y tabaco	3,2
3	Vestido y calzado	11,4
4	Vivienda	11,2
5	Menaje	6,5
6	Medicina	0,8
7	Transporte	14,6
8	Comunicaciones	1,6
9	Ocio y cultura	6,9
10	Enseñanza	0,1
11	Hoteles, cafés y restaurantes	11,8
12	Otros	4,4
Total		100,0

Fuente: Página web INE

6.6.2. Índices bursátiles

Un Indice Bursátil es un instrumento estadístico que refleja el cambio en el tiempo de los precios de un conjunto de acciones de empresas que cotizan en bolsa. Se trata, pues, de un índice de precios, aunque ahora los "bienes" son más homogéneos de lo que son en el IPC.

Los distintos índices bursátiles que se calculan dependen de las ponderaciones que utilicen para su elaboración. Los más habituales son lo índices tipo valor dados por:

$$I = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{i0}} \qquad (6.26)$$

donde:

p_{it} es el precio o cotización de la acción correspondiente a la empresa i en el periodo t.

q_{it} es el número de acciones de la empresa i en el periodo t

p_{i0} es el precio o cotización de la acción correspondiente a la empresa i en el periodo de referencia

q_{i0} es el número de acciones de la empresa i en el periodo de referencia.

5.6.2.1. Índice General de la Bolsa de Madrid.

Pero aunque este sea un estándar muy utilizado, sin embargo no es como se obtiene el Indice General de la Bolsa de Madrid (IGBM). El IGBM es un índice de Laspeyres donde las ponderaciones se calculan a 31 de diciembre de cada año y se mantiene fijas para todo el año siguiente. Esas ponderaciones se obtienen a partir de la capitalización bursátil de las empresas, entendiendo por tal el resultado de multiplicar el número de acciones de la empresa por la cotización de las mismas a 31 de diciembre. Esta forma de calcular este índice supone que las ponderaciones se cambian todos los años.

El IGBM se puede obtener como un índice complejo ponderado teniendo en cuenta a todas las empresas que lo integran de forma individualizada o bien como un media ponderada de los índices sectoriales en los que se agrupan las empresas. Por cualquiera de las vías que se calcule se aplican siempre la formula de Laspeyres y los resultados son los mismos.

En las Tabla 16 se da el número de empresas que se tienen en cuenta para la el cálculo del IGBM, agrupadas por sectores, así como sus respectivas ponderaciones, todo ello para el año 2001.

Esas 116 empresas no son las únicas que cotizan en la Bolsa de Madrid, pero si son las más representativas en términos capitalización y volumen de contratación, además de otras variables.

Tabla 16. Número de empresas del IGBM y ponderaciones por sectores.

	Empresas	Ponderación
Bancos y financieras	12	31,31
Eléctricas	7	12,36
Alimentación	15	2,08
Construcción	10	3,18
Cartera e inversión	2	2,68
Metal mecánica	7	1,39
Petróleo y químicas	12	6,03
Comunicaciones	11	29,84
Otras ind. y servicios	13	5,39
Nuevas tecnologías	27	5,74
Total	**116**	**100**

Fuente: Página web de Bolsa de Madrid.

En la Tabla 17 se dan los datos de capitalización y ponderaciones para uno de los sectores, en concreto para el de Bancos y Financieras.

Tabla 17. Capitalización y ponderaciones para las empresas del Sector de Bancos y Financieras de la Bolsa de Madrid en 2001.

	Capitalización	Ponderación para 2001	
	Euros	En el grupo	En el IGBM
Banco Santander Central Hispano	51476554502	43,24	13,54
BBVA	50654254882	42,55	13,32
Banco Popular Español	8056417704	6,77	2,12
Bankinter	2709681666	2,28	0,71
Corporación MAPFRE	1228586694	1,03	0,32
MAPFRE vida	988800000	0,83	0,26
Banco Pastor	835939692	0,70	0,22
Banco Zaragozano	801975000	0,67	0,21
Banco de Valencia	769256734	0,65	0,20
Banco de Andalucía	627757744	0,53	0,17
Banco Guipuzcoano	491499277	0,41	0,13
Catalana Occidente	398400000	0,33	0,10
Total bancos y financieras	**119039123893**	**100,00**	**31,31**

Fuente: Página web de Bolsa de Madrid.

Ejemplo 4. Obtenga el índice de cotización bursátil para el conjunto de empresas que aparecen en la tabla siguiente.

Cotizaciones diarias en euros

	BSCH	BBVA	B. Popular	Bankinter	B. Zaragozano	B. Andalucía
19/11/01	10,16	14,29	37,95	34,35	8,51	37,55
20/11/01	9,85	14,07	37,15	33,55	8,53	37,55
21/11/01	9,73	13,90	37,16	33,88	8,60	37,62
22/11/01	9,69	13,79	38,08	34,03	8,56	37,93
23/11/02	9,79	13,98	37,68	34,16	8,60	38,00
26/11/01	9,80	14,00	37,20	34,10	8,50	37,70
27/11/01	9,59	13,80	37,00	33,90	8,60	37,80

Fuente: Página web de Bolsa de Madrid.

Como puede observarse se trata de un grupo de empresas correspondientes al Sector Bancos y Financieras para los que se tiene su cotización diaria para siete días, algo más de una semana. Para calcular el correspondiente índice de cotización se hará uso de las capitalizaciones que aparecen en la Tabla 18, lo que permite deducir que sus ponderaciones son:

Tabla 18. Capitalización y ponderaciones

BSCH	51476554502	45,03
BBVA	50654254882	44,31
B. Popular	8056417704	7,05
Bankinter	2709681666	2,37
B. Zaragozano	801975000	0,70
B. Andalucía	627757744	0,55
Total	**114326641496**	**100,00**

Fuente: Página web de Bolsa de Madrid.

En la Tabla 19 se dan los índices simples de cotización para cada empresa y el correspondiente índice compuesto obtenido mediante la siguiente expresión:

$$I_{19}^{t} = \sum_{i=}^{6} I_{19}^{t}(i) w_i \qquad (6.27)$$

donde:

I_{19}^{t} es el valor del índice compuesto correspondiente al día t con base al día 19.
$I_{19}^{t}(i)$ es el valor del índice simple correspondiente a la empresa *i* para el día t con base al día 19.
wi es la ponderación (capitalización) de la empresa i.

Tabla 19. Indices bursátiles simples y compuesto.

	Sector	BSCH	BBVA	B. Popular	Bankinter	B. Zaragozano	B. Andalucía
19/11/01	**100,00**	100,00	100,00	100,00	100,00	100,00	100,00
20/11/01	**97,74**	96,95	98,46	97,89	97,67	100,24	100,00
21/11/01	**96,71**	95,77	97,27	97,92	98,63	101,06	100,19
22/11/01	**96,38**	95,37	96,50	100,34	99,07	100,59	101,01
23/11/02	**97,35**	96,36	97,83	99,29	99,45	101,06	101,20
26/11/01	**97,35**	96,46	97,97	98,02	99,27	99,88	100,40
27/11/01	**95,76**	94,39	96,57	97,50	98,69	101,06	100,67

La forma en la que se ha calculado este índice complejo es aplicable solo cuando se producen variaciones de precios, pero a lo largo del año se produce una serie de hechos que provocan variaciones en las cotizaciones que no son motivados por las propias fuerzas del mercado (equilibrio entre la oferta y la demanda) como son: pagos de dividendos, ampliaciones de capital, splits o desdoblamientos, reducciones de capital con devolución al accionista…. Por ello, el Índice de Madrid realiza una serie de ajustes para que no se vea influenciado por estos hechos. (Para más detalles puede consultarse la página web de la Bolsa de Madrid).

5.6.2.2. IBEX35.

Otro índice de cotización bursátil que ha adquirido un significado valor de referencia en el mundo financiero es el IBEX35. Se trata de un índice que recoge la evolución de las cotizaciones de los 35 valores cotizados en el Sistema de Interconexión Bursátil de las cuatro Bolsas Españolas más líquidos durante el período de control (el intervalo de seis meses contados a partir del séptimo mes anterior al inicio del semestre natural).

La formula de cálculo de este índice es:

$$IBEX35(t) = IBEX35(t-1)\sum_{i=1}^{35} \frac{Cap_i(t)}{\left[Cap^i(t-1) \pm J \right]} \qquad (6.28)$$

donde:

$Cap_i(t)$ es la capitalización en el periodo *t* (producto del número de acciones por la cotización de las mismas).

$Cap_i(t)$ es la capitalización en el periodo *t-1*.

J es la cantidad utilizada para ajustar el valor del Indice por ampliaciones de capital, etc.

Esta fórmula de cálculo del IBEX35 es idéntica a la del IGBM siempre y cuando el número de acciones en el pierdo *t* sea el mismo que el del periodo base.

El Indice tiene como valor base 3000 al cierre de mercado el día 29 de diciembre de1989.

5.6.3. Otros Indices.

Indices implícitos de precios . Como su nombre indica se trata de índices de precios que no se obtienen de forma directa, sino como resultado de la doble valoración de las distintas macromagnitudes de la Contabilidad Nacional.

En la cuantificación de estas macromagnitudes para años sucesivos es tradicional dar dos series distintas de valores. Para centrar las ideas vamos a utilizar como ejemplo una de esas macromagnitudes, el valor añadido bruto (VAB), entendiendo como tal la diferencia entre la producción de bienes y servicios y los consumos intermedios, para cada rama de actividad y para el total de la economía. Pues bien, esta magnitud, como cualquier otra de las cuentas nacionales, es el resultado de multiplicar las unidades físicas producidas por sus precios. Según este criterio esos valores vendrían dados por:

$$\sum_{i=1}^{N} p_{i0}q_{i0} \quad \sum_{i=1}^{N} p_{i1}q_{i1} \quad \sum_{i=1}^{N} p_{i2}q_{i2} \quad \ldots \quad \sum_{i=1}^{N} p_{it}q_{it} \qquad (6.29)$$

A esta serie de valores se le conoce como VAB a precios corrientes (nominales), es decir, valorada según los precios de cada periodo. Pero la Contabilidad Nacional también da esa serie valorada a los precios del año base. En este caso a ese agregado se le llama VAB a precios constantes (términos reales), pues éstos no cambian, son los del año cero. Formalmente esta serie es:

$$\sum_{i=1}^{N} p_{i0}q_{i0} \quad \sum_{i=1}^{N} p_{i0}q_{i1} \quad \sum_{i=1}^{N} p_{i0}q_{i2} \quad \ldots \quad \sum_{i=1}^{N} p_{i0}q_{it} \qquad (6.30)$$

Pues bien, si se divide la primera de esas series entre la segunda, término a término, el resultado es:

$$\frac{\sum_{i=1}^{N} p_{i0}q_{i0}}{\sum_{i=1}^{N} p_{i0}q_{i0}} \quad \frac{\sum_{i=1}^{N} p_{i0}q_{i1}}{\sum_{i=1}^{N} p_{i0}q_{i0}} \quad \frac{\sum_{i=1}^{N} p_{i0}q_{i2}}{\sum_{i=1}^{N} p_{i0}q_{i0}} \quad \ldots \quad \frac{\sum_{i=1}^{N} p_{i0}q_{it}}{\sum_{i=1}^{N} p_{i0}q_{i0}} \qquad (6.31)$$

Pero, como puede apreciarse, esos cocientes son en realidad un Indice de Precios de Paasche a los se les conoce como Indice Implícito de Precios. De igual forma se pueden obtener los índices implícitos de precios para las demás macromagnitudes de la Contabilidad Nacional. Se trata pues, de un índice que recoge la evolución de los precios de los bienes y servicios que por agregación dan lugar a esas grandes cifras de la economía de un país. A este índice se le conoce también como deflactor del VAB. El concepto de inflación y deflación serán objeto de estudio en el siguiente epígrafe.

Índice de Producción Industrial (IPI). Es un indicador coyuntural que mide la evolución mensual de la actividad productiva de las ramas industriales, excluida la construcción, contenidas en la Clasificación Nacional de Actividades Económicas 1993 (CNAE -93). Se trata de un índice de cantidad que mide la evolución de la producción física industrial por ramas de actividad. Este índice lo elabora el INE, aunque las distintas CC.AA. también cuantifican este índice para sus ámbitos de competencia.

Indice de Precios Industriales (IPRI) . Es un indicador coyuntural que, elaborado por el INE, mide la evolución mensual de los precios de los productos industriales fabricados y vendidos en el mercado interior, en el primer paso de su comercialización, es decir, de los precios de venta a salida de fábrica obtenidos por los establecimientos industriales en las transacciones que estos efectúan, excluyendo los gastos de transporte y comercialización y el IVA facturado.

5.7 Inflación y poder adquisitivo. Deflación de valores monetarios.

La inflación es un fenómeno económico de naturaleza monetaria que por sus consecuencias ha sido, y continua siéndolo, fuente de preocupación para todos los agentes que intervienen en la economía, tanto los privados como los públicos. Pero antes de hablar de los efectos de la inflación lo más conveniente será definirla. Para ello recurriremos a la que da Samuelson en su manual de Economía. En el mismo se dice que "*Entendemos por inflación un periodo de aumento general de los precios de los bienes de consumo y de los factores productivos, elevándose los precios del pan, los automóviles, el corte de pelo, y aumentando los salarios, las rentas de la tierra, etc*." (Samuelson, 1975). Lo sustantivo de esta definición es que el fenómeno en cuestión consiste en un aumento general y sostenido de los precios de todos los bienes y servicios tanto producidos como consumidos.

Ese incremento generalizado de precios tiene como consecuencia inmediata que la capacidad de compra del dinero se reduce de forma continuada. Es decir, la cantidad de un bien que puede adquirirse con una unidad monetaria dada (euro, libra, dólar, etc) es cada vez menor como resultado del incremento del precio de ese bien. Pero si en lugar de tratarse de un solo bien, la subida de precios afecta a todos los bienes de una economía, la situación sería similar, solo que agravada. Así pues, la inflación reduce la capacidad de compra del dinero o poder adquisitivo del mismo.

La siguiente cuestión sería definir un instrumento estadístico que permita cuantificar esa subida generalizada de precios. Es decir, se trata de buscar un índice de precios que recoja de forma adecuada el fenómeno de la inflación. A tal efecto, el índice que suele utilizarse de forma casi unánime es el IPC, aunque el mismo tiene algunas limitaciones como veremos más adelante.

Tabla 20. Evolución de la inflación en España y poder adquisitivo de la peseta.

Años	IPC (1961=100)	Valor de pta.	Años	IPC (1961=100)	Valor de pta.
1961	100,00	1000,00	1981	846,62	118,12
1962	105,71	945,97	1982	968,66	103,23
1963	114,96	869,84	1983	1086,60	92,03
1964	122,99	813,08	1984	1209,17	82,70
1965	139,24	718,19	1985	1315,75	76,00
1966	147,93	676,00	1986	1431,48	69,86
1967	157,38	635,40	1987	1506,60	66,37
1968	165,16	605,46	1988	1579,47	63,31
1969	168,74	592,61	1989	1686,75	59,29
1970	178,42	560,48	1990	1800,12	55,55
1971	193,10	517,86	1991	1906,94	52,44
1972	209,09	478,26	1992	2019,93	49,51
1973	232,94	429,29	1993	2112,21	47,34
1974	269,49	371,07	1994	2211,89	45,21
1975	315,16	317,29	1995	2315,27	43,19
1976	370,72	269,75	1996	2397,67	41,71
1977	461,69	216,60	1997	2444,91	40,90
1978	552,98	180,84	1998	2489,76	40,16
1979	639,56	156,36	1999	2547,28	39,26
1980	739,10	135,30	2000	2634,75	37,95

En la Tabla 20 se recogen los valores medios anuales del IPC para España en el periodo 1961-2000 con base 1961. Según el contenido de esta tabla, el nivel medio de los precios en ese periodo de cuarenta años creció por encima de 26 veces. Esto, dicho en otros términos , equivale a que si en 1961 un bien costaba 37,95 pesetas, el precio de ese mismo bien en el año 2000 era 1000 pesetas, o lo que es igual, con 1000 pesetas del año 2000 solo se podía comprar lo que en 1961 con 37,95. Estas cifras dan una idea bastante clara de cual ha sido la pérdida de la capacidad de compra de la peseta en España a lo largo de esos años.

Cualquiera de las dos columnas principales de esa tabla reflejan ese incremento de precios y la consiguiente reducción de la capacidad adquisitiva del dinero. Pero la segunda columna, la encabezada como valor de la peseta, merece algún comentario

adicional. La misma es el resultado de dividir la cantidad 1000 (1000 pesetas de cada año) por su correspondiente IPC. Esas mil pesetas de cada año es una serie monetaria valorada con los precios de cada año ($p_t q_t$). A este tipo de series monetarias se le conoce como series a precios corrientes o series monetarias nominales. En cambio, cuando una serie monetaria a precios corrientes se divide por un índice de precios adecuado, como se ha hecho en la tabla anterior, el resultado es una serie a precios constantes o en términos reales. A esta operación se le conoce como deflactar una serie, es decir, quitarle a una serie el efecto precios. Por eso, una vez que se ha deflactado la serie de 1000 pesetas anuales pasando la serie a términos reales, se observa como las mil pesetas del año 2000 equivalen solo a 37,95 pesetas del año 1961.

Pero para deflactar una serie monetaria nominal hay que trabajar con el índice de precios (conocido como deflactor) adecuado. Se ha señalado antes que es el IPC el se utiliza a tal efecto de forma generalizada. Pero también se ha indicado que presentaba algunos problemas. Como sabemos el IPC al ser un índice de Laspeyres viene dado por:

$$IPC = P_L = \frac{\sum_{i=1}^{N} p_{it} q_{io}}{\sum_{i=1}^{N} p_{i0} q_{io}} \qquad (5.32)$$

mientras que una serie monetaria en términos nominales o a precios corrientes viene dada por:

$$\sum_{i=1}^{N} p_{it} q_{it} \qquad (5.33)$$

y otra a precios constantes o términos reales sería:

$$\sum_{i=1}^{N} p_{i0} q_{it} \qquad (5.34)$$

Pues bien, si deflacionamos una serie con el IPC resulta que:

$$\frac{\sum_{i=1}^{N} p_{it} q_{it}}{IPC} = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}} \neq \sum_{i=1}^{N} p_{i0} q_{it} \qquad (5.35)$$

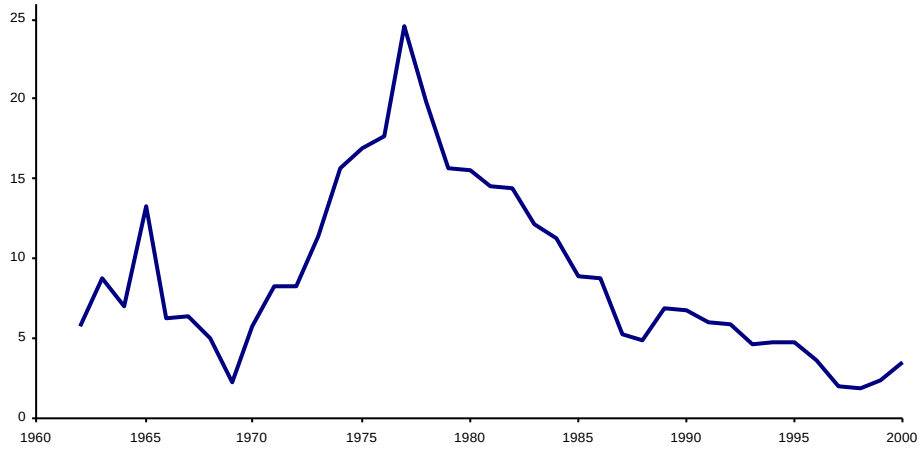
el resultado no es la serie en términos reales buscada, aunque se le parezca. Para conseguir esa serie habría que deflactar con un índice de precios de Paasche, pues en tal caso:

$$\frac{\sum_{i=1}^{N} p_{it} q_{it}}{P^P} = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}}} = \sum_{i=1}^{N} p_{i0} q_{it} \qquad (5.36)$$

Así pues, el deflactor adecuado de una serie monetaria nominal es un índice de precios de Paasche, como lo son los Índices Implícitos de Precios. Sin embargo, por problemas de cobertura (el IPC tiene una cobertura amplia que se adecua bastante bien a la mayoría de las series monetarias) o de información y cálculo (los índices de Laspeyres necesitan menos información que los de Paasche para su elaboración), es el IPC el que suele utilizarse para la deflación de series monetarias.

Vamos a finalizar este epígrafe realizando un pequeño análisis estadístico de la inflación en España. A tal efecto se han representado en el Grafico 1 las tasas medias anuales de inflación calculadas a partir de los datos del IPC de la Tabla 36. Como puede apreciarse las mayores tasas de inflación se dieron desde comienzos de los años setenta hasta mediados de los ochenta. Tanto es así que de 1973 a 1984 los precios se multiplicaron por más de 5 y en 1975 la tasa de inflación fue de casi un 25%. En cambio, durante los años noventa la inflación parecía que dejaba de ser un problema debido a su continuo decrecimiento, aunque de 1998 a 2000 la tasa de inflación casi se ha duplicado.

Grafico 1. Evolución de las tasas anuales medias de inflación en España



Pero una vez que se conoce como se han comportado los precios de forma conjunta, se podría plantear la cuestión de saber qué productos son los que tienen mayor o menor incidencia o repercusión en la variación global. Para responder a esta cuestión nos vamos a centrar en lo ocurrido en el año 2000.

Tabla 21. Variación anual del IPC en España por grupos de bienes. (Diciembre 2000 - Diciembre 1999)

Grupos	Variación anual	Ponderaciones
Alimentos Y Bebidas No Alcohólicas	3,2	215,051
Bebidas Alcohólicas Y Tabaco	3,8	32,182
Vestido Y Calzado	2,3	100,384
Vivienda	4,6	114,613
Menaje	3,0	63,574
Medicina	2,3	28,718
Transporte	6,3	157,331
Comunicaciones	-3,0	25,374
Ocio Y Cultura	5,1	65,238
Enseñanza	5,5	16,878
Hoteles, Cafés Y Restaurantes	4,6	113,259
Grupo Otros	4,2	67,398
General	**4,0**	**1000,000**

Fuente: Página web INE.

Para todo el año 2000 (de diciembre de 1999 a diciembre de 2000) la variación del IPC en España fue de 4 puntos. Esa variación determina la tasa de inflación, la cual se obtiene como:

$$T_{12}^{1} = \frac{IPC_1 - IPC_0}{IPC_0} x100 = \frac{\Delta IPC}{IPC_0} x100 \qquad (5.37)$$

donde:

$$\Delta IPC = IPC_1 - IPC_0 = \frac{\sum_{i=1}^{N} (p_{it} + \Delta p_{it}) q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} - \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} = \frac{\sum_{i=1}^{N} \Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} \qquad (5.38)$$

por lo que:

$$T_{12}^{1} = \frac{\frac{\sum_{i=1}^{N} \Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}}{\frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}} x100 = \frac{\sum_{i=1}^{N} \Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 \qquad (5.39)$$

es decir, la tasa de inflación global es función de las variaciones experimentadas por los precios de todos los bienes o grupos de ellos. Lo que nos interesa ahora es determinar la repercusión que en esa tasa o en la variación total han tenido cada uno de los bienes o grupos de ellos. Si nos fijamos en la Tabla 21 se puede observar que los tres grupos que más han incidido son Transporte (su variación de precios fue del 6,3), seguido de Enseñanza, Ocio y Cultura. Por el contrario, los tres que menos incidencia tuvieron fueron Comunicaciones (los precios de este

grupo decrecieron), Medicina y Vestido y Calzado. Pero aunque estas apreciaciones no son del todo erróneas tampoco reflejan fielmente la repercusión de los precios de cada grupo en la variación total, pues ésta es, en realidad, una media ponderada, como refleja (6.38). En consecuencia, habría que multiplicar la variación en precios de cada grupo por su correspondiente ponderación. Esto nos lleva a definir la repercusión de la forma siguiente:

$$R_i = \frac{\Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} \qquad (5.40)$$

de forma que la suma de las todas repercusiones sea igual a la variación total del índice dada por (5.38). De igual forma también se cumple que la suma de las repercusiones relativas es igual a la tasa de variación total del IPC, como vemos a continuación:

$$\sum_{i=1}^{N} \underline{R}_i = \sum_{i=1}^{N} \frac{\frac{\Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}}{\frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}}} x100 = \sum_{i=1}^{N} \frac{\Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 = IPC_0 \qquad (5.41)$$

De manera similar a como se han definido las repercusiones de cada bien o grupo de ellos sobre la variación total de precios, también se puede definir la participación de cada bien de la forma siguiente:

$$P_i = \frac{\frac{\Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}}}{\frac{\sum_{i=1}^{N} \Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}}} x100 = \frac{\Delta p_{it} q_{i0}}{\sum_{i=1}^{N} p_{it} q_{i0}} x100 \qquad (5.42)$$

Estos conceptos aplicados a los datos que aparecen en la Tabla 21 nos llevan a que:

Tabla 22. Repercusión y participación de los precios de cada grupo

	Variación	Ponderaciones	Repercusión	Participación
Alimentos y Bebidas no Alcohólicas	3,2	215,051	0,6882	17,33
Bebidas Alcohólicas y Tabaco	3,8	32,182	0,1223	3,08
Vestido y Calzado	2,3	100,384	0,2309	5,82
Vivienda	4,6	114,613	0,5272	13,28
Menaje	3	63,574	0,1907	4,80
Medicina	2,3	28,718	0,0661	1,66
Transporte	6,3	157,331	0,9912	24,97
Grupo Comunicaciones	-3	25,374	-0,0761	-1,92
Ocio y Cultura	5,1	65,238	0,3327	8,38
Enseñanza	5,5	16,878	0,0928	2,34
Hoteles, Cafés y Restaurantes	4,6	113,259	0,5210	13,12
Otros	4,2	67,398	0,2831	7,13
General	4	1000,000	3,97	100

De nuevo se observa que el grupo que mayor repercusión tuvo en la inflación española en el año 2000 fue el de Transportes, pero el segundo es Alimentos y Bebidas no Alcohólicas frente a Enseñanza como se ind icó antes. El tercero fue Vivienda. En cambio, los grupos de Enseñanza y Ocio y Cultura quedan relegados a posiciones más bajas, pues aunque sus precios aumentaron mucho, sin embargo en las cesta media de las familias españolas esas partidas de gasto tienen un peso relativamente bajo. En concreto, de mil pesetas gastadas por una familia, a Enseñanza no destina ni siquiera 17 pesetas (no hay que olvidar que en

España la mayor parte de la enseñanza es pública y gratuita), mientras que en Alimentación y Bebidas no Alcohólicas gasta más de 215 pesetas.

CAPÍTULO 6.- PROBABILIDAD

6.1 Introducción.

Como ya se ha visto, la *Estadística* es una *Ciencia* con la que se pretende buscar las regularidades existentes en el comportamiento de los datos. Sabemos que la Estadística se puede clasificar en dos grandes bloques: *Estadística Descriptiva e Inferencia Estadística*. Con el primero lo que se hace es dar un conjunto de métodos y herramientas que permiten estudiar esas regularidades cuando lo que observamos es toda la población. Es decir admitimos que es posible realizar esa operación de recuento exhaustivo. En tal caso lo que realizamos con la estadística es estudiar, describir, el comportamiento de una variable determinada. Esa observación exhaustiva nos permite realizar afirmaciones "categóricas" sobre las distintas características de la variable, tales como cual es su media , su dispersión, la forma de la distribución, etc.

Pero esa posibilidad de observación exhaustiva no siempre es posible. En la gran mayoría de los casos nos vemos limitados a realizar una observación parcial de la variable. Con ese conjunto limitado de datos intentaremos conocer las características de toda la población, es decir, intentaremos inferir su comportamiento. Así una empresa antes de lanzar un nuevo producto estará interesada en conocer cual puede ser su cuota de mercado, para lo cual realizará un sondeo de opinión entre algunos de sus potenciales clientes. Pero el resultado de ese sondeo, basado en una muestra (observación parcial), no le permite concluir cual será su verdadera cuota de mercado. La decisión que tome respecto a ese producto estará marcada por un cierto grado de incertidumbre.

Pero que duda cabe que, en esas situaciones, nuestras afirmaciones ya no pueden ser "categóricas" y las decisiones que se tomen puede que no sean las más acertadas como consecuencia de la información no contenida en la muestra. Más bien al contrario debemos admitir que nuestras conclusiones están sujetas a un margen de incertidumbre que es la consecuencia de nuestra observación parcial de la realidad. Ante tales circunstancias nuestro objetivo será doble: por un lado estudiar el comportamiento de la variable y de otro reducir en la medida de lo posible ese margen de incertidumbre o, al menos, intentar cuantificar esa falta de certeza en relación a las características de las variables. Una forma de cuantificar esa incertidumbre es

haciendo uso del concepto de probabilidad. De hecho la probabilidad es un concepto con el que convivimos de forma diaria, incluso sin percatarnos de él. Cada vez que hacemos uso de las expresiones *quizás*, *tal vez*, *es probable*, *puede que*, etc. estamos implícitamente hablando en términos probabilísticos. La incertidumbre es una acompañante inseparable de todas las ciencias sociales e incluso de las físicas como señaló Heisenberg con el enunciado del principio de incertidumbre de la mecánica cuántica.

6.2 Conceptos previos.

Pero antes de dar la definición de probabilidad es aconsejable introducir una serie de conceptos previos que nos serán de gran utilidad. Empezaremos con el de *fenómeno aleatorio*. Como sabemos un fenómeno es algo observable y que en la mayoría de los casos es, además, cuantificable. Podemos decir que la estadística tiene por objeto el estudio y comportamiento de fenómenos. Estos fenómenos son a su vez el resultado de una experimentación, por lo que podemos hablar indistintamente de *fenómenos y* experimentos aleatorios. De forma específica se dice que *un experimento aleatorio es aquel que puede concretarse en al menos dos resultados posibles, con incertidumbre en cuanto a cual de ellos tendrá lugar*.

Los experimentos se pueden clasificar en *deterministas y aleatorios*. Los primeros son aquellos que repetidos en idénticas condiciones nos llevan siempre al mismo resultado. Por el contrario, para el segundo tipo de experimentos nos encontramos que, incluso aunque las condiciones del experimento no cambien, el resultado del experimento es impredecible antes de realizarlo. (Antes de lanzar una moneda al aire no sabremos si saldrá cara o cruz. También son experimentos aleatorios la cotización de las acciones de una empresa, sus beneficios, sus ventas, su periodo de actividad, etc.). En general diremos que las características de un experimento aleatorio son las siguientes:

a) el experimento se puede repetir u observar de forma indefinida en circunstancias prácticamente muy similares.

b) Aunque no podemos predecir el resultado particular del experimento, si que podemos conocer el conjunto de todos los posibles resultados.

c) Si el experimento se repite pocas veces, los resultados parecen mostrar un comportamiento caótico, mientras que si se repite un número infinito de veces empieza a detectarse una regularidad en el comportamiento de los resultados.

Hemos señalado antes que una de las características del experimento aleatorio es que, aunque los resultados individuales no son predecibles con anterioridad, en cambio si que podemos saber cual es el conjunto de todos sus posibles resultados. Pues bien, a ese conjunto de posibles resultados le llamaremos espacio muestral y lo representaremos en adelante por la letra *E*. Así pues, *E* será un conjunto formado por los resultados del experimento. Estos resultados elementales de un experimento tienen la característica de que no son descomponibles. A partir de ellos surge el concepto de suceso o evento. Un suceso o evento será un conjunto de resultados elementales del experimento. Antes de continuar con el concepto de suceso o evento conviene señalar que un espacio muestral puede ser finito (si está formado por un conjunto finitos de resultados) o infinito. Dentro los espacios infinitos se puede diferenciar entre los infinitos numerable e infinitos no numerables. Tanto a los espacios finitos como a los infinitos numerables se les suele conocer como espacios discretos, mientras a que los infinitos no numerable se conoce también como continuos.

Habiéndose definido previamente el concepto de suceso, a continuación vamos a dar una tipología de los mismos dentro de la cual se distingue: suceso elemental, suceso compuesto (consta de dos o más sucesos elementales), suceso seguro o universal (coincide con el espacio muestral) y suceso imposible (no contiene ningún elemento del espacio muestral *E* y por tanto no ocurrirá nunca y lo denotaremos por $\emptyset$).

Ejemplo 1. En el experimento que consiste en lanzar un dado de seis caras vamos a concretar los conceptos de suceso elemental, suceso compuesto o evento, suceso seguro, suceso imposible, espacio muestral y naturaleza del mismo.

En este experimento si admitimos que cada una de las caras se identifican por los enteros que van del 1 al 6, de forma que a la cara uno se la identifica por el valor 1, a la dos por el valor 2, y así sucesivamente, entonces los sucesos elementales de este experimento, que representaremos por e_i, serán los enteros e_1 =1, e_2 =2, e_3 =3, e_4 =4, e_5=5, e_6 =6. A partir de éstos se pueden definir otros eventos. Así, el evento A = "número par" se define como A = {2, 4, 6}, el evento B = "número primo" viene dado por

B = {1, 2, 3, 5}, etc. A su vez el suceso seguro en este experimento es E = “que salga alguna cara” y está formado por E = {1, 2, 3, 4, 5, 6}. Sobre un experimento aleatorio se puede definir más de un suceso imposible, aunque todos ellos satisfacen la definición dada con anterioridad. Así en este ejemplo sería sucesos imposibles los siguientes: $\mathcal{O}$ = “que sal la cara siete”, $\mathcal{O}$ = “obtener la cara dos y medio”, etc. Finalmente el espacio muestral asociado a este experimento vendría dado por E = {1, 2, 3, 4, 5, 6}, es decir, el conjunto de todos los resultados posibles del mismo. En este caso se trata de un espacio finito y, por lo tanto, discreto.

Ejemplo 2. Sea el experimento que consiste en contar el número de mujeres en una muestra de 12 parlamentarios seleccionados al azar.

En este caso el espacio muestral correspondiente a este experimento viene dado por E = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}, que también es finito y discreto. Para este experimento también se puede definir distintos tipos de eventos como: A = “que el número de mujeres sea mayoría”; b = “que el número de mujeres sea al menos tres”; etc.

Ejemplo 3. Sea el experimento que consiste en contar el número de personas que llega a la caja de un supermercado durante un mes.

El espacio muestral de este experimento viene dado por E = {0, 1, 2, 3, 4, 5, }. En este caso estamos ante un espacio infinito numerable y, en consecuencia, también discreto.

Ejemplo 4. Sea el experimento que consiste en anotar el instante en que se recibe una llamada telefónica a lo largo de un día.

Si se admite que esa llamada puede ocurrir en cualquier instante de ese intervalo de 24 horas, entonces el espacio muestral será E = {el intervalo de tiempo correspondiente a las 24 horas}, que origina un espacio infinito no numerable, es decir, continuo.

Una vez que se ha dado el concepto de suceso o evento, a continuación se van a definir las *operaciones* más habituales que pueden realizarse con los mismos.

a) Suceso contenido en otro. Se dice que *A* está contenido en *B* y lo indicaremos por $A \varepsilon B$ si todos los elementos de *A* pertenecen a *B*.

Ejemplo 5. A partir del experimento definido en el Ejemplo 2, vamos a definir los sucesos A = "que haya 8 ó 9 mujeres" y B = "que haya mayoría de mujeres". En este caso se dice que $A \varepsilon B$.

b) Igualdad de sucesos. Se dice que *A* y *B* son dos sucesos iguales si se cumple simultáneamente que $A \varepsilon B$ y $B \varepsilon A$.

Ejemplo 6. Con el mismo experimento del Ejemplo 2 se puede definir los sucesos A = "mayoría de mujeres" y B = "al menos siete mujeres". Aquí se cumple que $A \varepsilon B$ y $B \varepsilon A$, por lo que A = B.

c) Unión de sucesos. Dados dos sucesos *A* y *B*, se define la unión de ambos como otro suceso, que indicaremos por $A \cup B$, que está formado por los elementos pertenecientes a *A*, o a *B* o a los dos a la vez.

Ejemplo 7. Con el mismo experimento del Ejemplo 2 se puede definir los sucesos A = "al menos siete mujeres" y B = "más de cinco mujeres pero menos de diez". En este caso:

$$A = \{7, 8, 9, 10, 11, 12\} \quad B = \{6, 7, 8, 9\}$$

Por lo que

$$A \cup B = \{7, 8, 9, 10, 11, 12\} \cup \{6, 7, 8, 9\} = \{6, 7, 8, 9, 10, 11, 12\}$$

d) Intersección de sucesos. Dados dos sucesos *A* y *B*, se define la intersección de ambos como otro suceso, que representamos por $A \nu B$, compuesto por resultados comunes a *A* y *B* simultáneamente.

Ejemplo 8. Con el mismo experimento del Ejemplo 2 se pueden definir los sucesos A = "al menos siete mujeres" y B = "más de cinco mujeres pero menos de diez". En este caso:

$$A = \{7, 8, 9, 10, 11, 12\} \quad B = \{6, 7, 8, 9\}$$

Por lo que

$$A \cup B = \{7, 8, 9, 10, 11, 12\} \cup \{6, 7, 8, 9\} = \{ 7, 8, 9\}$$

e) Sucesos disjuntos, incompatibles o mutuamente excluyentes. Dados dos sucesos *A* y *B*, se dicen que ambos son incompatibles, disjuntos o mutuamente excluyentes si la presencia de uno impide la del otro. En tal caso ocurre que $A \cup B = \emptyset$.

Ejemplo 9. Con el mismo experimento del Ejemplo 2 se pueden definir los sucesos A = “al menos siete mujeres” y B = “no más de cinco mujeres”. En este caso:

$$A = \{7, 8, 9, 10, 11, 12\} \quad B = \{0, 1, 2, 3, 4, 5\}$$

Por lo que

$$A \cup B = \{7, 8, 9, 10, 11, 12\} \cup \{0, 1, 2, 3, 4, 5\} = O$$

f) Complementario o contrario. Dado un suceso *A*, se define el complementario de *A* como otro suceso que ocurre cuando no ocurre *A* y que representaremos por $\overline{A}$.

Ejemplo 10. Con el mismo experimento del Ejemplo 2 se puede definir el sucesos A = “al menos siete mujeres”.

El complementario de este suceso es:

$$\overline{A} = \{ 0, 1, 2, 3, 4, 5, 6\}$$

6.3 Concepto de probabilidad.

El *concepto de probabilidad* es muy antiguo y a lo largo de la historia se ha definido de distintas formas, aunque todas ellas mantienen en común las características básicas del concepto. En general cuando hablemos de probabilidad lo haremos siempre en referencia a la probabilidad de un suceso y la entenderemos como una medida

cuantificada de la verosimilitud de ocurrencia de un suceso frente a los demás sucesos del experimento. Pero que duda cabe que esta definición no es del todo buena, pues se utiliza el término verosimilitud para definir la probabilidad, cuando el mismo es un sinónimo de lo que se quiere definir. También podría hablarse del grado de incertidumbre en la ocurrencia de los resultados de un experimento. En cualquier caso la probabilidad de un suceso es una medida cuantificable que toma valores entre cero y uno a diferencia del concepto de posibilidad que es una medida cualitativa.

Una vez que se ha dado el concepto de probabilidad en sentido amplio debemos señalar que a lo largo de la historia podemos encontrar tres formas distintas de definir o interpretar la probabilidad. Cada uno de ellas responde a un tipo de experimento distinto. En concreto, supongamos que queremos calcular la probabilidad de los siguientes sucesos:

- Obtener un seis al lanzar un dado honesto.
- Obtener un seis al lanzar un dado cargado.
- Que la tasa de crecimiento del VAB de España sea superior al 4%.

Para obtener esas probabilidades hay que recurrir a enfoque o definiciones distintas. En realidad esos enfoques sirven para establecer reglas de asignación de probabilidades a los sucesos más que para definir la probabilidad.

Probabilidad clásica o a priori (Regla de Laplace). Si el experimento que estamos realizando da lugar a un espacio muestral *E* que es finito y cuyos resultados son conocidos de antemano y equiprobables o simétricos, entonces, la probabilidad del suceso *A* perteneciente a *E* se define como el cociente de los resultados favorables a *A* respecto del total de resultados posibles.

$$P(A) = \frac{\text{Número de resultados favorables a A}}{\text{Número de resultados posibles}}$$

A esta expresión se le conoce como regla de Laplace.

Este concepto de probabilidad está íntimamente ligado a los juegos de azar. Esta definición satisface tres propiedades:

1) No negatividad, *P(A)* $\geq$ *0.*

2) Certeza, *P(E) =1*.

3) Aditividad. Si *A* y *B* son dos sucesos del espacio *E* y ambos son mutuamente excluyentes, entonces la probabilidad de *C =A* $\cup$ *B* será: *P(C) = (PA)+P(B).*

Antes de finalizar con este concepto de probabilidad hay que señalar la razón de su denominación. Así el adjetivo "clásica" hace alusión a que fue la forma en la que los primeros estadísticos abordaron este concepto. A su vez el término "a priori" se refiere a que la probabilidad de cualquiera de los sucesos de este tipo de experimentos es conocida incluso antes que los mismos tengan lugar. De hecho no es necesario realizar el experimento para conocer las probabilidades de sus resultados.

Probabilidad frecuencial o a posteriori. En este caso la probabilidad de un suceso *A* se define como el límite de una frecuencia relativa, cuando el experimento se realiza un número infinito de veces. Formalmente diremos que

$$P(A_i) = \lim_{n\to\infty} \frac{n(A_i)}{n}, \qquad \text{i} = 1,2,3,\ldots,\text{k}$$

Esta definición de probabilidad cumple también las tres propiedades enunciadas en el caso anterior.

Con este concepto de probabilidad lo que se pretende es dar respuesta a experimentos en los que no se cumplen los requisitos señalados antes, en especial el de equiprobabilidad o simetría de los resultados. Esta circunstancia conlleva que la probabilidad de cada resultado no sea conocido de antemano, siendo necesaria la realización del experimento para la cuantificación de la misma.

Con esta definición se puede determinar la probabilidad de: las caras de un dado cuando el mismo está cargado; pieza defectuosa en la producción de una empresa; accidente de tráfico; factura impagada; cliente moroso; que el cliente de un establecimiento comercial sea menor de 25 años; que los ingresos de una persona sea superior a la media; etc.

La probabilidad definida bajo este enfoque también satisface las tres propiedades dadas anteriormente.

Ejemplo 11. Los 1000 empleados de una empresa, según la edad y el sexo de los mismos, vienen dados en la siguiente tabla de doble entrada.

Sexo / Edad	Mujeres	Hombres	Total
Menos de 30 años	100	250	350
De 30 y más años	200	450	650
Total	300	700	1000

Obtenga la probabilidad de que elegido un empleado al azar el mismo sea:

a) Hombre
b) Mujer
c) Menor de 30 años
d) De 30 o más años
e) Mujer menor de 30 años
f) Hombre de 30 y más años

Antes de calcular esas probabilidades vamos a definir simbólicamente cada uno de esos sucesos:

A = el empleado seleccionado es hombre
B = el empleado seleccionado es mujer
C = el empleado seleccionado es menor de 30 años
D = el empleado seleccionado tiene 30 o más años

Definidos los sucesos de esta forma, las probabilidades pedidas son:

a) P(A) = (700/1000) = 0,7
b) P(B) = (300/1000) = 0,3
c) P(C) = (350/1000) = 0,35
d) P(D) = (650/1000) = 0,65
e) $P(B \cap C) = (100/1000) = 0,10$

f) $P(A \cap D) = (450/1000) = 0,45$

Probabilidad subjetiva. Hay determinados experimentos aleatorios que no son susceptibles de realizarse y sus resultados no son equiprobables. Imaginemos que se quiere determinar la probabilidad: de que la economía de España crezca en el próximo año un 3%; que las acciones de una empresa se revaloricen en un 10% en un mes; que una empresa presente suspensión de pagos; que un nuevo producto sea bien acogido en el mercado; que ocurra un accidente nuclear; etc.

En estas circunstancias, donde los experimentos solo se pueden realizar una vez o ninguna o que se puedan repetir pero en condiciones distintas, no son aplicables ninguna de las dos definiciones dadas anteriormente, por lo que no es posible asignar probabilidades mediante un procedimiento objetivo, debiendo recurrir a procedimientos de tipo subjetivo, a opiniones de expertos. En estos casos la probabilidad expresa un grado de creencia o confianza individual en relación con la ocurrencia o no de un determinado suceso. Se trata de un juicio personal sobre el resultado de un experimento aleatorio. Además debemos admitir la posibilidad de que distintos sujetos asignen probabilidades diferentes al mismo suceso. No obstante esta definición de probabilidad también satisface las tres propiedades vistas antes.

Probabilidad axiomática. Para dar esta definición es preciso, previamente, definir el concepto de α-*álgebra de Boole*. Un α-*álgebra de Boole*, que representaremos por $A=P(E)$, es una familia de sucesos no vacía, la cual contiene necesariamente los sucesos O y E y que, además, es cerrada para las operaciones de complementación y de unión de infinitos subconjuntos numerables de E, sien E el espacio muestral del experimento. En base a este concepto, la probabilidad axiomática se define como una función de conjunto, que llamaremos P, cuyo dominio es el α-*álgebra de Boole* y cuyo recorrido es el intervalo cerrado $[0,1]$ si además satisface los *tres axiomas siguientes* (*axiomas de Kolmogorov*):

1) *Axioma de no negatividad*. $P(A) \geq 0$, para todo $A \in A$.
2) *Axioma de certeza*. $P(E) = 1$.

3) *Axioma de aditividad*. Si A_1, A_2, …..es una sucesión numerable de sucesos pertenecientes a A, tales que entre si son mutuamente excluyentes, $A_i \nu A_j = 0$ para todo $i * j$, entonces $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

6.4 Teoremas básicos sobre probabilidad.

A continuación vamos enunciar una serie de teoremas sobre probabilidad, de gran utilidad, que se deducen de los axiomas anteriores.

1. Para cualquier suceso $A E A$ se verifica que la probabilidad de su complementario $P(\bar{A})$ =1-P(A).

Para demostrar este teorema partimos de que:

$$E = A \cup \bar{A} \qquad \text{y} \qquad A \cap \bar{A} = \varnothing$$

Por otro lado según los axiomas segundo y tercero se tiene que:

$$1 = P(A) + P(\bar{A})$$

por lo que:

$$P(A) = 1 - P(\bar{A})$$

2. La probabilidad del suceso imposible es nula. $P(0)=0$

Si en el teorema 1 se hace que A = 0, entonces E = $\bar{A}$, por lo que

$$P(\varnothing) = 1 - P(E) = 1 - 1 = 0$$

3. La probabilidad P es monótona no decreciente, es decir, para cualesquiera sucesos $A, B E A$, tales que $A \varepsilon B$, entonces $P(A) \leq P(B)$.

Para demostrar este teorema se parte de que

$$B = A \cup \left(B \cap \bar{A}\right) \quad \text{y que} \quad A \cap \left(B \cap \bar{A}\right) = \varnothing$$

Pero según los axiomas primero y tercero resulta que

$$P(B) = P(A) + P(B \cap \overline{A}) \geq P(A)$$

4. Para cualquier suceso $A \in \mathcal{A}$ se verifica que $0 \leq P(A) \leq 1$.

La primera desigualdad de este teorema es el primero de los axiomas. En cuanto a la segunda se tiene que $A \subset E$, por lo que, según el teorema anterior, resulta que:

$$P(A) \leq P(E) = 1$$

5. Regla de la suma. Para cualesquiera sucesos $A, B \in \mathcal{A}$ se verifica que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Para demostrar este teorema vamos a expresar los sucesos $(A \cup B)$ y A como la unión de los siguientes sucesos disjuntos:

$$(A \cup B) = B \cup \left(A \cap \overline{B}\right) \qquad A = (A \cap B) \cup \left(A \cap \overline{B}\right)$$

A su vez, la probabilidad de los mismos, según el tercero de los axiomas, viene dada por

$$P(A \cup B) = P(B) + P\left(A \cap \overline{B}\right) \qquad \text{y} \qquad P(A) = P(A \cap B) + P\left(A \cap \overline{B}\right)$$

De la segunda probabilidad se deduce que

$$P\left(A \cap \overline{B}\right) = P(A) - P(A \cap B)$$

Si ahora se sustituye este resultado en $P(A \cup B)$ se llega a que

$$P(A \cup B) = P(B) + P\left(A \cap \overline{B}\right) = P(B) + P(A) - P(A \cap B)$$

Ejemplo 12. Sean *A* y *B* dos sucesos tales que: *P(A∪B) = 3/4*; $P(\overline{A}) = 2/3$ y *P(A∩B) = 1/4*. Hallar: a) *P(A); b) P(B); c)* $P(A \cap \overline{B})$.

a) P(A) =1- $P(\overline{A})$ *=1- 2/3 = 1/3*

b) P(B) = P(A∪B) - P(A) + P(A∩B) = 2/3

c) $P(A \cap \overline{B})$ *= P(A) - P(A∩B) = 1/12.*

Ejemplo 13. La probabilidad de que las acciones de una empresa financiera coticen al alza es 0,8, mientras que esa probabilidad para una empresa del sector nuevas tecnologías es 0,4. A su vez, la probabilidad de que las dos coticen al alza es 0,3. Obtenga las siguientes probabilidades: a) que coticen al alza al menos una de las dos empresas; b) que ninguna de las dos cotice al alza; c) que solo cotice una al alza.

Para dar solución a este ejercicio vamos a proceder en primer lugar a definir los siguientes sucesos:
A = la empresa del sector financiero cotiza al alza.
B = la empresa del sector nuevas tecnologías cotiza al alza.
C = al menos una empresa cotiza al alza.
D = ninguna de las dos empresas cotiza al alza.
E = solo una empresa cotiza al alza.

a) A partir del enunciado sabemos que P(A) = 0,8; P(B) = 0,4 y P(A ∩ B) = 0,3. Con ello tenemos que:

P(C)= P(A∪B) = P(A) + P(B) – P(A ∩ B) = 0,8 + 0,4 – 0,3 = 0,9

b) El suceso D se puede expresar como:

$$D = \overline{A} \cap \overline{B} = \overline{A \cup B}$$

Este resultado nos lleva a que

$$P(D) = P\left(\overline{A \cup B}\right) = 1 - P(A \cup B) = 1 - 0{,}9 = 0{,}1$$

c) El suceso E se puede expresar como

$$E = \left(A \cap \overline{B}\right) \cup \left(\overline{A} \cap B\right)$$

Pero como se trata de la unión de dos sucesos disjuntos, entonces la probabilidad del suceso E es

$$P(E) = P\left(A \cap \overline{B}\right) + P\left(\overline{A} \cap B\right)$$

Ahora bien

$$A = (A \cap B) \cup \left(A \cap \overline{B}\right) \text{ por lo que } P\left(A \cap \overline{B}\right) = P(A) - P(A \cap B)$$

A su vez

$$B = (A \cap B) \cup \left(\overline{A} \cap B\right) \text{ por lo que } P\left(\overline{A} \cap B\right) = P(B) - P(A \cap B)$$

Todo ello nos permite escribir

$$P(E) = P\left(A \cap \overline{B}\right) + P\left(\overline{A} \cap B\right) = P(A) + P(B) - 2P(A \cap B) = 0{,}8 + 0{,}4 - 2(0{,}3) = 0{,}6$$

6.5 Probabilidad condicional, regla de la multiplicación e independencia de sucesos.

Hasta ahora hemos definido la probabilidad de un suceso *A* referida a todo el espacio muestral *E* del experimento. Supongamos ahora la existencia de otro suceso *B* definido sobre *E* y que no sea incompatible con *A*, es decir que *(A ν B) ∗ 0.* Esto significa que los sucesos *A* y *B* tienen partes en común. Supongamos adicionalmente que tenemos la certeza de que ha ocurrido el suceso *B*. Ahora estamos interesados en saber como

cambia la probabilidad de A sabiendo que ha ocurrido B. Sabiendo que ha ocurrido *B*, la probabilidad de que ocurra *A* se representa por *P(A / B)* y se le conoce como *probabilidad condicional*. En estas circunstancias, para calcular la probabilidad de *A* hay que cambiar el espacio de referencia el cual, ahora ya no es *E* sino *B*, y habrá que exigir que no sea un espacio nulo, es decir, debe cumplirse que *P(B) > 0*. Si sabemos que el suceso *B* ha ocurrido, entonces se sabe que el resultado del experimento es uno de los incluidos en *B*. Por tanto, para evaluar la probabilidad de que ocurra *A*, se debe considerar el conjunto de los resultados incluidos en *B* que también implique la ocurrencia de *A*. Este conjunto viene dado por la intersección de *A* y *B,* es decir *(A ∩ B)*. En tales circunstancias resulta natural definir la probabilidad condicional de *A* dado que ha tenido lugar *B* de la siguiente forma:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

En realidad para definir esta probabilidad se ha recurrido a la *regla de Laplace*, en el sentido de que si sabemos que ha ocurrido *B*, entonces, ahora, estos son los casos posibles del experimento, mientras que los favorables estarían constituidos por todos aquellos elementos que pertenecen simultáneamente a *A* y a *B,* es decir *(A ∩ B)*.

Esta definición de probabilidad tiene la particularidad de que ha implicado una redefinición de las probabilidades de *A* en base a la información que representa el conocimiento de la presencia del suceso *B*, el cual es ahora el nuevo espacio muestral de referencia y que al ser más pequeño que E supone una reducción de incertidumbre en relación con el suceso *A*.

Una vez dado el concepto de probabilidad condicional no resulta difícil demostrar que esta definición satisface los tres axiomas de la probabilidad.

A partir de esta definición del concepto de probabilidad condicional se puede expresar la correspondiente al suceso intersección como:

$$P(A \cap B) = P(A)P(B/A) = P(B)P(A/B)$$

A esta forma de dar la probabilidad de la intersección de dos sucesos se le conoce como regla del producto. Si en lugar de tener dos sucesos se tuvieran tres, entonces la probabilidad de la intersección de los tres vendrá dada por:

$$P(A \cap B \cap C) = P(A)P(B/A)P(C/A \cap B)$$

o por cualquiera de las otras cinco ordenaciones posibles. Esta regla puede extenderse para el caso de que el número de sucesos sea mayor que tres.

La definición de probabilidad condicional pone de manifiesto que la ocurrencia de un suceso B puede modificar la probabilidad de otro suceso A. Si esto no ocurriera se diría que los sucesos A y B son independientes. Antes de dar una definición formal de este concepto haremos uso de un ejemplo donde queden claras estas ideas.

Ejemplo 14. Supongamos que se tiene un dado de seis caras construido de forma honesta. En tal caso todas las caras son equiprobables y el espacio muestral asociado al experimento que consiste en lanzarlo al aire es E = {1, 2, 3, 4, 5, 6}. A partir de este espacio muestral vamos a definir los sucesos: A = “obtener número par”; B = “obtener un dos o un cinco”; C = “obtener un 4”.

Para este experimento aleatorio, las probabilidades de los sucesos definidos antes son: P(A) = 1/2; P(B) = 1/3; P(C) =1/6. Ahora bien, si nos dijeran que al lanzar el dado ha tenido lugar el suceso C, entonces P(A/C) = 1, dado que ($C \subset A$). Vemos como el conocer que ha tenido lugar C modifica la probabilidad de A. Por otro lado, si nos hubieran dicho que ha ocurrido B resulta ahora que:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/3} = \frac{1}{2} = P(A)$$

En este caso la presencia de B no ha alterado la probabilidad del suceso A. En estas circunstancias se dice que la probabilidad de A no depende de la presencia de B. Esta idea se puede expresar también diciendo que A y B son dos sucesos independientes. Es decir, los sucesos A y B se dicen que son independientes cuando la presencia de uno de ellos no afecta a la probabilidad del otro.

Si el resultado de este ejemplo lo lleváramos a la regla del producto definida antes se tiene entonces que:

$$P(A \cap B) = P(A)P(B)$$

Pues bien, cuando se cumple esta última igualdad se dice que los sucesos son independientes. Esta condición de independencia entre sucesos es equivalente a que P(A) = P(A/B), o bien que P(B) = P(B/A). Pero que dos sucesos sean independientes no significa que sean mutuamente excluyentes. Este segundo caso se da cuando esos sucesos no pueden ocurrir simultáneamente y, por lo tanto, su intersección es el suceso imposible, por lo que su probabilidad será nula.

Si en lugar de tener los sucesos A y B se tuvieran los sucesos A, B y C, entonces se diría que los tres son independientes si lo son dos a dos y los tres a la vez. Es decir si se cumple que:

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C)$$

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

Ejemplo 15. En un departamento hay cuatro ordenadores numerados del 1 al 4. Si se seleccionan dos ordenadores al azar y se definen los sucesos A = {1, 2}, B = {1, 3} y C = { 1, 4} resulta que P(A) = P(B) = P(C) = 1/2. Además

$$P(A \cap B) = \frac{1}{4} = P(A)P(B), \quad P(A \cap C) = \frac{1}{4} = P(A)P(C), \quad P(B \cap C) = \frac{1}{4} = P(B)P(C)$$

Este resultado nos permite decir que los sucesos son independientes por pares. En cambio:

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C)$$

Ello nos lleva a concluir que esos tres sucesos no son independientes.

Ejemplo 16. La probabilidad de que una empresa venda un producto defectuoso cuando la producción se somete a un proceso diario de control de calidad es 0,005. La

probabilidad de que un día no haya control de calidad es 0,05 y la probabilidad de que esa empresa venda un producto defectuoso es 0,02. Determinar:

a) La probabilidad de que se venda un producto defectuoso y que haya control de calidad.
b) La probabilidad de que habiéndose vendido un producto defectuoso haya habido control de calidad.
c) La probabilidad de que habiéndose vendido un producto defectuoso no haya habido control de calidad.
d) La probabilidad de que habiéndose vendido un producto no defectuoso haya habido control de calidad.
e) La probabilidad de que habiéndose vendido un producto no defectuoso no haya habido control de calidad.
f) La probabilidad de que no habiendo control de calidad se venda un producto defectuoso.
g) La probabilidad de que no habiendo control de calidad se venda un producto no defectuoso.

Antes de dar respuesta a cada uno de estos apartados vamos a definir los siguientes sucesos: *D* = "venta de producto defectuoso" y *C* = "hay control de calidad".

A su vez, el enunciado del ejercicio nos facilita la siguiente información:

$$P(D/C)=0{,}005;\quad P(D)=0{,}02;\quad P(\overline{C})=0{,}05$$

A partir de esta información resulta inmediato que:

$$P(\overline{D}/C)=1-P(D/C)=0{,}995;\quad P(\overline{D})=1-P(D)=0{,}98;\quad P(C)=1-P(\overline{C})=0{,}95$$

Con toda esta información tenemos que:

a)

$$P(D/C)=\frac{P(C\cap D)}{P(C)}\quad \text{por lo que}\quad P(C\cap D)=P(C)P(D/C)=(0{,}95)(0{,}005)=0{,}00475$$

b)

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{0{,}00475}{0{,}02} = \frac{475}{2000} = 0{,}2375$$

c)

$$P(\overline{C}/D) = 1 - P(C/D) = 1 - \frac{475}{2000} = \frac{1525}{2000} = 0{,}7625$$

d)

$$P(C/\overline{D}) = \frac{P(C \cap \overline{D})}{P(\overline{D})} = \frac{P(C)P(\overline{D}/C)}{P(\overline{D})} = \frac{(0{,}95)(0{,}995)}{(0{,}98)} = 0{,}9645$$

e)

$$P(\overline{C}/\overline{D}) = 1 - P(C/\overline{D}) = 0{,}0355$$

f)

$$P(D/\overline{C}) = \frac{P(\overline{C} \cap D)}{P(\overline{C})} = \frac{P(D)P(\overline{C}/D)}{P(\overline{C})} = \frac{(0{,}02)(0{,}7625)}{0{,}05} = 0{,}305$$

g)

$$P(\overline{D}/\overline{C}) = \frac{P(\overline{C} \cap \overline{D})}{P(\overline{C})} = 1 - P(D/\overline{C}) = 0{,}695$$

Para todos los apartados, puede apreciarse como influye de manera decisiva sobre la probabilidad inicial de los sucesos C y D (así como de sus respectivos complementarios) la información que se incorpora en el cálculo de las respectivas probabilidades condicionales. Así, mientras que P(C) = 0,95, en cambio, P(C/D) = 0,2375. Es decir, la probabilidad de que se realice un control de calidad es alta y, en esas circunstancias, es poco probable que se venda una pieza defectuosa (esa probabilidad no llega al 1%). Sin embargo, si se sabe que la pieza vendida es defectuosa entonces será poco probable que haya habido control de calidad, como de hecho se confirma con la nueva probabilidad. Este tipo de razonamiento es aplicable a todas las demás situaciones contempladas en este ejercicio.

Ejemplo 17. Una empresa que se dedica a la venta de sus productos por internet está interesada en conocer cuales son sus clientes potenciales. Para ello realiza una encuesta a 1000 personas atendiendo a su edad y al número de horas semanales que navegan en al red, obteniendo los resultados que se dan en la tabla siguiente.

Edad / Horas	Menores de 25 años	De 25 a 45 años	Mayores de 45 años	Total
Menos de 7 horas	100	250	100	450
De 7 a 14 horas	100	150	100	350
Más de 12 horas	100	50	50	200
Total	300	450	250	1000

A partir de la información de esta tabla se van a definir los siguientes sucesos:

A_1: persona menor de 25 años

A_2: persona de 25 a 45 años

A_3: persona mayor de 45 años

B1: navegar menos de 7 horas a la semana

B_2: navegar entre 7 y 14 horas a la semana

B_3: navegar más de 14 horas a la semana.

Con esta notación, la tabla anterior se puede expresar como:

Edad / Horas	A_1	A_2	A_3	Total
B_1	$(A_1 \cap B_1)=100$	$(A_2 \cap B_1)=250$	$(A_3 \cap B_1)=100$	$B_1=450$
B_2	$(A_1 \cap B_2)=100$	$(A_2 \cap B_2)=150$	$(A_3 \cap B_2)=100$	$B_2=350$
B_3	$(A_1 \cap B_3)=100$	$(A_2 \cap B_3)=50$	$(A_3 \cap B_3)=50$	$B_3=200$

Total	A_1=300	A_2=450	A_3=250	1000

Llegados a este punto se obtiene la siguiente tabla de probabilidades, siempre y cuando admitamos que la muestra anterior es representativa de la población de la que se ha extraído:

Edad / Horas	A_1	A_2	A_3	Total
B_1	$P(A_1 \cap B_1)=0,1$	$P(A_2 \cap B_1)=0,25$	$P(A_3 \cap B_1)=0,1$	$P(B_1)=0,45$
B_2	$P(A_1 \cap B_2)=0,1$	$P(A_2 \cap B_2)=0,15$	$P(A_3 \cap B_2)=0,1$	$P(B_2)=0,35$
B_3	$P(A_1 \cap B_3)=0,1$	$P(A_2 \cap B_3)=0,05$	$P(A_3 \cap B_3)=0,05$	$P(B_3)=0,2$
Total	$P(A_1)=0,3$	$P(A_2)=0,45$	$P(A_3)=0,25$	1

La lectura del contenido de esta tabla es sencillo e inmediato. Supongamos ahora que queremos determinar la probabilidad de que, seleccionada una persona al azar, ésta navegue más de 14 horas a la semana sabiendo que es menor de 25 años. Para determinar esa probabilidad no tiene sentido que trabajemos con todo el espacio muestral (las 1000 personas de la muestra), pues sabemos que es menor de 25 años. Así pues nuestro nuevo espacio muestral será el formado por el suceso A_1. Ahora el denominador de esa frecuencia relativa, que es la probabilidad, no es 1000, sino que es A_1=300. A su vez el numerador deja de ser B_3=200, pues dentro de ese colectivo de 200 personas que navegan más de 14 horas a la semana hay algunas que tienen más de 25 años, pero nosotros sabemos que nuestra persona seleccionada es menor de 25. Ahora el numerador es $(A_1 \cap B_3)=100$. Todo ello lleva a que la nueva probabilidad viene dada por el cociente:

$$P(B_3 / A_1) = \frac{P(B_3 \cap A_1)}{P(A_1)} = \frac{\frac{n(B_3 \cap A_1)}{1000}}{\frac{n(A_1)}{1000}} = \frac{n(B_3 \cap A_1)}{n(A_1)} = \frac{100}{300} = \frac{1}{3}$$

Además, como *P(B₃) = 0,20*, resulta que los sucesos *A₁* y *B₃* no son independientes. Pero como estos suceso no son independientes, se puede concluir que, para este

ejemplo, los atributos edad y horas de navegación en internet tampoco son independientes.

A partir de los datos de esta última tabla de doble entrada podemos definir los conceptos de probabilidad conjunta y probabilidad marginal. La primera es la que hace alusión a la presencia o aparición simultánea de más de un suceso elemental, es decir, la probabilidades de la intersección $P(A_i \cap B_j)$ para cualquier i o j. En cambio las probabilidades marginales son las de los sucesos elementales, $P(A_i)$ o $P(B_j)$.

Pero cualquier suceso elemental se puede expresar como la unión de un conjunto de sucesos mutuamente excluyentes. Para el ejemplo que estamos usando se tiene que el suceso A_2 sería:

$$A_2 = (A_2 \cap B_1) \cup (A_2 \cap B_2) \cup (A_2 \cap B_3)$$

y, en general, para el suceso A_i

$$A_i = (A_i \cap B_1) \cup (A_i \cap B_2) \cup \ldots \cup (A_i \cap B_k)$$

Esto lleva a que cualquier toda probabilidad marginal se pueda poner como una suma de probabilidades conjuntas. Es decir:

$$P(A_i) = P(A_i \cap B_1) + P(A_i \cap B_2) + \ldots\ldots\ldots + P(A_i \cap B_k) = \sum_k P(B_k)P(A_i/B_k)$$

A este resultado se le conoce como Teorema de la probabilidad total.

Ejemplo 18. Una empresa dedicada al montaje de ordenadores recibe procesadores procedentes de tres fabricantes distintos. Los procesadores que recibe pueden ser buenos o defectuosos y, por experiencia anterior, esta empresa trabaja con los siguientes datos:

	Empresa		
	H_1	H_2	H_3

Componente	Bueno (B)	0,23	0,30	0,39
	Defectuoso (D)	0,02	0,05	0,01

a) Si se elige un procesador al azar de entre todos los recibidos, ¿cuál es la probabilidad de que sea defectuoso?

b) Si se elige un procesador al azar de entre todos los recibidos, ¿cuál es la probabilidad de que proceda de la empresa H_1?

c) ¿Cuál es la probabilidad de que un procesador procedente de la empresa H_1 sea defectuoso?

d) ¿Cuál es la probabilidad de que un procesador defectuoso elegido al azar proceda de la empresa H_1?

e) ¿Es la calidad del procesador independiente del proveedor?

f) Teniendo en cuenta la calidad, ¿cuál de las empresas es más fiable?

En este ejemplo la información viene dada en forma de tabla de doble entrada que se puede completar en la manera siguiente:

	H_1	H_2	H_3	Total
B	0,23	0,30	0,39	0,92
D	0,02	0,05	0,01	0,08
Total	0,25	0,35	0,40	1,00

Con esta información las respuestas a cada uno de los apartados anteriores serían las siguientes:

a)

$$P(D) = 0,08 = \leftarrow_i P(H_i \nu D) = 0,02 + 0,05 + 0,01 = 0,08$$

b)

$$P(H_1) = 0,25$$

c)

$$P(D / H_1) = \frac{P(D \cap H_1)}{P(H_1)} = \frac{0,02}{0,25} = 0,08$$

d)

$$P(H_1 / D) = \frac{P(D \cap H_1)}{P(D)} = \frac{0,02}{0,08} = 0,25$$

e)

No, porque aunque $P(H_1/D) = P(H_1)$, sin embargo, $P(H_2/D) * P(H_2)$ y $P(H_3/D) * P(H_3)$, como es fácil de comprobar.

f)

La más fiable resulta ser H_3, según se desprende de las siguientes probabilidades condicionales:

$$P(D/H1) = 0,08;\ P(D/H_2) = 0,143;\ P(D/H_3) = 0,025$$

Ejemplo 19. En este ejemplo vamos a realizar una aplicación del concepto de probabilidad condicional en un contexto de muestreo, donde alguna de las preguntas del cuestionario sea de tal naturaleza que haya reticencias a contestarla de forma directa. En concreto, imaginemos que deseamos conocer la probabilidad de que una familia, en su declaración de renta, cometa fraude. Está claro que, si esta pregunta se hace de forma directa, lo más probable es que se tengan muchas respuestas falsas o muchas respuestas en blanco. Para evitar cualquiera de estas dos posibilidades y alcanzar el objetivo marcado se puede proceder de la forma siguiente. Se formulan dos preguntas: a) ¿es par el último dígito de su DNI? b) ¿ha cometido fraude en la declaración de la renta?. Cada entrevistado ha de responder solo a una de ellas en función del siguiente criterio: antes de responder lanza una moneda al aire y solo él conoce el resultado de ese lanzamiento. Si sale cara responderá a la pregunta a), mientras que si sale cruz responderá a b). Realizado este experimento el 40% de los entrevistados respondieron “si”. En estas circunstancias nuestro interés radica en determinar la probabilidad de que habiendo contestado a la pregunto b) haya dicho que

"si". Como puede comprobarse, con este procedimiento se garantiza al entrevistado que su respuesta es anónima y que solo él conoce la naturaleza de la misma. A este procedimiento se le conoce como muestreo con respuesta aleatorizada.

Veamos ahora cual es la probabilidad pedida. En primer lugar definamos los siguientes sucesos: A = el entrevistado responde "si", B = el entrevistado responde a la pregunta a), C = el entrevistado responde a la pregunta b).

Para estos suceso sabemos que: P(A) = 0,40; P(B) = 0,50 y P(C) = 0,50. Estas dos últimas probabilidades se deben a que son el resultado del lanzamiento de una moneda. Además sabemos que P(A/B) = 0,50, pues la mitad de los DNI termina en número par. Con toda esta información lo que se pretende es calcular P(A/C). Como los sucesos B y C se han construido de forma que sean mutuamente excluyentes y exhaustivos se tiene que:

$$(B \cap A) \cup (C \cap A) = A$$

por lo que

$$P(A) = P(B \cap A) + P(C \cap A) = P(B)P(A/B) + P(C)P(A/C)$$

Si en esta expresión se sustituyen los valores conocidos se tiene que:

$$0{,}40 = (0{,}50)(0{,}50) + (0{,}50)P(A/C)$$

Llegados a este punto es fácil determinar que la probabilidad buscada es: *P(A/C) = 0,30.*

Ejemplo 20. Un analista financiero está realizando un estudio para la captación de clientes en base a los registros de su antigua empresa. A partir de esos datos ha concluido que la probabilidad de que un cliente entre en el mercado de renta variable es 0,10. Además ha observado que para ese tipo de clientes, el 30% crean su propia cartera de valores y de ellos la mitad son partícipes de fondos de inversión. También

ha observado que de los que no tienen cartera propia, el 40% destinan sus ahorros a fondos de inversión.

Con esta información vamos definir los siguientes sucesos:

A: invierte en bolsa

B: invierte en cartera propia

C: Invierte en fondos.

Para ellos se sabe que:

$$P(A) = 0{,}10$$

$$P(B/A) = 0{,}30$$

$$P(C/A \nu B) = 0{,}50$$

$$P(C/A \nu \overline{B}) = 0{,}40$$

Determinar:

a) Probabilidad de A y B.

b) Probabilidad de A y B y C.

c) Probabilidad de C.

d) Probabilidad de B.

e) Probabilidad de A o B o C.

f) Probabilidad de no A, no B y no C.

Dadas las probabilidades iniciales se llega a que:

P($\overline{A}$)=0,90; P($\overline{B}$/A)=0,70; P($\overline{C}$/A$\cap$B)=0,50; P($\overline{C}$/A$\cap\overline{B}$)=0,60; P(B/$\overline{A}$)=0;

P($\overline{B}$/$\overline{A}$)=1; P(C/$\overline{A}\cap$B)=0; P($\overline{C}$/$\overline{A}\cap$B)=1; P(C/$\overline{A}\cap\overline{B}$)=0; P($\overline{C}$/$\overline{A}\cap\overline{B}$)=1

a) El suceso definido en este apartado, para este ejemplo, es equivalente a este otro:

$$(A \cap B) = \left[A \cap B \cap \left(C \cup \overline{C}\right)\right] = (A \cap B \cap C) \cup \left(A \cap \overline{B} \cap C\right)$$

por lo que la probabilidad pedida es:

$$P(A \cap B) = P\left[A \cap B \cap \left(C \cup \overline{C}\right)\right] = P(A \cap B \cap C) + P\left(A \cap \overline{B} \cap C\right) =$$

$$= P(A)P(B/A)P(C/A\cap B) + P(A)P(B/A)\overline{P}\left(C/A\cap B\right) =$$

$$= P(A)P(B/A)\left[P(C/A\cap B) + \overline{P}\left(C/A\cap B\right)\right] = P(A)P(B/A)$$

Pues la probabilidad entre corchetes vale en nuestro caso la unidad. Todo esto nos lleva a que:

$$= P(A\cap B) = P(A)P(B/A) = (0{,}10)(0{,}30) = 0{,}03$$

b)

$$P(A\cap B\cap C) = P(A)P(B/A)P(C/A\cap B) = (0{,}10)(0{,}30)(0{,}50) = 0{,}015$$

c)

$$P(C) = P(A\cap B\cap C) + \overline{P}\left(A\cap B\cap C\right) + P\left(\overline{A}\cap B\cap C\right) + P\left(A\cap B\cap C\right) =$$

$$= P(A)P(B/A)P(C/A\cap B) + \overline{P}\left(A\right)P\left(B/A\right)P\left(C/A\cap B\right) + P\left(A\right)P\left(B/A\right)P\left(C/A\cap B\right) +$$

$$+ P\left(\overline{A}\right)P\left(B/A\right)P\left(C/A\cap B\right) = (0{,}015) + (0{,}90)(0{,}00)(0{,}00) + (0{,}10)(0{,}70)(0{,}40) +$$

$$+ (0{,}90)(1{,}00)(0{,}00) = 0{,}015 + 0{,}028 = 0{,}043$$

d)

$$P(B) = P(A\cap B\cap C) + P\left(A\cap \overline{B}\cap C\right) + P\left(A\cap B\cap C\right) + P\left(A\cap B\cap C\right) =$$

$$= P(A)P(B/A)P(C/A\cap B) + P(A)P(B/A)\overline{P}\left(C/A\cap B\right) + P\left(A\right)\overline{P}\left(B/A\right)P\left(C/A\cap B\right) +$$

$$+ P\left(\overline{A}\right)P\left(B/A\right)\overline{P}\left(C/A\cap B\right) = (0{,}015) + (0{,}10)(0{,}30)(0{,}50) + (0{,}90)(0{,}00)(0{,}00) +$$

$$+ (0{,}90)(0{,}00)(0{,}00) = 0{,}015 + 0{,}015 = 0{,}03$$

e)

$$P(A\cup B\cup C) = P(A\cap B\cap C) + P\left(A\cap \overline{B}\cap C\right) + P\left(A\cap B\cap C\right) + P\left(A\cap B\cap C\right) +$$
$$+ P\left(A\cap B\cap C\right) + \overline{P}\left(A\cap B\cap C\right) + P\left(A\cap B\cap C\right) = P(A)P(B/A)P(C/A\cap B) +$$

$$+P(A)P(B/A)P(\overline{C}/A\cap B)+P(A)P(\overline{B}/A)P(C/A\cap \overline{B})+P(\overline{A})P(B/\overline{A})P(C/\overline{A}\cap B)+$$
$$+P(A)P(\overline{B}/A)P(\overline{C}/A\cap \overline{B})+P(\overline{A})P(B/\overline{A})P(\overline{C}/\overline{A}\cap B)+P(\overline{A})P(\overline{B}/\overline{A})P(C/\overline{A}\cap \overline{B})=$$
$$=(0{,}1)(0{,}3)(0{,}5)+(0{,}1)(0{,}3)(0{,}5)+(0{,}1)(0{,}7)(0{,}4)+(0{,}9)(0{,}0)(0{,}0)+(0{,}1)(0{,}7)(0{,}6)+$$
$$=(0{,}9)(0{,}0)(1)+(0{,}9)(1)(0{,}0)=0{,}10$$

f)

$$P(\overline{A}\cap\overline{B}\cap\overline{C})=\overline{P}(A)P(\overline{B}/\overline{A})\overline{P}(C/\overline{A}\cap B)=(0{,}9)(1)(1)=0{,}9$$

Ejemplo 21. La población activa de un país, según grandes sectores económicos, se distribuye de la forma siguiente: el 10% pertenece al sector primario, el 20% al industrial, el 5% a la construcción y el resto a los servicios. Por otro lado se sabe que de la población activa agraria el 35% está parada, de la industrial ese porcentaje es el 20% y en la construcción es el 15%. Además, de entre todos los parados los sector agrario son el 25%. Determine la probabilidad de paro que existe en el sector servicios.

Sean:
A = activo del sector primario
I = activo del sector industrial
C = activo del sector construcción
S = activo del sector servicios
P = activo parado

P(A) = 0,10; P(I) = 0,20; P(C) = 0,05; P(S) = 0,65; P(P/A) 0,35; P(P/I) = 0,20; P(P/C) = 0,15; P(A/P) = 0,25.

Lo que nos piden es *P(P/S)*. Para obtener esa probabilidad sabemos que:

P(P) = P(A ∨ P) + P(I ∨ P) + P(C ∨ P) + P(S ∨ P) = P(A)P(P/A) + P(I)P(P/I) +
+ P(C)P(P/C) + P(S)P(P/S) = (0,10)(0,35) + (0,20)(0,20) +(0,05)(0,15) +(0,65) P(P/S)=
= 0,035 + 0,04 + 0,0075 + (0,65) P(P/S).

Por otro lado sabemos que:

P(A ∨ P) = P(A)P(P/A) = (0,10)(0,35) = 0,035 = P(P)P(A/P) = P(P)(0,25)

De donde resulta que *P(P) = 0,14*. Haciendo uso de esta probabilidad resulta, finalmente, que:

$$P(P/S) = 0,088.$$

6.6 Teorema de Bayes.

Supongamos que tenemos un espacio muestral sobre el realizamos dos particiones tales cada una de ellas es exhaustiva. Tal podría ser el caso de un conjunto de *1000* personas que las clasificamos en relación a la actividad y la edad. Admitamos que esa clasificación nos lleva a la siguiente tabla:

	16-19 (H_1)	20-24 (H_2)	25-54 (H_3)	55+ (H_4)	Total
Activos (A_1)	25	35	425	50	535
No activos (A_2)	70	60	185	150	465
Total	95	95	610	200	1000

Ahora seleccionamos una persona al azar y nos preguntamos cual es la probabilidad de que sea Activo. En este caso, tal probabilidad viene dada por $P(A_1)=(535/1000)=0,535$. Al mismo resultado se habría llegado si hubiéramos hecho uso del teorema de la probabilidad total que se vio en el apartado anterior:

$$P(A_1)=(535/1000)=0,535=$$

$$=\sum P(H_i)\ P(A_1/H_i)=0.095(25/95)+0.095(35/95)+0.61(425/610)+02(50/200)=0535.$$

Para este ejemplo puede resultar innecesario la aplicación de esta última expresión, dada la abundante información de que se dispone. Pero si en lugar de conocer la tabla completa solo conociéramos la distribución porcentual de la población por edad (probabilidades de pertenecer a un grupo de edad concreto) y las tasas de actividad por edad (probabilidades condicionadas), entonces no se podría haber aplicado la definición de probabilidad frecuencial y la única solución habría sido el uso del teorema de la probabilidad total. La probabilidad del suceso A_1 es, en ese caso, la tasa de actividad total que, puede comprobarse, es una media ponderada de las tasas de actividad por edades, siendo las ponderaciones el peso relativo de cada grupo de edad.

Este resultado nos sirve para introducir el Teorema de Bayes. La aplicación de este teorema ha dado lugar al nacimiento de una rama de la Estadística. La que se conoce como Teoría Bayesiana. No es este el momento de entrar en más detalles respecto de esta cuestión, por lo que nos limitaremos exponer el teorema en sí. Para ello haremos uso del ejemplo.

Supongamos de nuevo que sobre un espacio muestral podemos realizar dos particiones que son, cada una de ellas, exhaustivas. Entonces, por la definición de probabilidad condicional tenemos que:

$$P(H_i / A_1) = \frac{P(H_i \cap A_1)}{P(A_1)} = \frac{P(H_i)\, P(A_1 / H_i)}{\sum_{i=1}^{k} P(H_i)\, P(A_1 / H_i)}$$

En el ejemplo con el que estamos trabajando *k=4*. A los posibles *k* resultados se les conoce habitualmente como *causas* o *hipótesis* y a $P(H_i)$ se les llama *probabilidades a priori*, mientras que a $P(H_i / A_1)$ se les llama *probabilidades a posteriori* y, finalmente, a $P(A_1/H_i)$ se les conoce como *verosimilitudes*.

La idea de este teorema es muy simple. Con el mismo lo que se pretende es modificar el conocimiento inicial que se tiene a cerca de una determinada realidad (las probabilidades a priori), haciendo uso de una información adicional que generalmente es de tipo muestral (las verosimilitudes). Se trata de ver en qué medida es información muestral nos lleva a cambiar nuestras hipótesis iniciales.

Ejemplo 22. Un analista de coyuntura económica quiere realizar predicciones a corto plazo sobre la evolución de la economía. Para ello utiliza como indicador adelantado el consumo de energía eléctrica. Por experiencia pasada sabe que cuando la economía crece durante un periodo a un ritmo superior al del periodo anterior (escenario A) la probabilidad de que el consumo eléctrico sea alto es 0,90. Si ese crecimiento es igual al del periodo anterior (escenario B) la probabilidad anterior es 0,50. Finalmente, si el crecimiento está por debajo al observado en el periodo anterior (escenario C), entonces aquella probabilidad se reduce al 0,20. Además se sabe que los pronósticos respecto del comportamiento de la economía asignan al escenario A una probabilidad

del 0,20 y al B del 0,60. Determinar: a) La probabilidad de que se de el escenario A y que el consumo eléctrico sea alto. b) La probabilidad de que el consumo eléctrico sea alto. c) Si el consumo es alto, ¿cuál es la probabilidad de los distintos escenarios?.

Antes de responder a las tres cuestiones planteadas vamos a representar simbólicamente cada uno de los sucesos definidos en el ejercicio así como a resumir las probabilidades que se nos dan.

A = tiene lugar el escenario A.
B = tiene lugar el escenario B.
C = tiene lugar el escenario C.
D = consumo eléctrico alto.

P(A) = 0,20; P(B) = 0,60; P(C) = 0,20; P(D/A) =0,90; P(D/B) = 0,50; P(D/C) = 0,20.

a) $P(A \cap D) = P(A)P(D/A) = (0,20)(0,90) = 0,18.$

b) $P(D) = P(A \cap D) + P(B \cap D) + P(C \cap D) = P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C) =$
$= (0,20)(0,90) + (0,60)(0,50) + (0,20)(0,20) = 0,52.$

c)

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)P(D/A)}{P(D)} = \frac{0,20x0,90}{0,52} = 0,346$$

$$P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{P(B)P(D/B)}{P(D)} = \frac{0,60x0,50}{0,52} = 0,577$$

$$P(C/D) = \frac{P(C \cap D)}{P(D)} = \frac{P(C)P(D/C)}{P(D)} = \frac{0,20x0,20}{0,52} = 0,077$$

En este ejemplo, la interpretación de los resultados sería la siguiente. La probabilidad a piori del suceso A es relativamente baja (solo del 0,20). Pero este suceso se asocia positivamente con D y como este ha tenido lugar, ahora, la probabilidad a posteriori es superior a la inicial. Es decir, las previsiones de las que se partía hay que revisarlas al alza pues ha tenido lugar un suceso que nos induce a pensar que la probabilidad de que la economía crezca es superior a la de partida.

La información de este ejercicio, así como los resultados del mismo, se puede resumir en la siguiente tabla:

Sucesos (Escenarios)	Probabilidades a priori	Verosimilitudes	Probabilidad total	Probabilidades a posteriori
A	0,20	0,90	(0,2)(0,9)= 0,18	0,346
B	0,60	0,50	(0,6)(0,5)= 0,30	0,577
C	0,20	0,20	(0,2)(0,2)= 0,04	0,077
Total	1,00		= 0,52	1,000

Ejemplo 23. Una empresa que se dedica a al envasado de café utiliza a tal efecto tres máquinas A, B y C. De ellas sabe, por controles de calidad previos, que la primera deposita menos cantidad de la establecida en un 2% de los paquetes, la segunda en 1% y la tercera en un 3%. El 40% del envasado lo realiza la máquina A y el 35% la B. Si se selecciona al azar un paquete, determinar la probabilidad: a) de que proceda de A si tiene menos cantidad de la establecida; b) de que no proceda de A si tiene la cantidad correcta.

Sean:
A = envasado por A
B = envasado por B
C = envasado por C
D = menos cantidad de la establecida.

P(A) = 0,40; P(B) = 0,35; P(C) = 0,25; P(D/A) =0,02; P(D/B) = 0,01; P(D/C) = 0,03.

a) Previamente es necesario obtener P(D).

P(D) = P(A ∨D) + P(B ∨D) + P(C ∨D) = P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C) =
= (0,40)(0,02) + (0,35)(0,01) + (0,25)(0,03) = 0,008 + 0,0035 + 0,0075 = 0,019.

$$P(A/D) = \frac{P(A \cap D)}{P(D)} = \frac{P(A)P(D/A)}{P(D)} = \frac{0{,}40x0{,}02}{0{,}19} = 0{,}421$$

b)

$$P(\overline{A}/\overline{D}) = \frac{P(\overline{A} \cap \overline{D})}{P(\overline{D})} = \frac{P(\overline{A \cup D})}{P(\overline{D})} = \frac{1 - P(A \cup D)}{1 - P(D)} = \frac{1 - P(A) - P(D) + P(A \cap D)}{1 - P(D)} =$$

$$= \frac{1 - P(A) - P(D) + P(A)P(D/A)}{1 - P(D)} = \frac{1 - 0{,}40 - 0{,}019 + 0{,}40x0{,}02}{1 - 0{,}019} =$$

$$= 0{,}6.$$

Ejemplo 24. El parlamento de un determinado país está integrado por diputados de tres partidos, A, B y C. La proporción de parlamentarios de esos partidos es de un 30%, un 50% y un 20%, respectivamente. La probabilidad de que un parlamentario del parido A vote afirmativamente a una determinada propuesta es 0,80, mientras que las probabilidades de votar en contra de esa propuesta para los parlamentarios de B y C son 0,7 y 0,9, respectivamente. Determinar la probabilidad de que un voto afirmativo provenga del partido A o del C.

Sean:

A = diputado del partido A

B = diputado del partido B

C = diputado del partido C

S = vota afirmativamente

N = vota en contra.

Si se admite que no hay votos nulos ni en blanco, entonces:

P(A) = 0,30; P(B) = 0,50; P(C) = 0,20; P(S/A) = 0,80; P(S/B) = 0,30; P(S/C) = 0,10.

$$P(A \cup C/S) = \frac{P(A \cap S) + P(C \cap S)}{P(S)} = P(A/S) + P(C/S) = \frac{P(A)P(S/A) + P(C)P(S/C)}{P(S)} =$$

$$= \frac{(0{,}30)(0{,}80) + (0{,}20)(0{,}10)}{(0{,}30)(0{,}80) + (0{,}50)(0{,}30) + (0{,}20)(0{,}10)} = 0{,}634$$

Ejemplo 25. En un sistema de alarma, la probabilidad de que se produzca una situación de peligro es 0,10. Si esta tiene lugar, la probabilidad de que la alarma funcione es 0,95. La probabilidad de que funcione la alarma sin que haya situación de peligro es 0,03. Determinar la probabilidad de que : a) habiendo funcionado la alarma no ya situación de peligro; b) se de una situación de peligro y la alarma no funcione; c) no habiendo funcionado la alarma, haya peligro.

Sean:
P = se da una situación de peligro.
F = funciona la alarma.

P(P) = 0,10; P(F/P) =0,95; P(F/ $\overline{P}$) = 0,03.

a)

$$P\left(\overline{P}/F\right)=\frac{P\left(\overline{P}\cap F\right)}{P(F)}=\frac{P\left(\overline{P}\right)P\left(F/\overline{P}\right)}{P\left(\overline{P}\right)P\left(F/\overline{P}\right)+P(P)P(F/P)}=$$

$$=\frac{(0{,}90)(0{,}03)}{(0{,}90)(0{,}03)+(0{,}10)(0{,}95)}=0{,}2195$$

b)

P(P ν $\overline{F}$) = P(P)P($\overline{F}$/P) = P(P)[1-P(F/P)] = (0,10)[1-(0,95)] = 0,005.

c)

$$P\left(P/\overline{F}\right)=\frac{P\left(P\cap\overline{F}\right)}{P\left(\overline{F}\right)}=\frac{P(P)P\left(\overline{F}/P\right)}{P(P)P\left(\overline{F}/P\right)+P\left(\overline{P}\right)P\left(\overline{F}/\overline{P}\right)}$$

$$=\frac{(0{,}10)(0{,}05)}{(0{,}10)(0{,}05)+(0{,}90)(0{,}97)}=0{,}000572$$

Ejemplo 26. Suponga que Vd. es el responsable de una agencia de alquiler, para todo el estado, de un modelo especifico de automóvil. Su agente de servicio en una determina ciudad no ha sido totalmente digno de confianza, porque en el pasado

suspendió el servicio en un 10% de las ocasiones. El efecto de dicha suspensión es que la probabilidad de que un cliente cancele el contrato pasa de 0,20 a 0,50. Si un cliente canceló un contrato, ¿cuál es la probabilidad de que alguna vez haya sido afectado por la suspensión del servicio?.

Sean:
A = suspende servicio
C = cancela contrato.

P(A) = 0,10; P(C/A) = 0,50; P(C/ $\bar{A}$)= 0,20.

$$P(A/C) = \frac{P(A \cap C)}{P(C)} = \frac{P(A)P(C/A)}{P(A)P(C/A) + P(\bar{A})P(C/\bar{A})} =$$

$$= \frac{(0,10)(0,50)}{(0,10)(0,50) + (0,90)(0,20)} = 0,21$$

Ejemplo 27. Un gran almacén está considerando cambiar su política de otorgamiento de créditos para reducir el número d clientes que finalmente no pagan sus cuentas. El gerente sugiere que en el futuro le sea cancelado el crédito a cualquier cliente que se demore un mes o más en sus pagos en dos ocasiones distintas. La sugerencia del gerente se basa en el hecho de que, en el pasado, el 90% de todos los clientes que finalmente no pagaron sus cuentas se habían demorado en sus pagos por lo menos en dos ocasiones. Supongamos que, tras una auditoria, se encuentra que el 2% de todos los clientes que compran a crédito finalmente no pagan sus cuentas, y que de aquellos que finalmente si las pagan, el 45% se han demorado al menos en dos ocasiones. Determine la probabilidad de que un cliente, que ya se demoró por lo menos en dos ocasiones, finalmente no pague su cuenta y, con la información obtenida, analice la política que ha sugerido el gerente.

Sean:
D = demora de pago en al menos dos ocasiones
P = paga

P(P) = 0,98; P(D/P) =0,45; P(D/ $\overline{P}$) = 0,90.

$$P\left(\overline{P}/D\right)=\frac{P\left(\overline{P}\cap D\right)}{P(D)}=\frac{P\left(\overline{P}\right)P\left(D/\overline{P}\right)}{P\left(\overline{P}\right)P\left(D/\overline{P}\right)+P(P)P(D/P)}=$$

$$=\frac{(0{,}02)(0{,}90)}{(0{,}02)(0{,}90)+(0{,}98)(0{,}45)}=0{,}0{,}0392$$

A la vista de este resultado se tiene que solo 4 de cada cien clientes (de forma aproximada), que incurren en demora de al menos dos ocasiones, terminan por no pagar. Así, si se acepta el plan del gerente, y se le retira el crédito a todos aquellos clientes que se demoran en pagar en al menos dos ocasiones, resultará que de esos 100 clientes a los que se le va a retirar el crédito, solo 4 son lo que realmente terminan por no pagar, mientras que los otras 96 restantes si que pagan. Este plan conlleva la posibilidad de perder a 96 clientes admisibles por castigar a 4 inadmisibles.

INTRODUCCIÓN A LA ESTADÍSTICA EMPRESARIAL

Relación de Ejercicios nº 1 (Temas 1 – 4). ***Curso 2002/2003***

---

1) El aplazamiento en el cobro de las últimas cien ventas facturadas por un establecimiento se había agrupado en cuatro intervalos, recordándose sólo los siguientes datos de la distribución:

a) El primer intervalo tiene 6 semanas como extremo superior, una frecuencia relativa de 0,2 y una amplitud de 4 semanas.
b) En el segundo intervalo se acumulan 60 ventas.
c) Las marcas de clase del segundo y cuarto intervalos son 8 y 50 semanas, respectivamente.
d) El tercer intervalo presenta una frecuencia de 30 ventas y una amplitud de 30 semanas.

Con esta información reconstruya la distribución de frecuencias, y represente el histograma correspondiente.

2) Una determinada licenciatura ofrece dos posibles especialidades. Tras la realización de un examen se obtienen los siguientes resultados en cada una de ellas.

Especialidad A		**Especialidad B**	
Calificaciones	**Nº de alumnos**	**Calificaciones**	**Nº de alumnos**
0-3	70	0-3	10
3-5	32	3-5	12
5-8	84	5-8	18
8-10	14	8-10	60

a) ¿Cuál es la nota media de todos los alumnos? ¿Qué especialidad resulta más homogénea?
b) ¿Cuál es la nota más frecuente en la especialidad A?
c) Un alumno de la especialidad A ha obtenido 6 puntos y otro de la B 7,25. ¿Cuál ha logrado una mejor calificación relativa dentro de su grupo?
d) En la especialidad B se ha decidido conceder una ayuda al 20% de los alumnos con mejores calificaciones. ¿Qué nota tendrán como mínimo los estudiantes que se beneficien de dichas ayudas?

3) En un barrio de determinada ciudad se da la siguiente distribución de la renta:

Renta en 10^3 €	Nº de familias
3 - 9	40
9 - 15	30
15 - 21	20
21 - 40	8
40 - 65	2

a) Obtenga el promedio que considere más adecuado. Razone su elección.
b) Si se define la clase media como el grupo de familias que tienen rentas comprendidas entre un 10% por debajo y un 30% por encima de la media, ¿Qué número de familias componen dicha clase media?.
c) En otro barrio, la renta media por familia es de 12.000 € y la desviación típica es de 2.400 € ¿Cuál de los dos barrios puede considerarse más homogéneo en rentas familiares?
d) ¿Qué nivel de renta tienen como mínimo las familias de mayores ingresos que perciben el 30% del total de las rentas del barrio?. ¿Cuántas familias componen este grupo?.
e) Estudie analítica y gráficamente la concentración de la renta en el barrio y extraiga conclusiones.

4) La población total de la provincia de Málaga en 2001 era 1.302.240 habitantes y estaba distribuida de la siguiente forma:

Tamaño de municipios (10^3 habitantes)	Número de municipios	Porcentaje De población
0 – 2	45	3,2
2 – 10	37	10,1
10 – 50	14	29,0
50 – 100	2	8,2
Más de 100	2	49,5

Se pide:
a) Calcule el tamaño medio, mediano y modal de los municipios de la provincia.
b) ¿Qué porcentaje de la población vive en municipios de menos de 70.000 habitantes?

5) Una empresa de hostelería presenta la siguiente distribución de salarios por hora:

Salarios por hora (Euros)	Nº de empleados
5 - 8	10
8 - 10	15
10 - 12	10
12 - 15	
15 - 18	5

Si la empresa tiene un total de 45 trabajadores obtenga:

a) El salario más frecuente y el salario medio por hora.
b) Calcule e interprete una medida de la asimetría de la distribución de los salarios por hora en la empresa.
c) En 1996 debido a la crisis del sector la dirección de la empresa reduce en un 20% los salarios. Calcule el salario medio tras dicha reducción sin calcular la nueva distribución.
d) Suponga ahora que la crisis fue de tal magnitud que en vez de reducir los salarios la decisión que se toma finalmente es la de reducir el número de empleados de cada intervalo en un 20%. ¿Cuál es ahora el salario medio? ¿Y si se redujeran los salarios y el número de empleados simultáneamente?

6) En un convenio laboral se acuerda subir un 10% el volumen total de salarios. Un empresario con 250 empleados les paga un total de 200.000 € mensuales. Se sabe además que dicha variable presenta una desviación típica de 120 €. La subida del 10% representa para el empresario un incremento total de las nóminas del personal de 20.000 € mensuales. Dicha subida puede ser:

a) Proporcional: Se aumenta el sueldo de cada empleado en un 10%
b) Lineal: Cada empleado percibe un aumento de 80 €. (20.000/250)

¿Qué alternativa conduce a reducir las diferencias salariales?

7) Se ha obtenido la siguiente información sobre la renta anual de las 1000 familias de un determinado municipio andaluz.

Renta familiar (Euros)	Nº de familias
Menos de 6.000	200
6.000- 18.000	
18.000 – 42.000	
Más de 42.000	

Además sabemos que:

a) Las rentas familiares en el municipio oscilan entre un mínimo de 3.600 € y un máximo de 60.000 €.
b) Hay 320 familias con rentas inferiores a 18.000 € y 200 familias cuyas rentas superan los 42.000 €.

Se pide:

a) ¿Qué porcentaje de familias obtienen rentas superiores a la media?
b) ¿Cuál es la renta mínima del 40% de las familias con mayores ingresos?
c) Dé una medida de la concentración de la renta en el municipio e interprete su significado.

8) Se han seleccionado 1000 alumnos de la Universidad de Málaga que presentan la siguiente distribución de notas en una determinada asignatura:

Notas	% de alumnos
0 - 5	32
5 - 7	38
7 - 9	21,4
9 - 10	8,6

Se pide:

a) ¿Supera el aprobado la nota más frecuente?
b) ¿Cuál es la nota mínima del 60% de los alumnos con mejores notas?
c) ¿Qué porcentaje de alumnos obtiene una puntuación superior a 8?
d) Si la nota media de la asignatura en otra Universidad es 6,5 con varianza 10,35 ¿cuál de las dos Universidades tiene una nota media más representativa?

9) La distribución de las empresas del sector alimenticio de la provincia de Málaga según su tamaño (nº de empleados) es la siguiente:

Empleados	Nº de empresas
0 - 4	45
4 - 10	35
10 - 20	12
más de 20	8

Sabiendo que el volumen total de empleados de las empresas con más de 20 trabajadores es de 400 se pide:

a) Calcule el promedio más adecuado para representar el comportamiento de las empresas del sector. Argumente la elección realizada.
b) ¿Cuál es el tamaño de empresa más frecuente?
c) La Junta de Andalucía ha decidido conceder subvenciones a las empresas más pequeñas, con el fin de fomentar su competitividad en la Unión Europea. Dada la limitación de recursos con la que cuenta dicha institución, sólo podrán beneficiarse de dichas ayudas el 25% de las empresas más pequeñas. ¿Cuántos empleados tendrán, como máximo, las empresas que reciban dichas subvenciones?

10) La distribución de la Renta Familiar Disponible (RFD) en España en el año 1993 se recoge en la siguiente tabla:

RFD (10^{12} ptas.)	Nº provincias
Menos de 0,52	23
0,52 –1,04	16
1,04 – 1,56	9
1,56 – 3,21	2
3,21 – 6,84	2

a) Si clasificamos a las provincias en cuatro grupos de igual tamaño de menor a mayor nivel de renta, ¿en qué grupo se encuentra la provincia de Málaga con un nivel de renta de 1,16 billones de ptas.?, ¿entre qué valores de rentas estará comprendido dicho grupo?
b) Calcule la renta familiar disponible media y dé una medida de su representatividad.
c) ¿Cuál es la renta familiar disponible más frecuente?.
d) ¿Qué número de provincias de rentas más bajas se lleva el 50% de la renta nacional?

11) Un país tiene dos regiones con la misma población. Se han tomado sendas muestras sobre las rentas percibidas, obteniéndose la siguiente información:

Región A		**Región B**	
Renta (Euros)	**Nº de Familias**	**Renta (Euros)**	**Nº de Familias**
600 - 1200	24	*300- 900*	10
1200 - 1800	36	*900 - 1500*	42
1800 - 2500	20	*1500 - 2000*	35
2500 - 4000	20	*2000 - 3000*	20
4000- 6000	50	*3000 - 5000*	13

a) Determine la renta media de las muestras de cada región y la renta media para el conjunto del país. ¿Cuál de las dos rentas medias es más representativa?. ¿Por qué?.
b) Si en la región B clasificamos a una familia en el grupo en que se encuentra el 50% de las menos favorecidas, ¿cuál sería el nivel máximo de renta que podría percibir sin salir de ese grupo?.
c) Calcule una medida de la asimetría de la distribución de la renta en la región A.

12) Se les ha preguntado a 1000 establecimientos de comercio minorista de una determinada ciudad por sus ventas anuales de un cierto producto de alimentación. Los resultados obtenidos son los que refleja la tabla siguiente:

Ventas (€)	**Establecimientos**
Hasta 600	400
De 600 a 1500	225
De 1500 a 3000	175
De 3000 a 6000	120
De 6000 a 9000	75
Más de 9000	5

Además se sabe que el total de ventas para ese producto ascendió a 2 millones de euros.
Con toda esa información:

a) Obtenga la media aritmética de esa variable. Analice su representatividad en términos absolutos y realtivos. Observando la distribución de frecuencias, ¿considera que es el promedio más adecuado?
b) Determine el procentaje de establecimientos, de entre los que menos venden, cuyas ventas acumuladas represetan la cuarta parte de las ventas totales. Determine también el valor de la variable que deja a su derecha el 10% de los establecimientos que más venden.
c) Si al año siguiente el precio de ese producto aumenta un 5%, determine cual sería el nuevo volumen medio de ventas así como su desviación estándar.

13) En una empresa trabajan 20.000 personas, cuyos salarios mensuales se distribuyen según figuran en la tabla adjunta:

Salarios (€)	**Nº de empleados**
600-1200	12.000
1200-2400	6.000
2400-3000	1.000
3000-6000	800
Más de 6000	200

Se sabe que la masa salarial total es de 30 millones de euros. Se pide:

a) Obtenga el promedio que considere más adecuado, razonando la respuesta.
b) Estudie la dispersión de esta distribución usando una medida relativa. Comente el resutlado.
c) Estudie la concentración en la distribución de los salarios utilizando el índice índice correspondiente.

14) La distribución los hogares según su tamaño, medido por el número de sus miembros, de un determinado barrio es la que se recoge en la tabla siguiente:

Tamaño del hogar	Nº de hogares
1	40
2	70
3	110
4	90
5	48
6	42
7	40
8	35
9	20
10	8

A partir de esta información dertermine:
a) ¿Cuál es el tamaño medio del hogar?
b) ¿Cuál es el tipo de familia más frecuente?

c) Si solo hubiera plazas de aparcamiento para el 50% de las familias y éstas se asignaran a las de mayor tamaño, ¿a partir de qué tamaño de familia se le asignarían plaza de garage?
d) Si en otro barrio el coeficiente de variación es 1, ¿en cuál de los dos barrios la media es más representativa?

Responda a las cuestiones anteriores y compare los resultados si los datos se hubieran agrupado en la forma siguiente:

Tamaño del hogar	Nº de hogares
De 0 a 2	110
De 2 a 4	200
De 4 a 6	90
De 6 a 8	75
De 8 a 10	28

15) Una determinada empresa de alimentación presenta la siguiente distribución de salarios mensuales entre hombres y mujeres:

Hombres		Mujeres	
Salarios (€)	Nº personas	Salarios (€)	Nº personas
0-600	50	0-600	30
600-1200	10	600-1200	5
1200-3000	2	1200-2400	1

a) ¿Qué grupo percibe mayor salario, por término medio? ¿En cuál de ellos es superior el salario más frecuente?
b) Compare la dispersión de los salarios de hombre y mujeres
c) Si se le pregunta a un hombre por su salario y contesta que es 1200 euros, mientras que el de una mujer es 1100 euros, en términos relativos ¿cuál de los dos trabajadores está mejor retribuido?
d) Compare la concentración de ambas distribuciones.

Resultados de la Relación de Ejercicios nº 1

2. a) 5,417 . La especialidad B (CV_A = 55,4% ; CV_B = 35,43%).
   b) 5,91
   c) El de la especialidad A (Z_A=0,588 ; Z_B=0,02).
   d) 9,33

3. a) 11.000 €.
   b) 23 familias.
   c) El segundo barrio (CV = 20%).
   d) 20270 €.; 12 familias.
   e) G = 0,325

4. a) $\overline{X}$ = 13022 habitantes. M_e=3081 habitantes. M_o=2.000 habitantes.
   b) El 45,58%.

5. a) M_o= 9 €/hora; $\overline{X}$ = 10,22 €./hora.
   d) A_P= 0,4011
   c) 8,18 €./hora
   d) 10,22 €./hora; 8,18 €./hora.

6. La subida lineal, debido a que implica un menor coeficiente de variación.

7. a) Un 50%, $\overline{X}$ =27000 €.
   b) 32000 €.
   c) G = 0,3122

8. a) No (Mo=6,25)
   b) 5,42
   c) Un 19,3%
   d) La de Málaga, ya que tiene un coeficiente de variación menor.

9. a) Cinco empleados (aproximadamente)
   b) Cuatro empleados.
   c) Dos empleados (aproximadamente)

10. a) En el cuarto grupo; entre 1,04 y 6,84 billones de ptas.
    b) $\overline{X}$ = 0,87 billones de ptas.; CV= 112,76%
    c) M_o= 0,52 billones de ptas.
    d) 42 provincias (aproximadamente)

11. a) $\overline{X}_A$= 2890,6 €. $\overline{X_B}$ = 1830,4 €. $\overline{X_T}$ = 2419,4 €.
    La de la región B (CV_B=50,42%) es más representativa.
    b) 1914 €.
    d) A_p = 0,88

12. a) $\overline{X}$ = 2000 €. S = 2858,3 € CV = 142,9 %
    b) El 69 %, aproximadamente. 5500 €.
    c) Nueva media = 2100 €. Nueva desviación = 3001,25 €.

13. a) Me = 1100 €.
    b) S = 1218,6 € CV = 81,24 %
    c) G = 0,295

14. a) Media = 4,34 personas; b) Mo = 3 personas; c) Me = 4 personas; d) S = 2,28 personas, CV = 52,58 %; e) media 4,35 personas, Mo = 2,9 personas, Me = 3,415 personas, S = 2,29 personas, CV =52,69 %.

15. a) Media(H) =454,8 €, Media(M) =311,25 €, Mo(H) = 600 €, Mo(M) = 600 €
    b) S(H) =372,3 €, S(M) = 311,25 €, CV(H) =81,85 %, CV(M) = 73,23 %
    c) z(H) = 2,00, z(M) = 2,17
    d) G(H) = 0,288, G(M) = 0,253.

INTRODUCCIÓN A LA ESTADÍSTICA EMPRESARIAL

Relación de Ejercicios nº 2 (tema 5) ***Curso 2002/2003***

1) Las ciento treinta agencias de una entidad bancaria presentaban, en el ejercicio 2002, los siguientes datos correspondientes a las variables:
X: Saldo medio de las cuentas a 31/12 (en euros).
Y: Cuentas a plazo / Total cuentas

X \ Y	Menos de 0'1	De 0,1 a 0'3	Más de 0'3
Menos de 120	48	--------	--------
De 120 a 300	21	11	--------
De 300 a 600	7	8	2
De 600 a 1500	7	5	1
Más de 1500	6	k	1

Calcule:
a) El valor de k.
b) Las distribuciones marginales de X e Y.
c) La mediana de X y el tercer cuartil de Y.
d) La distribución de las agencias, en términos relativos, según X, cuando la ratio Y se encuentra comprendida entre 0,1 y 0,3.
e) Distribución, en términos relativos, de Y para agencias con saldo medio por encima de las 100.000 ptas.

2) La siguiente tabla de doble entrada representa la distribución de dos variables X e Y para un grupo de 40 observaciones.

X \ Y	9	7	5
2 – 6	3	8	4
6 – 10	7	1	5
10 – 16	k	2	9

Se pide:
a) Obtener el valor de k. ¿Qué valor de la variable X deja a su derecha el 38% de las observaciones?.
b) Obtener la media de la distribución X y valorar su representatividad.
c) Determinar la distribución de f. relativas de la variable Y condicionada a un valor de X mayor o igual que 6.
d) ¿Son variables independientes?. Razone la respuesta.

3) El ingreso anual disponible y los gastos de consumo de 12 familias, seleccionadas aleatoriamente, en una zona de nuestra ciudad han sido en miles de euros los siguientes:

Gastos de consumo	Ingreso anual disponible
7	8
12	15
18	20
30	35
20	28
24	25
8	8
11	13
8	7
6	6
10	12
18	15

Haciendo uso de esta información, responda a las siguientes cuestiones:

a) Ajuste por mínimos cuadrados una recta en la que el consumo sea función de los ingresos.
b) Comente el significado económico de los coeficientes de la recta ajustada.
c) Proporcione una medida de la bondad del ajuste.
d) Halle el incremento porcentual que experimentaría el consumo de una familia cuyos ingresos fueran de 8000 euros, ante un incremento porcentual unitario de estos últimos.
e) ¿Qué valor alcanzaría la propensión marginal al consumo si las variables consideradas vinieran expresadas en miles de dólares?.
f) Exprese la recta ajustada tomando como unidad el millón de pesetas.

Nota: Utilice la aproximación 6 euros = 1000 ptas.

4) De una distribución de dos variables X e Y se conocen los siguientes datos:

$$r = 0'9; \qquad S_x = 1'2; \quad S_y = 2'1; \qquad \bar{x} = 5; \qquad \bar{y} = 10$$

a) Calcule la recta de regresión de Y sobre X.
b) Valore la bondad del ajuste.
c) Siendo la variable Z = 2 X+3, obtenga la recta de regresión de Y sobre Z.

5) De las estadísticas de gasto en Marketing en millones de ptas. (X) y las ventas en millones de ptas. (Y) de una empresa se han tenido los siguientes datos relativos a los seis últimos años (1991-1996):

$$\sum \boldsymbol{Y}_j = 3.580 \qquad \sum \boldsymbol{X}_i = 25'7 \qquad \sum\sum \boldsymbol{X}_i \boldsymbol{Y}_j = 15.547 \qquad \sum \boldsymbol{X}_i^2 = 115'01 \qquad \boldsymbol{X}_{96} = 5'5$$

a) Estime la recta de regresión donde las ventas sean función del gasto en marketing.
b) ¿Cuál sería el incremento porcentual que experimentarían las ventas si el gasto medio en marketing aumentara en un 1%?.
c) Estime las ventas para 1997, si se espera aumentar el gasto en marketing en un 2% con respecto al año anterior.

6) Mediante un ajuste mínimo cuadrático se ha observado que la relación entre el precio en euros (X) y la cantidad demandada en Kilogramos (Y) de un cierto artículo, viene dada por el modelo X* = 1 - 0'154Y.
Sabiendo que la distribución de Y tiene una media de 2 Kg. y una desviación típica de 12 Kg., mientras que la de X tiene una desviación típica de 3euros.
a) Obtenga la recta de regresión de la cantidad respecto al precio e indique cuál es el tanto por ciento de las variaciones en la cantidad demandada que no viene explicada por el precio.
b) Calcule la elasticidad media demanda-precio y explique su significado.

7) La recta de regresión de Y (demanda de un producto en Kilogramos) con respecto a X (precio de ese producto en euros), viene dada por :

$$Y^* = 4'5 - 0'2*X \text{ , siendo } R^2 = 0'81.$$

a) ¿Cuál es el significado del coeficiente de regresión?.
b) ¿Cuál es el valor del coeficiente de correlación lineal entre X e Y, r_{XY}?. Comente el significado de su valor.
c) Según la ecuación anterior, ¿Cuál es la demanda del producto cuando su precio es de 10 euros/Kg.?. Comente la fiabilidad de este resultado, sabiendo que los valores observados de X estaban comprendidos entre 3 y 7 euros/Kg.

8) A partir de la siguiente tabla de doble entrada y teniendo en cuenta que la población observada consta de 40 elementos, se pide:

X Y	1 - 3	3 - 6	6 - 8	8 -12
-2	3	1	1	2
0	2	---	---	3
3	---	1	1	4
4	2	4	3	5
6	1	---	5	K

a) Determinar el valor de K.
b) ¿ El coeficiente de correlación lineal, r_{XY} , es –1?. Justifique la respuesta.
c) Calcule la mediana de X.
d) Calcule la moda de Y.
e) ¿Qué valor de la variable Y deja a su derecha el 27% de la población?.
f) ¿ Son independientes?

9) En relación a los precios (p) y cantidades (q) demandadas de un bien, se dispone de la siguiente información:

- El precio medio es de 10 euros y su varianza de 11'6.
- La media y la varianza de la cantidad demandada es de 57 toneladas y 1096 toneladas^2 respectivamente.
- Para los valores medios indicados, cuando el precio se incrementa en un 1%, la cantidad demandada disminuye en un 1'65%.

Determine:

a) La función de demanda q = a + bp.
b) El porcentaje de la variación de la demanda que viene explicado por el precio.
c) La demanda esperada si el precio del bien fuese de 20euros. Comente la fiabilidad de dicha predicción.

10) Se han estudiado las calificaciones de cien alumnos en la asignatura de Matemáticas y las horas mensuales dedicadas al estudio, obteniéndose los siguientes resultados: $\bar{x}=110$ $\bar{y}=6$ $S_x=10$ $S_y=0'5$

Además, se sabe que el coeficiente de correlación lineal entre estas variables es de 0'85.Se pide:

a) ¿Qué porcentaje de las calificaciones no viene explicado por las horas dedicadas al estudio?.
b) ¿Qué nota se puede predecir para un alumno que ha dedicado 125 horas mensuales al estudio?.

11) Se quiere estimar el modelo lineal que relaciona la demanda de un bien (Y) con la renta disponible (X). Para un conjunto de observaciones sobre aquellas variables, se han obtenido los siguientes resultados:

$$\bar{x} = 30 \; ; \; \bar{y} = 13 \; ; \; S^2_x = 200 \; ; \; S^2_y = 53 \; ; \; S_{xy} = 102$$

Se pide:

a) Estimación de los parámetros del modelo Y* = a + bx, indicando su significado económico.
b) Coeficiente de determinación. Descomponer la varianza de Y en varianza explicada por la regresión y varianza residual.
c) ELasticidad demanda-renta para una renta disponible de 25 unidades monetarias. Comentar el resultado.

12) Para el periodo 1994-1998 (ambos años incluidos) se dispone de la información sobre inversión en una determinada provincia (variable X en cientos de millones de pesetas) y empleo generado (variable Y en cientos de personas).

$$\sum Y_i = 593 \qquad \sum X_i = 174 \qquad \sum X_i^2 = 6.190 \qquad \sum Y_i^2 = 72.469 \qquad \sum X_i Y_i = 21.154$$

Se pide:

a) Ajuste e interprete el modelo de regresión lineal donde el empleo generado sea función de la inversión.
b) Si la inversión prevista para 2003 es de 24. millones de euros, ¿cuántos empleos se esperan crear?. ¿Es fiable dicha predicción?.
c) ¿Qué incremento porcentual de empleo se produce si la inversión media aumenta en un 1% en la provincia?.

Años	Fondos (10^3pts)
1951	23′411
1952	25′207
1955	30′132
1956	31′416
1957	33′244
1958	36′329

13) En un intento de reconstruir las series de dotaciones a fondos de autoseguros que destinaba una gran compañía en los años cincuenta, sólo han podido recopilarse los datos de 6 años.

a) Supuesta una tendencia lineal de la serie, estime las cifras correspondientes a 1950, 1953 y 1954. Dé una medida de la bondad del ajuste.

b) Determine el crecimiento medio anual de la serie en términos porcentuales.

14) Los últimos datos disponibles de la serie de Parados en el sector servicios de Andalucía según el INE han sido los siguientes:

Período	Parados en servicios en Andalucía (103 persones)
1999. 4º trimestre	224’43
2000. 1º trimestre	219’51
2000. 2º trimestre	216’75
2000. 3º trimestre	207’96

a) ¿Cuál ha sido la variación porcentual entre los últimos dos trimestres considerados?.
b) ¿Y la variación porcentual durante el periodo considerado?.
c) ¿ Cuál ha sido la tasa trimestral media de reducción del paro?

Soluciones de los Ejercicios del tema 5

1. a) k= 6
   c) Me (x) = 35.938 ptas; Q3 (y) = 0´11

2.- a) K = 1, P62 =9.01
   b) $\bar{x}$ = 8 C.V = 47.05%
   d) Dependientes

3.- a) y* = 1´658 + 0´792 x
   b) a: consumo autónomo. b: propensión marginal al consumo.
   c) R^2 = 0'92
   d) 0´792%
   e) La propensión marginal al consumo no cambiaría
   f) y* = 1´658/6 + 0´792 x

4.- a) y*= 2´125 + 1´575 x
   b) R^2= 0´81
   c) y*= - 0´2375 + 0´7875 z

5.- a) y* = 408´25 + 44´02 x
   b) 0´32%
   c) 655´20 millones de ptas.

6. a) y* = 3´705 - 2´464 x
   b) -0´85%

7.- a) Disminuye la demanda 0.2 unidades por cada incremento del precio en un euro
   b) – 0.9 . La recta se ajusta “bastante bien” a la nube de puntos de los datos (X,Y) observados.

8. a) k=2
   b) No. Una justificación (sin cálculos): a un valor de X corresponden varios de Y.
   c) Me (x) = 4
   d) Mo (y) = 7´3333
   e) 9´3
   f) No son independientes

9. a) q*= 151´05 - 9´405 q
   b) 93´6%
   c) -37´05 Tm.

10. a) 27´75%
   b) 6´64

11. a) $y^* = -2´3 + 0´51\ x$
b) $R^2 = 0´98$; $S^2y = 53 = (S^2y^* = 51´94) + (S^2e = 1´06)$
c) 1´22%

12. a) $y^* = -15{,}025 + 3{,}84\ x$.
b) 13.857; $R^2 = 0{,}929$
c)1,13%.

13.- a) Fondos 1950= 21.547 ptas.; Fondos 1953= 26.767 ptas.; Fondos 1954= 28.507 ptas.; $R^2 = 0.989$.
b) Crecimiento medio anual: 1.740 pesetas.

14. a) Disminución del4’06%.
b) Disminución del7’34%.
c) 3’6%

INTRODUCCIÓN A LA ESTADÍSTICA EMPRESARIAL

Relación de Ejercicios nº 3. (Tema 6) *Curso 2002/2003*

1. El salario mensual medio en euros de un tipo de administrativos en los años 1993 a 2002 ha sido:

1993	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
875	883	902	936	962	962	1000	1043	1044	1050

a) Elabore un índice que muestre las variaciones de los salarios a partir del año base 1993.
b) ¿Cuál sería la serie de índices simples con base en 2000?.
c) Elabore un índice que muestre las variaciones inter-anuales de los salarios. ¿En qué año se registra el mayor crecimiento?.
d) Si el índice de precios mostró un incremento del 10% en el período 1993-1997, ¿cuánto percibieron en términos reales este tipo de empleados en el año 1997?.

2. En la tabla adjunta figura el índice simple que muestra la evolución del Valor Añadido Bruto (VAB) andaluz para el período 1995-1999. Dicho índice se ha confeccionado a partir de una serie a precios corrientes. Asimismo se proporciona el deflactor implícito del PIB (índice de precios).

Años	*Índices simples del VAB andaluz (precios corrientes)*	*Deflactor VAB andaluz*
1995	100,0	96,7
1996	105,9	99,2
1997	112,7	100,0
1998	120,6	101,8
1999	130,6	105,0

Haciendo uso de dicha información y sabiendo que en 1995 el PIB andaluz se estimó en 53.521.372 miles de euros corrientes, se pide:

a) Obtener la serie completa del VAB andaluz en miles de euros a precios corrientes.
b) Obtener los índices en cadena del VAB a precios corrientes.
c) Obtener la serie del VAB andaluz a precios de 1995. Comente los resultados.

3. Los gastos familiares en sanidad, expresados en millones de euros, durante los años 1986 a 1990 y el índice de precios referido a esos años, están recogidos en la tabla adjunta.

Año	*Gasto fam. sanidad*	*Í. Precios (1992=100)*
1997	20’1	153’6
1998	21’5	166’2
1999	22’1	181’6
2000	22’4	200’5

2001	23'3	222'5

Se pide:

a) Obtenga los incrementos inter-anuales del gasto a precios corrientes.
b) Calcule el incremento del gasto a precios constantes de 1997 para el período 1997-2001. Comente los resultados.
c) En 2002 el gasto en sanidad se incrementa un 10% en términos corrientes y se produce una subida de los precios de estos bienes y servicios del 5%. ¿qué incremento porcentual experimentará el gasto en sanidad en términos reales entre 2001 y 2002?.

4. Dadas las siguientes tres series de índices de precios al por mayor, calcule los índices del período 1993-2001, tomando como base 1999.

Años	*Índice (base 1988)*	*Índice (base 1994)*	*Índice (Base 1999)*
1993	115		
1994	122	100	
1995		104	
1996		105	
1997		108	
1998		114	
1999		121	100
2000			107
2001			112

5. Conocidos los precios y cantidades de los artículos de consumo correspondientes a los años 1999 a 2001, determine con base en 1999 los siguientes índices complejos:

a) Los índices de precios mediante la fórmula de Laspeyres, Paasche y Fisher.
b) Los índices de cantidades mediante la fórmula de Laspeyres, Paasche y Fisher.
c) Los índices de valor.

Años	*Artículo A*		*Artículo B*		*Artículo C*	
	Precio	*Cantidad*	*Precio*	*Cantidad*	*Precio*	*Cantidad*
1999	2	10	5	12	10	3
2000	2	12	6	10	11	2
2001	3	15	6	5	12	3

6. Los índices elementales de precios son los de la tabla adjunta. Se sabe que el valor de estos índices con cantidades de 2002 y precios de 1999 es de 2, 3 y 5. Hallar el índice complejo de precios de Paasche para 2002.

Años	I^A	I^B	I^C
1999	100	100	100
2000	103	102	103
2001	107	108	109

2002	109	111	112

7. Una asociación de empresas de alimentación está estudiando la evolución de la producción agrícola en una comarca y cuenta con la siguiente información:

Años	***Índice simple de la producción (€ corrientes)***	***Índice de precios agrícolas (base 1991=100)***
1997	100	200
1998	105'4	260
1999	107'0	295
2000	108'8	360
2001	113'0	400

a) Sabiendo que en 2001 la producción agrícola ascendió a 565 miles de euros, obtenga las serie de la producción agrícola en euros corrientes.
b) Estime la tendencia lineal para la producción agrícola en euros constantes de 1991. Comente el resultado de los parámetros y realice una predicción para 2002.

8. De un sistema de índices de precios de consumo se tiene la siguiente información estadística sobre los grupos de artículos que componen la cesta de la compra:

Grupo	W_i (%)	$I_{i\ 93}^{94}$	$I_{i\ 93}^{95}$	$I_{i\ 93}^{96}$	$I_{i\ 93}^{97}$
I. Alimentación	40	108	119	127	130
II. Vestido	9	113	120	132	140
III. Vivienda	17	110	114	119	125
IV. Transporte	20	121	132	146	154
V. Otros	14	103	110	113	118

Se pide:

a) Obtener la serie del IPC para el período 1993-1997, sabiendo que el índice usado es de tipo Laspeyres, y que las ponderaciones corresponden al año 1993.
b) Evaluar la variación porcentual de los precios entre 1994 y 1996, y entre 1995 y 1997.
c) Si el alquiler de un piso se pactó en 1995 en 100.000 pesetas, ¿cuál será su valor actualizado en 1997 de acuerdo con la evolución de los precios en el grupo correspondiente?. ¿Y si se pactó una revisión anual según el IPC?. Interprete los resultados desde el punto de vista del arrendador y el inquilino.
d) Si una persona ganaba 115.000 pesetas al mes en 1995, ¿cuánto debería ganar en 1997 para no perder poder adquisitivo?.

9. Una empresa se dedica a la distribución de dos tipos de refrescos, A y B. Las cantidades vendidas así como el precio de los mismos en los años 2000 y 2002 aparecen en la tabla adjunta:

	Precios (€ por litro)		***Cantidades (litros)***	
	2000	2002	2000	2002

Refresco A	1,25	1,75	10.500	35.000
Refresco B	1,00	1,25	8.000	10.000

A partir de dicha información determine:

a) ¿Cuál fue la variación porcentual de la cantidad vendida del refresco A entre 2000 y 2002?. ¿Y la de la cantidad del refresco B?.
b) Para los dos refrescos, calcule e interprete un índice complejo de cantidad de Laspeyres para 2002 con base 2000.
c) De igual forma, calcule e interprete un índice complejo de valor para 2002 con base 2000.

10. Disponemos de la siguiente información relativa a índices de precios al consumo (IPC):

Grupos	Índices (1996)	Ponderaciones	Índices 31-XII-97
1. Alimentos, bebidas y tabaco	100	405'20	135
2. Vestido y calzado	100	81'65	131
3. Vivienda	100	140'05	132
4. Menaje y servicios del hogar	100	77'50	121
5. Serv. médicos y conservación de la salud	100	33'75	122
6. Transportes y comunicaciones	100	97'45	125
7. Esparcimiento, cultura y enseñanza	100	69'45	128
8. Otros gastos de consumo	100	94'45	124
Total	100	1.000'00	

a) ¿Cuál ha sido el incremento del IPC en el año 1997?.
b) Halle las repercusiones y participaciones de cada uno de los grupos del IPC en la variación sufrida en el índice global a 31-XII-97.
c) ¿Cuál es el grupo más afectado por la subida de precios?.

11. Uno de los criterios establecidos por el Tratado de Maastricht de acceso a la Unión Monetaria consideraba que la tasa de inflación no debía superar la media de las tasas de variación anual de los tres países menos inflacionistas de la Unión Europea más punto y medio. Comprobar, a partir de los datos del índice de precios de consumo armonizado (IPCA) que figuran en las siguientes tablas, que España cumplía este criterio en enero de 1998.

Índices de Precios de Consumo armonizado para España. Base 1996, enero 1998

Grupos	Índice 1998 Base 96	Ponderación 1996 (%)	% variación en un año (98/97)
1. Alimentación y bebidas no alcohólicas	101'1	27´5	0'7
2. Bebidas alcohólicas y tabaco	116'5	3'2	10'8
3. Vestido y calzado	102'9	11´4	1'5
4. Vivienda	104'5	11´2	2'5
5. Menaje	102'5	6´5	1'4
6. Medicina	102'0	0'8	0'6
7. Transporte	102'9	14´6	0'8
8. Comunicaciones	100'5	1´6	0'3
9. Ocio y cultura	103'2	6´9	2'8
10. Enseñanza	105'5	0'1	2'5
11. Hoteles, cafés y restaurantes	104'8	11'8	3'0
12. Otros	102'9	4'4	1'7

Índices de precios de consumo armonizado. Base 1996, enero 1998

Índice de los países de la Unión Europea.

Países	Índices 1998 Base 96	Tasa anual 98/97 (%)
Alemania	101'7	0'8
Austria	101'7	1'1
Bélgica	101'8	0'5
Dinamarca	102'4	1'7
España		
Finlandia	101'9	1'8
Francia	101'3	0'6
Grecia	107'1	4'3
Holanda	102'0	1'6
Irlanda	101'5	1'2
Italia	103'1	1'9
Luxemburgo	102'2	1'5
Portugal	102'7	1'6
Reino Unido	102'1	1'5
Suecia	102'5	2'1
Unión Europea	102'2	1'3

12. El salario/hora de los trabajadores de un determinado sector productivo, a precios corrientes, figura en la tabla adjunta, en la que aparece igualmente el índice de precios al consumo para esos años:

Años	Salario/Hora	IPC
1996	52	140
1997	58	162
1998	60	175
1999	63	190
2000	64	200
2001	84	205

Se pide:

a) Determinar la evolución real del salario/hora a precios constantes de 1996.
b) Determinar la evolución porcentual del salario/hora en términos monetarios y en términos reales entre 1996 y 1998.
c) Calcular los índices en cadena de la serie salario/hora a precios corrientes.

13. Los salarios de un grupo de trabajadores han aumentado de un año a otro en un 50%, mientras que el índice de precios al consumo lo ha hecho en un 10% para ese mismo período de tiempo. ¿Cuál ha sido el cambio real en la capacidad adquisitiva de los trabajadores?.

14. Dados los índices en cadena para los cuatro trimestres de 1994, hallar el porcentaje de variación de la ocupación hotelera entre el primer y el último trimestre del año.

TRIMESTRES	I.C.
1	-
2	120
3	150
4	50

15. Dada la serie de índices simples de precios de cinco productos para el año 1990, con base en 1986:

110, 115, 117, 107 y 112,

y sabiendo que los valores de las cantidades consumidas en 1986 de los productos considerados a precios del mismo año fueron;

6, 8, 10, 11 y 14,

respectivamente, calcular el correspondiente índice complejo ponderado de precios. ¿A qué tipo corresponde a Laspeyres o a Paasche?.

16. Una fábrica de equipo industrial ha tenido las siguientes ventas anuales en los últimos seis años, expresadas en miles de euros:

Años	***Ventas***	***Deflactor***
1997	540	105
1998	565	109
1999	580	109
2000	610	118
2001	625	123
2002	660	140

a) Determine las tasas de variación inter-anuales de las ventas en términos corrientes.
b) A fecha 1-1-2003 dicha fábrica desea saber cuál es la valoración total en euros de 2002 de las ventas en el período considerado.

17. Sabiendo que las ponderaciones porcentuales del índice de precios al consumo (IPC) de un país son las siguientes:

Grupos	***Ponderación***
1. Alimentos, bebidas y calzados	29'4

2.	Vestido y calzado	11’5
3.	Vivienda	12’3
4.	Menaje y servicios para el hogar	6’6
5.	Medicina y conservación de la salud	3’1
6.	Transporte y comunicaciones	16
7.	Esparcimiento, enseñanza y cultura	7’3
8.	Otros bienes y servicios	13’8

a) Calcule el IPC correspondiente a 1996 con base en 1994, si el IPC para 1995 con base en 1994 fue de 108 y los datos sobre los incrementos/disminuciones porcentuales de precios en 1996 con respecto a 1995 fueron los siguientes para cada uno de los grupos:

Grupo	1	2	3	4	5	6	7	8
Incremento/disminución	11%	12%	8%	-5%	8%	-2%	11%	9%

b) ¿Cuál sería la pérdida o ganancia porcentual del poder adquisitivo de un salario que en 1994 era de 120.000 pesetas mensuales y en 1996 asciende a 123.000 pesetas?.

18. Las ponderaciones de los doce grupos del Índice de Precios al Consumo armonizado (IPCA) de la Unión Europea, base 1996, junto con los incrementos porcentuales de precios experimentados, en el año 1997, en un determinado país se recogen en la siguiente tabla:

Grupos	Ponderación (%) 1996	Incrementos 1997/96
1. Alimentación y bebidas no alcohólicas	27´5	2’1
2. Bebidas alcohólicas y tabaco	3’2	1’6
3. Vestido y calzado	11´4	1’5
4. Vivienda	11´2	3’0
5. Menaje	6´5	2’5
6. Medicinas	0’8	0’9
7. Transporte	14´6	1’9
8. Comunicaciones	1´6	2’8
9. Ocio y cultura	6´9	2’1
10. Enseñanza	0’1	2’5
11. Hoteles, restaurantes y cafetería	11’8	2’9
12. Otros	4’4	1’1

a) Calcule el IPCA de 1997, base 1996, del país considerado.
b) Obtenga e interprete la participación y repercusión del grupo de Comunicaciones en el índice general.

19. Se conocen los siguientes datos del Índice de Precios de Consumo Armonizado (IPCA) base 1996, referentes a España en enero de 1998.

Índices de precios de consumo armonizado. Base 1996, enero 1998

Índices nacionales de grupos

Grupos	Índice 1996	Ponderación (%) 1996
1. Alimentación y bebidas no alcohólicas	101’1	27’5
2. Bebidas alcohólicas y tabaco	116’5	3’2
3. Vestido y calzado	102’9	11’4

4. Vivienda	104'5	11'2
5. Menaje	102'5	6'5
6. Medicina	102'0	0'8
7. Transporte	102'9	14'6
8. Comunicaciones	100'5	1'6
9. Ocio y cultura	103'2	6'9
10. Enseñanza	105'5	0'1
11. Hoteles, cafés y restaurantes	104'8	11'8
12. Otros	102'9	4'4

y además:

Índices generales de los doce meses de 1997

	Índice 1997	% variación anual 1997/1996
Enero	101'3	2'8
Febrero	101'2	2'5
Marzo	101'3	2'2
Abril	101'3	1'6
Mayo	101'4	1'3
Junio	101'4	1'4
Julio	101'6	1'5
Agosto	102'1	1'7
Septiembre	102'6	1'9
Octubre	102'6	1'8
Noviembre	102'7	1'9
Diciembre	103'0	1'9

a) Calcule el IPCA de España para enero de 1998, base 1996, y la variación anual con respecto al mismo mes del año anterior.
b) ¿Cuál ha sido el grupo con mayor repercusión?.
c) Obtenga la serie mensual completa del IPCA desde julio de 1996 hasta enero de 1997.
d) ¿Cuál ha sido el mes de mayor inflación en el año 1997?.

Resultados de la Relación de Ejercicios nº 3

1. a) 100, 100'91, 103'09, 106'97, 109'94, 109'94, 114'29, 119'2, 119'31, 120.
b) 83'89, 84'66, 86'48, 89'74, 92'23, 92'23, 95'88, 100, 100'10, 100'67.
c) ---, 100'91, 102'15, 103'77, 102'78, 100, 103'95, 104'30, 100'10, 100'57.
Entre 1999 y 2000.
d) 874'6 €.

2. a)

Años	VAB corriente
1995	53.521.372
1996	56.660.194
1997	60.296.317
1998	64.535.040
1999	69.873.647

b)

Años	I.C.
1995	-
1996	105,9
1997	106,4
1998	107,0
1999	108,3

c)

Años	VAB constante
1995	53.521.372
1996	55.225.337
1997	58.316.223
1998	61.311.689
1999	64.339.013

3. a) ---, 6'97%, 2'79%, 1'36%, 4'02%.
b) Disminución del 19,9%.
c) 4'76%. Ojo con los decimales, puede salir un 3,5% si se realiza de otra forma el cálculo.

4. 77'9, 82'64, 85'95, 86'78, 89'25, 94'21, 100, 107, 112.

5.

	INDICE DE PRECIOS		
AÑOS	LASPEYRES	PAASCHE	FISHER
1999	100	100	100
2000	113'64	112'77	113'20
2001	125'45	130'59	127'99

	INDICE DE CANTIDADES		
AÑOS	LASPEYRES	PAASCHE	FISHER
1999	100	100	100
2000	85'45	84'80	85'12
2001	77'27	80'43	78'83

Años	INDICE VALOR
1999	100
2000	96'36
2001	100'91

6. 111'1.

7. a) 500; 527; 535; 544; 565.
b) serie en euros constantes de 1991: 250; 202,69; 181,35; 151,11; 141,25.
Ajuste: $y^* = 185,3 - 26,9t$; origen temporal t=0 en 1999; $y_{02}^* = 104,6$ miles de euros de 1991.

8. a) 100; 110,69; 119,58; 127,93; 133,17.

b) 15,6%; 11,36%.
c) 109.650 y 111.360 pesetas, respectivamente.
d) 128.064 pesetas.

9. a) 233,33%; 25%.
b) 254,44 (variación: 154%)
c) 349,11 (variación: 249%)

10 a) 30,2%.
b) repercusiones: 14,2; 2,53; 4,48; 1,63; 0,74; 2,44; 1,95; 2,27;
participaciones: 46,94; 8,38; 14,83; 5,39; 2,46; 8,06; 6,44; 7,50.
c) alimentación, bebidas y tabaco.

11. En España crecieron los precios un 1,83%. La tasa de variación anual de precios de los tres países menos inflacionistas fue 0,663; por tanto, la cota de inflación estaba en 3,13 (0,663 + 1,5) y el incremento de precios en España estuvo por debajo.

12. a) 52, 50'13, 48, 46'42, 44'8, 57,37.
b) Términos monetarios: Δ del 15'38%. Términos reales: ∇ del 7'7%.
c) ---, 111'54, 103'45, 105, 101'59, 131'25.

13. El aumento real en la capacidad adquisitiva de los trabajadores ha sido del 36'36%.

14. Se ha producido una disminución del 10%.

15. Precio de Laspeyres: 112'143.

16. a) ---, 4'63%, 2'65%, 5'17%, 2'46%, 5'60%.
b) 720, 724'36, 743'59, 723'61, 711, 660.

18. a) 102,2
b) 0,0448; 2,069%.

19. a) 103,2; 1,87%.
b) 100,099; 100,393; 100,687; 100,786; 100,785; 101,079.
c) Agosto y Septiembre con un incremento del 0,49%. El que menos ha sido Febrero con un –0,099%

INTRODUCCIÓN A LA ESTADÍSTICA EMPRESARIAL

Relación de Ejercicios nº4 (Tema7). ***Curso 2002/2003***

---

1) Una cooperativa de ambulancias tiene dos vehículos destinados al transporte de enfermos. Debido a la disponibilidad de los vehículos, por averías o por otras causas, se calcula que la probabilidad de que cada vehículo esté disponible cuando se necesita es 9/10. Si la probabilidad de una ambulancia es independiente de la otra, calcule la probabilidad de los siguientes sucesos:

a) Ambas ambulancias estén disponibles.
b) Ninguna esté disponible.
c) El servicio requerido sea satisfecho.

2) La probabilidad de que un estudiante A apruebe un examen de Estadística es de 0,8; la de otro estudiante B es 0,4; y la probabilidad de que aprueben los dos es de 0,3. Calcule la probabilidad de los siguientes sucesos:

a) Al menos uno de los dos aprueba el examen.
b) Ninguno aprueba el examen.
c) Sólo uno aprueba el examen.

3) Se ha realizado un estudio de mercado para conocer las preferencias de los consumidores entre tres marcas diferentes de detergentes (A, B, C). La investigación se ha realizado estudiando dos compras consecutivas en un grupo de consumidores. A partir de las siguientes tablas de probabilidad:

Probabilidades conjuntas

Marca comprada en la ocasión 1	Marca comprada en la ocasión 2			
	A2	B2	C2	
A1				0,431
B1				
C1				0,258
	0,431		0,237	

Probabilidades condicionadas
P(Marca comprada en la 2ª ocasión/Marca comprada en la 1ª ocasión)

Marca comprada en la ocasión 1	Marca comprada en la ocasión 2			
	A2	B2	C2	
A1	0,595	0,245		
B1	0,311		0,146	
C1	0,301		0,477	

Se pide:
a) Complete las tablas dejando constancia de los cálculos efectuados.
b) ¿Son los sucesos "marca comprada en la primera ocasión" y "marca comprada en la segunda ocasión" independientes? ¿P or qué?
c) Teniendo en cuenta que el estudio se realiza sobre 682 consumidores, ¿cuántos eligieron la marca A en la primera compra?

4) Una empresa vende sus productos en tres ciudades. Los porcentajes de venta son: 50% en A, 30% en B y 20% en C. La probabilidad de que se produzca un impagado es, respectivamente, 0,01 en A, 0,02 en B y 0,08 en C. Habiéndose dado un impagado, ¿de qué ciudad es más probable que proceda?

5) La población de matriculados en primer curso en las escuelas universitarias de la Universidad de Málaga en 1975 se distribuyó de la forma siguiente: 0,2 en Empresariales, 0,3 en Peritos y 0,5 en Magisterio. Los que terminaron sus estudios en cada centro fueron:
P(T/E) = 0,8 (terminados que eran de Empresariales)
P(T/P) = 0,6 (terminados que eran de Peritos)
P(T/M) = 0,8 (terminados que eran de Magisterio)
Transcurrido el tiempo suficiente para que todos hayan podido terminar, se les reúne y se elige uno al azar. Se le pregunta si terminó los estudios y contesta que sí. Determine la probabilidad de que la persona elegida se haya diplomado en Ciencias Empresariales.

6) Una entidad bancaria califica a sus clientes, a la hora de conceder préstamos, en dos grupos: clientes "preferentes" y clientes "no preferentes". En su Memoria de 2001 aparecen los siguientes datos:
- El 30% de los préstamos fueron fallidos (no se pagaron a tiempo).
- El 40% de los préstamos fallidos fueron concedidos a clientes "preferentes".
- El 55% de los préstamos no fallidos fueron concedidos a clientes "preferentes".

Calcule:
a) Probabilidad de que un préstamo concedido a un cliente "preferente" resulte fallido.
b) Probabilidad de que un préstamo concedido a un cliente "no preferente" no sea fallido.

7) Se sabe que si el Producto Nacional Bruto (PNB) aumenta, la probabilidad de que el valor de unas acciones aumente es de 0,8. Si el PNB se mantiene constante, la probabilidad de que suban las acciones es 0,2, y si el PNB disminuye, la probabilidad de que aumente el valor de las acciones es de 0,1. Si para el futuro se asignan las probabilidades 0,4; 0,3 y 0,3 a los sucesos: suba el PNB, se mantenga constante y disminuya, respectivamente, responda a las siguientes cuestiones:

a) Determine la probabilidad de que aumente el valor de las acciones.
b) Supuesto que las acciones hayan subido, determine la probabilidad de que el PNB haya subido efectivamente.
c) Supuesto que haya subido el PNB, determine la probabilidad de que las acciones bajen de valor.

8) Un grupo independiente de investigación ha realizado un estudio de las probabilidades de que un accidente en una planta nuclear traiga consigo escapes radioactivos. El grupo considera que los únicos tipos posibles de accidentes en el reactor son incendios, desgaste del material y error humano; que dos o más accidentes nunca ocurrirán juntos y además la probabilidad de ocurra escape radioactivo y no haya pasado ninguno de los accidentes anteriores es nula.
Se han realizado estudios que indican que si hubiera incendio, el escape de radiación ocurrirá el 10% de las veces; si hubiera fallo mecánico, el escape de radiación ocurrirá un 40% de las veces y si hubiera un error humano el escape de radiación ocurrirá un 50% de las veces. Los estudios también demuestran que:
a) La probabilidad de que el incendio y el escape radioactivo ocurran juntos es de 0,0005
b) La probabilidad de que un fallo mecánico y un escape radioactivo ocurran juntos es de 0,001
c) La probabilidad de que un error humano y un escape radioactivo ocurran juntos es de 0,0007

Haciendo uso de esa información se pide:
a) ¿Cuál es la probabilidad de escape radioactivo?
b) ¿Cuáles son las probabilidades, respectivas, de tener un incendio, un fallo mecánico y un error humano bajo las cuales se basaron las probabilidades anteriores?
c) ¿Cuáles son las probabilidades respectivas de que dado un escape radioactivo éste ha ya sido causado por un incendio, un fallo mecánico o un error humano?

9) En las cuatro provincias de una comunidad autónoma se dan las siguientes cifras de población activa y paro:

Provincia	Pob. Activa	Tasa de paro
A	200.000	5%
B	600.000	8%
C	800.000	3%
D	400.000	10%

Se pide:
a) La probabilidad de que elegida una persona activa al azar, ésta sea de la provincia A.
b) La probabilidad de que elegida al azar una persona activa de esta comunidad esté en paro.
c) Si se ha elegido una persona activa al azar y resulta no estar parada, ¿cuál es la probabilidad de que proceda de la provincia A?
d) Si se ha elegido una persona activa al azar y resulta que está en paro, ¿de qué provincia es más probable que proceda?.

10) Sobre la población activa de una provincia tenemos los siguientes datos: el 30% son obreros no cualificados, el 60% son obreros especialistas y el resto son técnicos medios o superiores. Actualmente, el paro afecta al 40% de los no cualificados y al 20% de los especialistas, constituyendo los obreros no cualificados el 48% del total de los parados. Determine el porcentaje de paro que existe entre los técnicos.

11) Analizadas las estadísticas de los visitantes a los museos de una ciudad durante un año determinado se ha observado que 1.000.000 de personas han visitado el total de los museos. En particular se sabe que 700.000 personas han visitado el museo A y 500.000 han visitado el museo B y no se tiene información del resto. Obtenga:

a) La probabilidad de que un visitante visite el museo A.
b) La probabilidad de que un visitante visite el museo B.
c) La probabilidad de que visite los dos museos A y B.
d) La probabilidad de que visite al menos uno de los dos museos.

12) El dueño de una tienda de ropa para hombres ha observado el comportamiento de sus clientes durante un largo periodo de tiempo. Como consecuencia de esa observación afirma que la probabilidad de que un cliente que entre a la tienda compre una camisa es 0,4, pero de los que compran una camisa el 50% compran también una corbata, y solamente un 10% compran la corbata cuando no han comprado la camisa. Obtenga las probabilidades de que los clientes compren lo siguiente:

a) Una camisa y una corbata.
b) Una corbata.
c) Una camisa o una corbata.
d) Una corbata pero no una camisa.

13) Para analizar el volumen de fraude en las declaraciones de IVA, un grupo de investigación realiza el siguiente experimento. Sobre una amplia muestra de empresas se les preguntan dos cuestiones: A) ¿Termina su CIF en número par?, B) ¿Ha cometido algún tipo de fraude en la declaración de IVA?. Para evitar la falta de respuesta motivada por la naturaleza de la segunda pregunta, a los que tienen que responder se les hace la siguiente propuesta: lance una moneda al aire, de forma que si sale cara responda la pregunta A y si sale cruz responda la pregunta B. En ningún caso la empresa encuestada indica qué pregunta ha contestado. Después de realizado el experimento, resulta que el 37% de los entrevistados dan como respuesta sí. En estas condiciones: ¿cuál es la probabilidad de una persona a la que se le preguntó B diga sí?

14) Las probabilidades a priori de los eventos A_1 y A_2 son $P(A_1)=0,4$ y $P(A_2)=0,6$. También se sabe que $P(A_1 A_2)=0$. Suponga que $P(B/A_1)=0,2$ y que $P(B/A_2)=0,5$. Se pide:

a) Calcule $P(A_1 \cup A_2)$.
b) Calcule $P(A_1 B)$ y $P(A_2 B)$.
c) Calcule $P(B)$.
d) Calcule $P(A_1/B)$ y $P(A_2/B)$.

15) El número de camiones que pasan por una carretera donde hay un surtidor de gasolina está en relación 3 a 2 respecto de otra clase de vehículos. La probabilidad de que pasando un camión éste llegue al surtidor a abastecerse es de 0,1. Respecto a otra clase de vehículos dicha probabilidad es 0,2. Si llega un vehículo a abastecerse, ¿qué probabilidad hay de que sea un camión?

SOLUCIONES:

1.- a) 0,81; b) 0,01; c)0,99.
2.- a) 0,9; b) 0,1; c)0,56.
3.- a) Ver tablas. b) No son independientes ya que $P(A_1 \quad A_2) \quad P(A_1)*P(A_2)$. c) 294.

Probabilidades conjuntas

Marca comprada en la ocasión 1	Marca comprada en la ocasión 2			
	A2	B2	C2	
A1	P(A1 A2) **0,256**	P(A1 B2) **0,106**	P(A1 C2) **0,069**	P(A1) **0,431**
B1	P(B1 A2) **0,097**	P(B1 B2) **0,169**	P(B1 C2) **0,045**	P(B1) **0,311**
C1	P(C1 A2) **0,078**	P(C1 B2) **0,057**	P(C1 C2) **0,123**	P(C1) **0,258**
	P(A2) **0,431**	P(B2) **0,332**	P(C2) **0,237**	**1**

Probabilidades condicionadas
P(Marca comprada en la 2ª ocasión/Marca comprada en la 1ª ocasión)

Marca comprada en la ocasión 1	Marca comprada en la ocasión 2			
	A2	B2	C2	
A1	P(A2/A1) **0,595**	P(B2/A1) **0,245**	P(C2/A1) **0,16**	
B1	P(A2/B1) **0,311**	P(B2/B1) **0,543**	P(C2/B1) **0,146**	
C1	P(A2/C1) **0,301**	P(B2/C1) **0,221**	P(C2/C1) **0,477**	

4.- De la ciudad C.

5.- 0,2162.

6.- a) 0,2376; b) 0,6364.

7.- a) 0,41; b) 0,78; c) 0,2.

8.- a)0,0022; b) P(I)=0,005; P(ME)=0,0025; P(EH)=0,0014;

c) P(I/R)=0,2273; P(ME/R)=0,4545; P(EH/R)=0,3182.

9.- a) 0,1; b) 0,061; c) 0,101; d) De la B.

10.- El 10%.

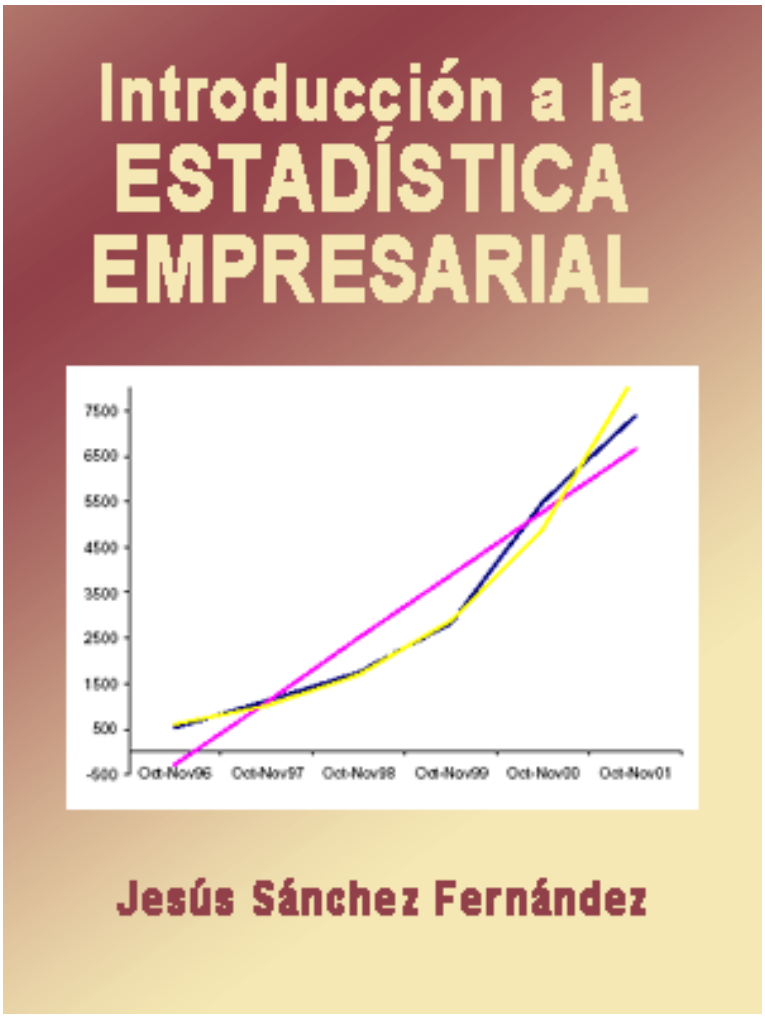
11.- a) 0,7; b) 0,5; c) 0,2 P(A B) 0,5; d) 0,7 P(AUB) 1.

12.- a) 0,2; b) 0,26; c) 0,46; d) 0,06.

13.- 0,24.

14.- a) 1; b) 0,08 y 0,3; c) 0,38; d) 0,21 y 0,79.

15.- 0,4286.



Puede enviar sus comentarios al libro directamente al autor:
j_sanchez10@terra.es

Para citar este libro puede utilizar el siguiente formato:

Sánchez Fernández, J. (2004) Introducción a la Estadística Empresarial
Edición electrónica en http://www.eumed.net/cursecon/libreria/index.htm

editado por eumed.net