ESTADÍSTICA PARA ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

Ángel Muñoz Alamillos
Juan Antonio Vicente Virseda
Azahara Muñoz Martínez

CONTIENE DVD

UNED

EDICIONES ACADÉMICAS

ÁNGEL MUÑOZ ALAMILLOS
JUAN ANTONIO VICENTE VÍRSEDA
AZAHARA MUÑOZ MARTÍNEZ

ESTADÍSTICA PARA ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

UNED
EDICIONES ACADÉMICAS

ISBN: 978-84-92477-37-1
Depósito legal: M. 36.790-2010

Compuesto e impreso por Fernández Ciudad, S. L.
Coto de Doñana, 10.18320 Pinto (Madrid)

Impreso en España - *Printed in Spain*

ÍNDICE

PRÓLOGO

Este libro está especialmente recomendado para el estudio de la asignatura de Introducción a la Estadística en el Grado de Administración y Dirección de Empresas. Se divide en 8 capítulos y está planteado para que su estudio pueda ser afrontado por alumnos que sólo dispongan de unos conocimientos matemáticos elementales.

Sus autores forman parte del equipo docente de la asignatura en la Universidad Nacional de Educación a Distancia, complementando su formación económica con la puramente estadística.

En el primer capítulo se introduce al alumno en la Ciencia Estadística y en sus aplicaciones a la dirección y gestión de empresas, explicando los conceptos estadísticos más elementales. El capítulo se complementa con una revisión de las fuentes estadísticas publicadas en España y que tienen interés y relación directa con la actividad empresarial.

En el segundo capítulo se estudian los conceptos estadísticos fundamentales y la representación gráfica de las variables estadísticas; como en el resto del libro termina con el planteamiento y resolución de una serie de problemas relacionados con la materia tratada.

Los capítulos tercero y cuarto, se dedican, respectivamente, al estudio de las medidas de posición (media aritmética, geométrica y armónica, moda, mediana, cuartiles y percentiles) y de las medidas de dispersión (rango o amplitud, recorrido, rango entre percentiles, recorrido relativo, recorrido semi-intercuartílico, desviaciones media y mediana, varianza, desviación típica o estándar, etc...) en las distribuciones unidimensionales.

En el capítulo quinto se aborda el estudio de las distribuciones bidimensionales, explicando su representación gráfica y sus principales estadísticos descriptivos, y los conceptos de regresión y de correlación, así como la utilización de los mismos para el estudio de la dependencia entre variables y para predecir el comportamiento de variables dependientes.

En el capítulo sexto se estudian los números índices, explicando las principales propiedades y las formulaciones de los índices más habituales (Laspeyres, Paasche y Fisher) y otros menos utilizados (como los de Drovisch-Bowley, Edgeworth-Marshall y Walch).

En el capítulo séptimo se introduce el concepto de series temporales estudiando su representación gráfica, el ajuste de tendencias, las variaciones estacionales y cíclicas y la desestacionalización de las series; con ello termina la parte dedicada a la Estadística Descriptiva.

Finalmente, en el capítulo octavo se abordan los conceptos básicos de la teoría de la probabilidad; estos conceptos son la base para abordar el tercer gran apartado de la Estadística, la denominada Inferencia Estadística, es decir, para inferir datos a una población a partir de los resultados extraídos de una muestra de la misma.

El libro incluye un DVD, en el que se incorporan diversos enlaces con páginas Web de interés, un programa informático de manejo sencillo y la grabación visual de diversas lecciones útiles para aprender a utilizar la hoja de cálculo Excel y el paquete informático SPSS para tratamiento de información estadística; estas lecciones complementan las instrucciones que se dan en algunos epígrafes del libro para la adecuada utilización de estos programas y para el análisis y la interpretación de los resultados estadísticos que proporcionan los mismos.

Los Autores

Capítulo 1

INTRODUCCIÓN

1.1. DEFINICIÓN Y RAMAS DE LA ESTADÍSTICA

La Estadística suele definirse como la ciencia que tiene por objeto recoger de forma agrupada la información que se produce de fenómenos repetitivos o no ocasionales. En su origen la Estadística se desarrolló en la economía y en la política, de hecho hasta bien entrado el siglo XIX la «Estadística» se utilizaba para hacer mención a informaciones de tipo socioeconómico sobre la realidad de un Estado (establecimiento de registros de población, nacimientos, defunciones, etc.; censos de edificios y de elementos de riqueza, etc.); etimológicamente la palabra alude precisamente a esta acepción de «Ciencia de los Estados».

Hoy, la Estadística no queda reservada al estudio del Estado sino que es algo más amplio y útil en múltiples ciencias y áreas del conocimiento humano.

Actualmente, en su definición más común se le reconoce como *«la ciencia que se ocupa de la obtención de información y que proporciona instrumentos para la toma de decisiones cuando prevalecen condiciones de incertidumbre»* o *«La rama del método científico que se ocupa de los datos obtenidos contando o midiendo las propiedades de determinados colectivos»;* a estos colectivos, se le denomina en Estadística «Poblaciones».

El concepto de incertidumbre o de falta de certeza proviene de la existencia en la naturaleza de fenómenos aleatorios; en este sentido se diferencia entre fenómenos deterministas y fenómenos aleatorios; un fenómeno es determinista cuando al experimentarlo en las mismas condiciones se obtienen siempre los mismos resultados (las leyes físicas y químicas clásicas suelen tener este tipo de comportamientos, al dejar caer un objeto desde un árbol siempre cae al suelo a una determinada velocidad derivada de la ley de la gravedad, la unión de 2 moléculas de hidrógeno y una de oxígeno en determinadas condiciones siempre proporciona agua, etc..); un fenómeno es, por el contrario, aleatorio cuando al experimentarlo en las mismas condiciones produce un resultado variable que no puede predecirse a priori con exactitud (el tiempo de vida de una lámpara, el lanzamiento de una moneda, el resultado de un partido de fútbol o de un examen, etc.).

Los fenómenos aleatorios son las más comunes en la naturaleza y en el comportamiento humano y social; bien por su propia naturaleza, bien por su complejidad, bien por la falta de información o de modelos explicativos adecuados, tanto en la vida cotidiana como en la toma de decisiones empresariales nos enfrentamos a factores de «aleatoriedad»; el resultado de una contienda electoral, la cuantía de una subvención

pedida a la administración, la adjudicación de un concurso público, la duración del viaje en coche al lugar de vacaciones, el tiempo que nos hará el fin de semana, el número de visitantes que tendremos en una exposición, etc.., son fenómenos aleatorios, a cuyo análisis y comprensión ayudan las técnicas estadísticas.

La Estadística suele dividirse en dos grandes apartados: La ***Estadística Descriptiva,*** que recoge un conjunto de técnicas y procedimientos para organizar, resumir y tratar sistemáticamente datos disponibles de sucesos ya acaecidos y La **Estadística Inferencial o *Inferencia Estadística,*** que, basada en la teoría matemática de la *Probabilidad,* estudia los métodos empleados para inferir algo acerca de una población basándose en la información aportada por una parte del colectivo (muestra).

1.2. LA ESTADÍSTICA OFICIAL EN ESPAÑA Y EUROPA

La estadística oficial en España comienza su andadura con la creación de la Comisión de Estadística del Reino. El 3 de noviembre de 1856, el general Narváez, presidente del Consejo de Ministros de Isabel II, firma un Decreto por el que se crea una Comisión, compuesta por personas de reconocida capacidad, para la formación de la Estadística General del Reino. Unos meses más tarde, el 21 de abril de 1857, la Comisión pasa a denominarse Junta de Estadística. Su primer trabajo es el Censo de Población, con fecha de referencia del 21 de mayo del mismo año.

La Ley de Instrucción Pública de 9 de septiembre de 1857 establece que la Estadística será una disciplina académica, pasando a impartirse en la Universidad. Un Decreto del 12 de septiembre de 1870, durante el gobierno provisional del general Serrano, crea el Instituto Geográfico. Tres años más tarde, 19 de junio de 1873, pasa a denominarse Instituto Geográfico y Estadístico, asumiendo todas las tareas de recogida de información numérica para el Estado.

En 1877, el Instituto Geográfico y Estadístico aprueba su Reglamento. Las estadísticas pasan a depender del Ministerio de Fomento en el año 1890. Un Decreto de 1 de octubre de 1901 establece la formación de las estadísticas oficiales y la publicación de las mismas. El Instituto Geográfico y Estadístico se transforma en Dirección General y se crean departamentos en los Ministerios para completar su labor.

En 1924, el Consejo del Servicio Estadístico, creado en 1921 es reformado, cuatro años antes de que pase a depender del Ministerio de Trabajo y Previsión. En 1931, la adscripción se hace al Ministerio de la Presidencia.

Durante la Guerra Civil (1936-1939) comienza a funcionar el Servicio Sindical de Estadística en coordinación con los Servicios de Estadística del Estado, dentro de la llamada zona nacional.

La Ley de 31 de diciembre de 1945, publicada en el BOE del 3 de enero de 1946, crea el Instituto Nacional de Estadística (INE, www.ine.es), que tiene como misión la elaboración y perfeccionamiento de las estadísticas demográficas, económicas y sociales ya existentes, la creación de otras nuevas y la coordinación con los servicios estadísticos de las áreas provinciales y municipales.

Además de regular la coordinación entre otros servicios estadísticos como el Servicio Sindical de Estadística, la Ley crea el Consejo Superior de Estadística. El INE se

organiza en Servicios Centrales, Delegaciones provinciales y Delegaciones en los Ministerios. El 9 de mayo de 1989 se promulga la Ley de la Función Estadística Pública que hace del INE un organismo autónomo potenciando las nuevas tecnologías estadísticas, la coordinación con las Comunidades Autónomas, la elaboración del Plan Estadístico Nacional y las relaciones con la Unión Europea en materia estadística.

El actual Estatuto del INE, aprobado por Real Decreto 508/2001 de 11 de mayo (BOE 12-05-2001), le asigna las funciones de coordinación general de los servicios estadísticos de la Administración General del Estado, la vigilancia, control y supervisión de las competencias de carácter técnico de los servicios estadísticos estatales, y las demás previstas en la Ley 12/1989, de 9 de mayo, de la Función Estadística Pública.

Con la creación del Estado de las autonomías, a raíz de la Constitución de 1978, los gobiernos regionales han ido asumiendo de forma paulatina la competencia exclusiva en materia estadística para fines no estatales y dentro de sus ámbitos geográficos, creando distintos organismos de estadística regionales; destacan el Instituto Vasco de Estadística (EUSTAT), el Instituto de Estadística de Cataluña (IDESCAT) y el Instituto Gallego de Estadística (IGE), Instituto de Estadística de la Comunidad de Madrid, etc.[1]

Por su parte los Ayuntamientos son los encargados de la creación, mantenimiento, revisión y custodia del Padrón Municipal de Habitantes, correspondiendo en todo caso al INE la coordinación de todos los municipios y la realización de las comprobaciones oportunas. La resolución de las posibles discrepancias entre los Ayuntamientos con el INE, corresponde al Consejo de Empadronamiento, Órgano colegiado con representantes de la Administración General del Estado (INE, Oficina del Censo Electoral y Ministerio de AAPP). Asimismo, cabe destacar la importante labor estadística que comienzan a realizar los ayuntamientos correspondientes a los grandes municipios, como Madrid, Barcelona o Sevilla, como consecuencia de la creciente demanda de información estadística de ámbito municipal e inferior.

Desde esta perspectiva, resulta imprescindible una armonización metodológica de las estadísticas que permita su comparabilidad; a ello han contribuido decisivamente organismos internacionales como la Oficina de Estadísticas Europea (EUROSTAT), el Banco Central Europeo, la Organización Mundial del Turismo (OMT) o la Organización para la Cooperación y el Desarrollo Económico (OCDE).

El EUROSTAT, con sede en Luxemburgo, es la oficina estadística de la Comisión Europea, que produce datos sobre la Unión Europea y promueve la armonización de los métodos estadísticos de los estados miembros. Se creó en 1953 para satisfacer las demandas de la Comunidad del Carbón y el Acero. A lo largo de los años su tarea se ha ampliado; cuando se fundó la Comunidad Económica Europea en 1958 se convirtió en una Dirección General (DG) de la Comisión Europea. Dos de sus papeles particularmente importantes son:

— La producción de datos macroeconómicos que apoyan las decisiones del Banco Central Europeo en la política monetaria de la eurozona.

[1] Las direcciones Web de cada uno de ellos están disponibles en la página del INE (http://www.ine.es/serv/estadist.htm#0001).

— La producción y organización de datos regionales y su clasificación por zonas (NUTS) para orientar las políticas estructurales de la Unión Europea.

Puede decirse que el principal papel de EUROSTAT es proporcionar estadísticas al resto de Direcciones Generales de la Comisión Europea y otras instituciones Europeas con el fin de que éstas puedan definir, analizar y mejorar las políticas comunitarias; también se ocupa de desarrollar sistemas estadísticos en los países candidatos a entrar en la Unión.

EUROSTAT no genera datos. Son las autoridades estadísticas de los Estados Miembros las que generan, verifican y analizan los datos nacionales y los envían a EUROSTAT, que asegura que su utiliza una metodología homogénea que permite controlar y comparar la información.

1.3. INTERÉS DE LA ESTADÍSTICA PARA EL ANÁLISIS ECONÓMICO

Los gestores de las empresas deben disponer de información en la que apoyar su toma de decisiones; la estadística es, precisamente, un instrumento que facilita tanto la recogida de esta información como su posterior tratamiento y análisis.

Se trata pues de un instrumento auxiliar en la labor de administración de las empresas, que permite:

— Obtener información de la realidad que directa o indirectamente afecta al trabajo que desarrolla la empresa, proporcionando métodos y técnicas para la recogida de datos y su codificación o sistematización.
— Sistematizar y reducir dicha información para hacer más fácil su comprensión mediante tablas y gráficos.
— Buscar reiteraciones de determinados fenómenos con objeto de predecir su posible repetición en el futuro.
— Relacionar comportamientos entre diversas variables a fin de efectuar estimaciones con diversas hipótesis sobre el comportamiento de las variables.

En este libro aportaremos los conocimientos mínimos imprescindibles para abordar este tipo de tareas.

1.4. CONCEPTOS ESTADÍSTICOS FUNDAMENTALES

Los principales conceptos básicos que se utilizan en estadística son los siguientes:

Población

Se denomina población al conjunto de elementos que cumplen ciertas propiedades y entre los que se desea estudiar el fenómeno en cuestión; este conjunto de elementos son de distinta naturaleza: personas, hogares, empresas, edificios, tornillos, caras de una moneda o de un dado, cartas de una baraja, etc.

Si estudiamos las caras de una moneda la población será de dos elementos (cara y cruz), si es un dado, de 6 elementos, etc.

Muestra

Muestra es cualquier subconjunto de individuos pertenecientes a una población determinada.

El interés de las muestras es que, en determinadas condiciones, las técnicas estadísticas permiten que los resultados que se obtengan del análisis de una muestra puedan ser extendidos («inferidos») al conjunto de la población a la que pertenece; estas técnicas nos evitan el enorme trabajo que puede suponer el estudio de toda una población de grandes dimensiones.

Para ello las muestras deben ser *representativas* de la población, tener un tamaño suficientemente grande y cumplir otras condiciones estadísticas que estudiaremos más adelante.

Individuo

Individuo o Unidad de Investigación es cada uno de los elementos de una muestra o de una población.

Un ejemplo de población serían el conjunto de visitantes al Museo del Prado en el mes de junio del año 2009. Cada uno de estos visitantes es un individuo o un elemento de dicha población; una muestra elegida aleatoriamente de esta población podrían ser la selección de los 10 primeros visitantes que entraron cada día a partir de las 11 y de las 13 horas de la mañana (horas, ambas, elegidas al azar).

Sí tenemos 30 días de apertura del museo y realizamos una encuesta a 20 individuos cada día, tendremos una muestra de 600 individuos elegidos al azar; esta muestra es suficientemente grande para extraer determinadas conclusiones estadísticas generales, de forma que, sí en la muestra seleccionada un 30% de los individuos son extranjeros y el 70% nacionales, podemos inferir ambas proporciones al colectivo y afirmar que el 30% de los visitantes al Museo del Prado durante el mes de referencia eran extranjeros, evitando con ello el arduo trabajo de tener que entrevistar a todos y cada uno de los visitantes para conocer su nacionalidad.

La teoría de la inferencia estadística nos permitirá extraer estas conclusiones con un determinado error y un cierto nivel de confianza o probabilidad de equivocarnos; de esta forma será habitual indicar que el dato anterior se ha obtenido, por ejemplo, con un 2,5% de error y con un nivel de confianza del 95%; estas afirmaciones quieren decir que la teoría estadística asegura que sí hiciésemos el experimento (repitiésemos la encuesta) 100 veces en idénticas condiciones, 95 veces obtendríamos un resultado que estaría comprendido entre el 28,25 (un 5% menos del 30% obtenido) y el 30,75% (un 5% más del 30% obtenido); las otras 5 veces puede darnos cualquier cosa diferente.

Hasta ahí y sólo hasta ahí llega la inferencia estadística, lo que no es poco como ayuda en la toma de decisiones empresariales; ya que saber, con un 95% de certeza, el

país de procedencia de los visitantes a una ciudad, el grado de satisfacción de los clientes con los servicios de una empresa, el motivo de un determinado producto, etc., puede ser fundamental para planificar una acción de marketing o para tomar decisiones de cambio en la organización de una determinada empresa.

Las investigaciones estadísticas pueden ser ***censales y muestrales***; baste recordar aquí, que según los casos, determinadas investigaciones sólo pueden hacerse con carácter muestral (sólo puede analizarse una muestra de sangre de un individuo o una muestra de agua del mar, ya que una investigación censal sería en ambos casos imposible), mientras que en otros, son necesarias investigaciones censales (un catastro, que se haga con el fin de grabar los edificios con un impuesto de bienes inmuebles, debe tener carácter censal porque debe disponerse de información detallada de todos y cada uno de los individuos a fin de fijar la cuota impositiva correspondiente).

Parámetros

Son las características poblacionales que deseamos investigar y que suelen ser desconocidas a priori. Por ejemplo, la edad de los viajeros de una compañía aérea, el precio medio en el mercado de un determinado producto, la opinión de los clientes sobre los productos fabricados por una empresa, etc.

Variables

Cuando estas características o parámetros son *numéricas*, es decir, cuando se pueden medir, se denominan ***Variables*** (años de edad, renta anual en euros, etc.).

Las variables pueden ser ***discretas o continuas***, según tomen un número infinito no numerable (variables continuas) o un número finito o infinito numerable de valores (discretas) en un intervalo.

Ejemplos de variables discretas son el número de hijos de una familia, el número de coches de un país, el número de turistas que visitan una ciudad, etc.-, ejemplos de variables continuas son la temperatura, la edad, el peso o la altura de las personas, la medición de la distancia entre dos puntos, etc.

La mayor parte de las variables continuas pueden tratarse como discretas; así, por ejemplo, si valoramos la edad de las personas en años, despreciando las unidades de tiempo menores (días, minutos, segundos, milésimas de segundo, etc.), una variable continua se convierte en discreta; lo mismo podríamos pensar para el peso en kilogramos o para la distancia entre dos puntos medida en kilómetros, etc.

También, y como veremos más adelante, aunque se pierda información sobre la distribución, es bastante habitual agrupar los valores de la variable en *intervalos* de amplitud constante o variable.

Otra característica de las variables es su referencia o no a un determinado período de tiempo.

Tendremos en este sentido dos tipos de datos:

— Variables temporales o históricas; son las referidas a distintos momentos del tiempo y adoptan en general la forma de serie; por ejemplo, la serie mensual de parados inscritos en el INEM.
— Variables atemporales, también llamadas de corte transversal; están referidas a un momento o período concreto y más o menos largo; por ejemplo, las personas que visitaron Toledo el mes de febrero de 2010 o la cifra de negocio de una empresa durante el año 2009.

Atributos

Cuando los parámetros o características de la población no son susceptibles de medirse numéricamente reciben el nombre de ***Atributos*** (el color del pelo, el sexo, la profesión, el estado civil, el grado de satisfacción del cliente con un servicio, etc.).

Los atributos, a diferencia de las variables, presentan ***Modalidades o Categorías*** (el atributo sexo puede adoptar las modalidades de varón o mujer, el atributo intención de voto se concreta en los nombres de los partidos políticos que se presenten a las elecciones sondeadas o el de opinión sobre la satisfacción de un cliente con el servicio recibido en una gestión bancaria, puede adoptar las modalidades que se determinen: excelente, bueno, malo, regular, etc.).

Los atributos pueden clasificarse como **ordenables,** que son los que sugieren una ordenación, por ejemplo, el grado de satisfacción con el trato recibido (excelente, bueno, regular, malo), el nivel de estudios (alto, medio, bajo), y **no ordenables,** que son los que sólo admiten una ordenación alfabética o casual, por ejemplo el estado civil, el color del pelo, el país de procedencia o la nacionalidad de un turista, etc.

El atributo más simple es el que sólo presenta dos modalidades: presencia/ausencia; favorable/desfavorable, etc.

1.5. LAS FUENTES DE LA INFORMACIÓN ESTADÍSTICA

Las fuentes de información estadística son aquellas que proporcionan los datos sometidos al análisis estadístico; suelen clasificarse en dos tipos: *fuentes de información* ***secundarias*** *y* ***primarias***; las primeras son aquellas ya existentes y elaboradas por otros agentes o investigadores, públicos o privados, ajenos a nuestra investigación; las fuentes directas o primarias son las elaboradas específicamente para la propia investigación.

Dado que el coste de la información primaria es normalmente bastante más alto, en cualquier estudio estadístico es conveniente comenzar realizando una investigación sobre las posibles fuentes de información secundarias previamente existentes; su disponibilidad facilita la planificación de la operación, a veces incluso evita la recogida de cierta información inicialmente prevista y que se descubre de que ya está disponible; en todo caso la información preexistente puede permitir la comparación o relativización de nuestros resultados con los generados por otros investigadores.

En el siguiente epígrafe, se describen las principales fuentes estadísticas oficiales que existen en España relativas al análisis económico; buena parte de esta información es asequible por Internet y el resto puede consultarse en centros de documentación o comprarse a precios relativamente reducidos.

1.6. LAS ESTADÍSTICAS ECONÓMICAS EN ESPAÑA

1.6.1 Introducción

Antes de abordar los principales conceptos estadísticos, hemos considerado adecuado introducir un apartado en el que se resuma el primero de los tres aspectos señalados con anterioridad, es decir, el conocimiento de las principales estadísticas económicas disponibles en España; estas estadísticas permiten la realización de investigaciones y análisis sobre la situación actual y la evolución de la economía española en aspectos que son necesarios para la dirección y administración de empresas.

El INE almacena toda la información estadística que produce en formatos electrónicos en Internet a través de un sistema denominado INEbase. La organización primaria de la información sigue la clasificación temática del Inventario de Operaciones Estadísticas de la Administración General del Estado (IOE). La unidad básica de INEbase 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.

A las operaciones estadísticas se puede acceder directamente a través de la lista completa de operaciones de INEbase o a través de los menús temáticos. Estos menús permiten conocer toda la información disponible de cada tema: operaciones para las que se presentan resultados, junto con una pequeña descripción de las variables publicadas, la periodicidad y disponibilidad de los datos y el ámbito geográfico; publicaciones y estudios relacionados; enlaces a otras webs donde ampliar la información de fuentes externas; y un enlace al IOE para conocer todas las operaciones del Sistema Estadístico Español relacionadas con el tema.

Para cada operación estadística en INEbase existe una página que da acceso a toda la información relativa a la misma: los resultados detallados completos, la última nota de prensa publicada, el calendario de disponibilidad de datos y toda la información metodológica o descriptiva que ayuda a la mejor comprensión e interpretación de los datos (metodologías, cuestionarios, clasificaciones, notas explicativas, ...).

Los resultados detallados incluyen los últimos resultados publicados y además la historia reciente de la estadística. Los ficheros de datos se pueden visualizar directamente desde INEbase o descargar en formato Pc-Axis para un tratamiento posterior utilizando el programa Pc-Axis cuya descarga se puede hacer de forma gratuita.

El alumno debe entrar en la Web del INE (www.ine.es) y familiarizarse con la información estadística disponible.

Las operaciones que figuran en INEBase se agrupan en las siguientes áreas temáticas:

— Entorno físico y medio ambiente.
— Demografía y población.
— Sociedad.
— Economía.
— Agricultura.
— Industria, energía y construcción.
— Servicios.
— Clasificaciones.
— Internacional.
— Síntesis estadística.
— Banco de series Tempus.

Dentro del área temática correspondiente a la economía, las operaciones estadísticas se estructuran en los siguientes apartados:

1. Empresas

   1.1. Directorio central de empresas: explotación estadística.

2. Cuentas económicas

   2.1. Contabilidad nacional de España.
   2.2. Contabilidad nacional trimestral de España.
   2.3. Cuentas trimestrales no financieras de los sectores institucionales.
   2.4. Contabilidad regional de España.
   2.5. Cuenta satélite del turismo de España.
   2.6. Balanza de pagos.

3. Estadísticas financieras y monetarias.

   3.1. Efectos de comercio devueltos impagados.
   3.2. Hipotecas.
   3.3. Sociedades mercantiles.
   3.4. Estadística del procedimiento concursal.
   3.5. Estadística de transmisiones de derechos de la propiedad.
   3.6. Suspensiones de pagos y declaraciones de quiebras.
   3.7. Ventas a plazos.
   3.8. Magnitudes monetarias y financieras.
   3.9. Mercado bursátil.
   3.10. Tipos de interés.

4. Comercio exterior.

   4.1. Índices de comercio exterior de servicios.
   4.2. Resultados de comercio exterior.
   4.3. Tipos de cambio.

5. Información tributaria

   5.1. Mercado de trabajo y pensiones en las fuentes tributarias.
   5.2. Estadísticas de impuestos.

En lo referente a las operaciones que analizan datos de empresa, la información se estructura, tal como puede observarse en la lista de operaciones, según el sector de actividad al que pertenecen. Las operaciones que engloban son las siguientes:

1. Ciencia y tecnología.

   1.1. Investigación y desarrollo tecnológico.
   1.2. Estadística sobre actividades de I+D.
   1.3. Encuesta sobre innovación tecnológica en las empresas.
   1.4. Indicadores de alta tecnología.
   1.5. Encuesta de recursos humanos en Ciencia y Tecnología.
   1.6. Estadísticas de propiedad industrial.

2. Nuevas tecnologías de la información y la comunicación.

   2.1. Encuesta de uso de TIC y comercio electrónico en las empresas.
   2.2. Encuesta de tecnologías de la información en los hogares.
   2.3. Indicadores del sector TIC.
   2.4. Tecnología de la información en la enseñanza no universitaria.

3. Agricultura, ganadería, silvicultura y pesca.

   3.1. Censo Agrario.
   3.2. Encuesta sobre la estructura de las explotaciones agrícolas.
   3.3. Producción agrícola.
   3.4. Producción ganadera.
   3.5. Indicadores económicos agrarios.
   3.6. Medios de producción.
   3.7. Silvicultura. Estadísticas pesqueras.
   3.8. Agricultura y ganadería ecológica.

4. Industria.

   4.1. Índices de producción industrial.
   4.2. Índices de precios industriales.
   4.3. Índices de entradas de pedidos en la industria.
   4.4. Índices de cifras de negocios en la industria.
   4.5. Índices de precios de exportación e importación de productos industriales.
   4.6. Encuesta industrial de empresas.
   4.7. Encuesta industrial de productos.
   4.8. Fabricación de vehículos.

5. Energía
   - 5.1. Encuesta de consumos energéticos.
   - 5.2. Otros resultados sobre la energía.

6. Construcción y vivienda.
   - 6.1. Índice de precios de vivienda.
   - 6.2. Censo de Población y Viviendas 2001.
   - 6.3. Censo de Población y Viviendas 1991.
   - 6.4. Índices de precios de materiales e índices nacionales de la mano de obra.
   - 6.5. Estadísticas de la construcción.

7. Encuestas globales del sector Servicios.
   - 7.1. Indicadores de actividad del sector servicios.
   - 7.2. Índices de precios del sector servicios.
   - 7.3. Índices de comercio exterior de servicios.
   - 7.4. Encuesta anual de servicios.
   - 7.5. Estadística de productos en el sector servicios.
   - 7.6. Estadísticas de filiales de empresas extranjeras en el sector servicios.

8. Comercio.
   - 8.1. Índices de comercio al por menor.
   - 8.2. Encuesta anual de comercio.
   - 8.3. Estadística de productos en el sector comercio.
   - 8.4. Encuesta de uso de TIC y comercio electrónico en las empresas.
   - 8.5. Encuesta de comercio al por menor.
   - 8.6. Encuesta de comercio al por mayor.

9. Transporte y actividades conexas, comunicaciones.
   - 9.1. Estadística de transporte de viajeros.
   - 9.2. Transporte de mercancías por carretera.
   - 9.3. Red de carreteras, vehículos, conductores y accidentes.
   - 9.4. Transporte ferroviario.
   - 9.5. Transporte marítimo.
   - 9.6. Transporte aéreo.
   - 9.7. Transporte por tubería.
   - 9.8. Estadísticas de turismos de servicio público.
   - 9.9. Servicios postales y de telecomunicaciones.

10. Hostelería y turismo.
    - 10.1. Encuesta de ocupación hotelera.
    - 10.2. Encuesta de ocupación en acampamentos turísticos.
    - 10.3. Encuesta de ocupación en apartamentos turísticos.

10.4. Encuesta de ocupación en alojamientos de turismo rural.
10.5. Índice de precios hoteleros.
10.6. Índice de ingresos hoteleros.
10.7. Índice de precios de acampamentos, apartamentos turísticos y alojamientos de turismo rural.
10.8. Encuesta sobre la estructura de las empresas hoteleras.
10.9. Encuesta sobre la estructura de empresas de agencias de viajes.
10.10. Turismo receptor.
10.11. Turismo nacional.
10.12. Ingresos y pagos por turismo.
10.13. Albergues y ciudades de vacaciones.

11. Otros servicios empresariales, personales y comunitarios.

11.1. Encuesta de servicios audiovisuales.
11.2. Encuesta de servicios técnicos.
11.3. Encuesta de servicios personales.
11.4. Encuesta de servicios informáticos.
11.5. Encuesta de servicios industriales de limpieza.

En el ámbito social, desde el punto de vista del consumo y de los precios, cabe destacar:

— Encuesta de presupuestos familiares.
— Índice de Precios al Consumo

Y, por último, respecto al mercado laboral:

— Encuesta de Población Activa.
— Encuesta Anual de Coste Laboral.

No todas estas operaciones son realizadas por el INE, aparecen aquí todas las más importantes que se realizan en este ámbito, así la Información tributaria y los resultados de Comercio Exterior depende del Ministerio de Economía y Hacienda, las relativas al sector agrario del Ministerio de Medio Ambiente y Medio Rural y Marino, las Estadísticas de la construcción de la Dirección General de Programación Económica del Ministerio de Fomento y la Balanza de Pagos, las Magnitudes Monetarias y Financieras y los Tipos de interés y Tipos de cambio del Banco de España. En estos casos, sólo se publican los principales resultados, indicando el enlace correspondiente al organismo elaborador de la estadística para más detalle.

Al margen del INE, buena parte de los organismos estadísticos regionales han comenzado en la actualidad a elaborar sus propias estadísticas económicas, realizando encuestas a empresas y hogares, elaborando sus propios directorios de empresas y locales y estimando las cuentas económicas relativas a su ámbito territorial.

Los servicios estadísticos de los principales Ayuntamientos disponen también de importante información a la que puede accederse desde sus respectivas páginas Web.

Cabe también señalar por su importancia a nivel económico y social, la información estadística derivada de los registros administrativos del Ministerio de Trabajo e Inmigración, en concreto, de los ficheros de centros de cotización y afiliados a la seguridad social y los registros de paro y contratos del INEM, así como la Encuesta de Coyuntura Laboral, que analiza el mercado laboral desde la perspectiva de la demanda (se encuesta a los centros de cotización).

También merecen especial mención el Instituto de Estudios Turísticos (IET), dependiente del Ministerio de Industria, Turismo y Comercio, que como productor de información, es el encargado de las operaciones estadísticas Movimientos Turísticos en Fronteras (FRONTUR), Encuesta de Gasto Turístico (Egatur) y Movimientos Turísticos de los Españoles (FAMILITUR), generando datos sobre las llegadas de visitantes extranjeros a nuestro país, sus peculiaridades, gastos que realizan y viajes realizados por los españoles y sus características.

Otra fuente de información de gran interés es la Estadística Mercantil del Registro Mercantil Central, que publica datos estadísticos referidos a los actos societarios de constitución, extinción, transformación, fusión, aumento y reducciones de capital, situaciones concursales, depósitos de estados financieros y sistemas de administración societaria, así como datos muy relevantes sobre los diversos objetos sociales que constituyen el tejido productivo de nuestro país.

Por último, puede destacarse, otras instituciones como el Banco de España (www.bde.es) o la Intervención General de la Administración del Estado (IGAE; http://www.igae.pap.meh.es) que elabora, a partir de las liquidaciones presupuestarias, las cuentas donde se registra la actividad económica desarrollada por las administraciones públicas.

En las páginas que siguen haremos un repaso de las principales operaciones estadísticas españolas en el ámbito socioeconómico y en el relativo a la información económica y financiera de las empresas, explicando los rasgos metodológicos básicos con los que se elaboran y la información que proporcionan.

1.6.2. Estadísticas económicas

Siguiendo la clasificación de las operaciones señaladas en INEBase, exponemos a continuación las principales características.

1.6.2.1. Estadísticas sobre empresas

A) El Directorio Central de Empresas (DIRCE)

El DIRCE es un directorio económico que el INE ha elaborado siguiendo las recomendaciones del Reglamento nº 2186/93. El DIRCE, cuyos antecedentes son el Proyecto de Integración de Directorios Económicos de finales de 1989, se creo con el ejercicio de referencia de 1994, manteniéndose desde entonces hasta el presente.

El DIRCE se creó a partir de las siguientes fuentes básicas:

a) Impuesto de Actividades Económicas.
b) Licencias Fiscales de la Comunidad Foral de Navarra.
c) Directorio de Locales del País Vasco.
d) Registro de Retenedores a Cuenta de la Declaración Anual de Retenciones sobre las Rentas del Trabajo (DART).
e) Cuentas de Cotización de la Seguridad Social.

La existencia de fuentes procedentes de los registros tributarios hace que el DIRCE esté sujeto a la Ley General Tributaria, sin que pueda transmitirse entre Administraciones, aunque sean depositarias del deber de Secreto Estadístico establecido por las diferentes Leyes Estadísticas, ni pueda ser difundido en los términos que establece la Ley de Protección de Datos.

El DIRCE cubre todas las actividades económicas excepto la producción agraria y pesquera, los servicios administrativos de la Administración Central, Autónoma y Local (incluida la Seguridad Social), las actividades de las comunidades de propietarios y el servicio doméstico. El resto de las actividades de las Administraciones Públicas (sanidad, enseñanza, producción destina a venta), y de las Instituciones Privadas sin fines de lucro solo se cubren de forma parcial.

La unidad de referencia del DIRCE es la empresa, que se define como: «Organización sometida a una autoridad rectora que puede ser, según los casos, una persona física, una persona jurídica o una combinación de ambas y constituida con miras a ejercer en uno o varios lugares una o varias actividades de producción de bienes o servicios», ofreciéndose también la información relativa a sus unidades locales.

Su objetivo básico es hacer posible la realización de encuestas económicas por muestreo. Se actualiza una vez al año, generándose un nuevo sistema de información a 1 de enero de cada período.

Se publica una explotación estadística de los resultados para empresas y unidades locales, desglosados por comunidades autónomas según condición jurídica, actividad económica principal y estrato de asalariados asignado. El DIRCE genera información asociada a: altas, permanencias y bajas, clasificadas estas según sector económico, condición jurídica y estrato de asalariados, es decir, se describe la demografía del sector empresarial español. Las fuentes estadísticas utilizadas en su elaboración son los ficheros de centros de cotización a la Seguridad Social, el fichero del Impuesto de Actividades Económicas y las propias encuestas a empresas realizadas por el INE.

1.6.2.2. Cuentas económicas

A) Contabilidad Nacional de España (CNE)

El objetivo más relevante de los Sistemas de Cuentas Económicas es ofrecer una representación cuantificada de la realidad económica, referida a ámbitos espaciales y temporales determinados, que sea lo más actual, sistemática, completa y fiable posible.

Periódicamente se introducen en la CNE los cambios de base contable, destinados a actualizar las mediciones de las ponderaciones, conforme se producen cambios en las metodologías de referencia y en las fuentes y procedimientos de estimación de las cuentas nacionales.

El Sistema de Cuentas Nacionales de la economía española está adaptado al Sistema Europeo de Cuentas Nacionales y Regionales (SEC95), que aplican de forma armonizada todos los Estados miembros de la Unión Europea (UE), en cumplimiento de lo dispuesto en el Reglamento del Consejo de la CEE nº 2223/96 de 25 de junio de 1996.

Además del marco input-output, se elaboran los siguientes cuadros contables:

— Cuentas económicas integradas.
— Cuenta de bienes y servicios.
— Cuentas del total de la economía y de los sectores institucionales.
— Cuentas del sector *Instituciones financieras y sus subsectores.*
— Cuentas del sector Administraciones públicas y sus subsectores.
— Cuentas del sector Resto del mundo y sus subsectores.
— Cuenta de producción y explotación por ramas de actividad.
— Agregados por ramas de actividad.
— Clasificación del gasto en consumo final de los hogares por finalidad (COICOP).
— Gasto en consumo final de las Administraciones públicas.
— Formación bruta de capital.
— Operaciones de bienes y servicios con el Resto del Mundo.
— Tablas anexas.

B) Contabilidad Nacional Trimestral de España (CNTR)

La Contabilidad Nacional Trimestral de España (CNTR) es una estadística de síntesis de carácter coyuntural, cuyo objetivo primordial es proporcionar una descripción cuantitativa coherente del conjunto de la actividad económica española en el pasado inmediato, mediante un cuadro macroeconómico trimestral, elaborado desde la óptica de la oferta, la demanda y las rentas. Asimismo, incluye estimaciones del nivel de empleo, en términos de Contabilidad Nacional, abarcando los conceptos: personas, puestos de trabajo, puestos de trabajo equivalentes a tiempo completo. Las estimaciones de la CNTR se ofrecen en dos versiones: datos brutos o no ajustados y datos corregidos de estacionalidad y calendario. El sistema de la CNTR incluye la Cuenta Trimestral del Resto del Mundo que cuantifica, entre otras cosas, la capacidad o necesidad de financiación de la economía en su conjunto.

Todas las estimaciones de la CNTR se ajustan a los mismos principios de coherencia y equilibrio contable que la Contabilidad Nacional de España (CNE) anual. Ambas estadísticas se elaboran en consonancia con el Sistema Europeo de Cuentas Nacionales y Regionales (SEC95).

C) Cuentas trimestrales no financieras de los sectores institucionales

El objetivo de las Cuentas trimestrales no financieras de los sectores institucionales (CTNFSI) es describir las relaciones de comportamiento entre las unidades institucionales que forman la economía nacional (hogares e instituciones sin fines de lucro al servicio de los hogares, sociedades no financieras, instituciones financieras y administraciones públicas) y entre aquellas y las unidades que forman el resto del mundo.

Las CTNFSI forman parte del objetivo global de elaboración de un sistema de cuentas anuales y trimestrales para la Unión Europea y la zona del euro. Este sistema incluye los principales agregados macroeconómicos y las cuentas financieras y no financieras de los sectores institucionales. El objetivo es aportar coherencia entre todos esos conjuntos de cuentas y, en relación con las cuentas del resto del mundo, entre los datos de la balanza de pagos y los de las cuentas nacionales.

Las CTNFSI se elaboran de acuerdo con el Reglamento del Parlamento Europeo y del Consejo número 1161/2005. No obstante, el marco conceptual y normativo que subyace es conforme con los principios del Sistema Europeo de Cuentas Nacionales y Regionales (SEC-95), actualizados por las enmiendas que, con posterioridad a su adopción, se han incorporado bajo la forma de Reglamentos del Parlamento Europeo y del Consejo.

D) Contabilidad Regional de España (CRE)

La Contabilidad Regional de España (CRE) es una operación estadística que el INE viene realizando desde el año 1980 y cuyo principal objetivo es ofrecer una descripción cuantificada, sistemática y lo más completa posible de la actividad económica regional en España (comunidades autónomas y provincias), durante el período de referencia considerado.

La información que proporciona permite analizar y evaluar la estructura y evolución de las economías regionales, y sirve de base estadística para el diseño, ejecución y seguimiento de las políticas regionales.

Las cuentas regionales son una especificación de las cuentas nacionales; es decir, la Contabilidad Nacional de España (CNE) constituye el marco de referencia conceptual y cuantitativa en el que se integra la CRE.

E) La Cuenta Satélite del Turismo de España (CSTE)

Esta operación intenta armonizar todas las informaciones económicas referidas al sector para evaluar su impacto económico real, superando las dificultades que entraña la propia delimitación sectorial y las derivadas de la aplicación de diferentes metodologías de estimación; la responsabilidad de su elaboración recae en la Subdirección General de Cuentas Nacionales del INE.

Para la medición del impacto económico del sector hay que abordar los problemas propios de la evaluación de una actividad transversal en la que ni desde la

perspectiva de los productos, ni desde la perspectiva de las unidades de producción o actividades productivas, puede establecerse de manera unívoca una delimitación del turismo, ya que, como reconoce la propia metodología de esta operación estadística:

— Los gastos realizados por los turistas no son fácilmente identificables y mensurables, ya que, si bien se concentran en servicios como los de alojamiento, transporte, agencias de viaje, pueden abarcar un número mucho más amplio de productos y materializarse en casi cualquier tipo de bien o servicio disponible.
— Las industrias que concentran la mayor parte de la oferta de productos para los turistas sirven también otras motivaciones y usos distintos del turismo: un restaurante sirve también comidas a las personas que residen o trabajan cerca del establecimiento, sin que ello tenga nada que ver con la *actividad turística*. Lo mismo podría decirse del transporte de pasajeros, etc.

La CSTE está compuesta por un conjunto de cuentas y tablas, basadas en los principios metodológicos de la contabilidad nacional, que presenta los distintos parámetros económicos del turismo en España, para una fecha de referencia dada. Comprende tres tipos de elementos:

— Cuentas y tablas de oferta, en las que se trata de caracterizar la estructura de producción y costes de las empresas turísticas.
— Tablas de demanda, en las que se trata de caracterizar, desde el punto de vista económico, los diferentes tipos de turistas, el turismo nacional frente al internacional, el tipo de bienes y servicios demandados, etc.
— Tablas que interrelacionan la oferta con la demanda, que permiten obtener unas mediciones integradas de la aportación del turismo a la economía a través de variables macro como el PIB, la producción o el empleo.

F) La Balanza de Pagos

La Balanza de Pagos española es elaborada por el Banco de España con el fin de registrar de manera sistemática todas las transacciones económicas entre España y el resto del mundo. Esta información se elabora a partir de registros administrativos y tiene periodicidad mensual, trimestral y anual.

Para su elaboración, el Banco de España, de acuerdo con las directrices y recomendaciones establecidas por el Quinto Manual de Balanza de Pagos del Fondo Monetario Internacional, emplea un sistema basado en el registro de las transacciones internacionales comunicadas por determinados agentes económicos que están obligados a informar directamente al propio Banco todas aquellas operaciones que hayan realizado con unidades no residentes (transferencias bancarias identificadas como viajes, compra y venta de moneda extranjera en bancos y oficinas de cambio, movimiento de divisas entre bancos nacionales y extranjeros y pagos con tarjeta de crédito).

La información es accesible a partir de la página Web del banco de España, en concreto: http://www.bde.es/informes/be/balpag/balpag.htm.

1.6.2.3. Estadísticas financieras y monetarias

A) Efectos de comercio devueltos impagados (ECI)

El objetivo de la ECI es determinar mensualmente el número y el importe de los efectos comerciales de las entidades de crédito, en cartera y recibidos en gestión de cobro de clientes, que hayan vencido durante el mes de referencia y de éstos, los que hayan resultado impagados. Se obtienen datos sobre número y cuantía de los diferentes efectos, desagregados por clase de entidad: bancos, cajas de ahorros y cooperativas de crédito y territorialmente a escala nacional por comunidades autónomas y provincias.

B) Hipotecas

Esta estadística ofrece información sobre constituciones de hipotecas, es decir, sobre el número de nuevas hipotecas que se constituyen durante el mes de referencia sobre bienes inmuebles y el importe de los nuevos créditos hipotecarios correspondientes a dichas hipotecas. A partir de 2006 se publica información sobre cambios y cancelaciones registrales de hipotecas. Toda esta información se desagrega en base a diversas variables como naturaleza de la finca hipotecada o entidad prestamista.

Toda la información sobre constituciones, cambios y cancelaciones registrales de hipotecas se obtiene a partir de la información contenida en los Registros de la Propiedad de todo el territorio nacional

C) Sociedades mercantiles

El objetivo de esta estadística es obtener información de carácter mensual sobre las sociedades creadas, las sociedades disueltas y de aquellas en las que se ha producido modificaciones de capital.

Se obtienen datos por provincias, comunidades autónomas y total nacional a partir de los datos suministrados por el Registro Mercantil Central que recoge información de todo el territorio nacional, incluidas Ceuta y Melilla.

D) Estadística del procedimiento concursal

Esta estadística sustituyó a la Estadística de Suspensiones de Pago y Declaraciones de Quiebra del INE, y tiene como objetivo proporcionar información trimestral sobre el sobre el número de deudores concursados, así como del tipo de concur-

so (voluntario o necesario), de la clase de procedimiento (ordinario o abreviado), y de la existencia de propuesta anticipada de convenio y de su contenido (quita, espera, quita y espera u otra proposición), a partir de la información recogida mensualmente de los nuevos Juzgados de lo Mercantil, de los Juzgados de 1ª Instancia y los Juzgados de 1ª Instancia e Instrucción con competencia mercantil.

E) Estadística de transmisiones de derechos de la propiedad

Esta estadística proporciona información mensual sobre las transmisiones de derechos de la propiedad, es decir, sobre el número de derechos de la propiedad que se transmiten a nivel nacional, por provincias y comunidades autónomas. Toda la información sobre transmisiones de derechos de la propiedad se obtiene a partir de la información contenida en los Registros de la Propiedad de todo el territorio nacional.

F) Estadística sobre suspensiones de pagos y declaraciones de quiebra

El objetivo de esta estadística es recoger información sobre el número de expedientes de suspensiones de pagos y declaraciones de quiebra que se inician en los Juzgados de Primera Instancia e Instrucción en el territorio nacional.

La información que se ofrece hace referencia al número de empresas afectadas, activo y pasivo de dichas empresas y clase y causa de los procedimientos, según la actividad económica y condición jurídica de la empresa.

G) Estadística de Ventas a plazos

El objetivo es ofrecer información sobre el volumen de las ventas a plazos de aquellos bienes muebles inscritos en los Registros de Venta a Plazos, analizándose el número de contratos de compraventa de bienes muebles, valor al contado por tipo de financiación, aplazamiento y grupos de bienes. A partir de agosto de 1999 se dejó de realizar esta estadística, por entrar en vigor la Orden de 19 de julio de 1999, que aprobaba la Ordenanza para el Registro de Venta a Plazos de Bienes Muebles. Dicha Ordenanza no contempla la inscripción, en los Registros de algunos de los datos que el INE recogía.

H) Magnitudes monetarias y financieras

Estas estadísticas son elaboradas por el Banco de España, y tienen como objetivo ofrecer la información relativa al activo y pasivo de las instituciones financieras, Banco de España y otras instituciones financieras monetarias diferentes del Banco de España, y depósitos de otros sectores residentes en las entidades de depósito por entidad.

I) Mercado bursátil

En este apartado se ofrece información elaborada por distintos organismos sobre la actividad bursátil, en concreto: cuantía de valores negociados en las Bolsas de Valores, índices de cotización de acciones en la Bolsa de Madrid, mercado de deuda en anotaciones en cuenta, acciones y obligaciones por volumen de cotización.

La Comisión Nacional del Mercado de Valores (http://www.cnmv.es/) publica también información económica financiera y estadística elaborada a partir de sus propios Registros Oficiales.

J) Tipos de interés

Esta estadística es elaborada por el Banco de España y tiene como objetivo ofrecer información sobre los tipos de interés legales del mercado financiero y del mercado monetario, tipos de interés legales aplicados por las entidades de crédito y tipos de interés de la Banca y Cajas de ahorro a residentes de la UEM.

1.6.2.4. Comercio exterior

A) Índices de comercio exterior de servicios (ICES)

Los ICES tienen como principal objetivo proporcionar indicadores de la evolución a corto plazo del valor de las exportaciones e importaciones de servicios no turísticos realizadas entre las unidades residentes y las no residentes en España.

Se trata de una operación estadística continua derivada de los datos trimestrales recogidos en la Encuesta de Comercio Internacional de Servicios (ECIS).

Los ICES se presentan para los diez principales tipos de servicios (exceptuando turismo) de acuerdo a la Clasificación Ampliada de la Balanza de Pagos de Servicios (CABPS): transporte, comunicaciones, construcción, seguros, servicios financieros, servicios de informática y de información, royalties y derechos de licencia, servicios empresariales, servicios personales, culturales y recreativos, y servicios gubernamentales (no incluidos anteriormente). Asimismo, se presentan índices de exportaciones e importaciones de servicios por zonas geográficas/países de contrapartida.

B) Resultados de comercio exterior

La fuente estadística básica para estudiar las operaciones de comercio exterior es el Fichero Territorial de Aduanas, elaborado por el Departamento de Informática Tributaria de la Agencia Estatal de Administración Tributaria (AEAT) con datos procedentes del Departamento de Aduanas e Impuestos Especiales de la AEAT. Los ficheros con los datos de todas las operaciones realizadas están disponibles en Internet en la página Web www.aet.es.

La unidad informante es la empresa que efectúa la operación de exportación/expedición o de importación/introducción que está obligada a documentarla en el denominado Documento Aduanero (DUA), recogiéndose información sobre el importe de la operación, el peso de las mercancías, la provincia y el país de origen/destino, la provincia correspondiente a la aduana y el modo de transporte utilizado.

En 1988 la Comunidad Europea elaboró e implantó un nuevo Arancel de Aduanas: la Nomenclatura Combinada y, en el marco de ésta, para facilitar su informatización, se creó el Arancel Integrado Comunitario o TARIC, cuyas subpartidas quedan identificadas por un código compuesto por once cifras en total. La versión nacional de la TARIC entró en vigor el 1 de enero de 1988.

Desde el 1 de enero de 1993, fecha del establecimiento del Mercado Único en la Unión Europea, las operaciones de comercio exterior en sentido estricto se refieren únicamente al intercambio de bienes y mercancías con países terceros, es decir, fuera de la Unión Europea. Por el contrario, los realizados con Estados miembros de la UE constituyen el comercio intracomunitario, en el que los términos de exportación e importación se sustituyen por los de expedición e introducción respectivamente.

A partir de esa fecha las operaciones del Comercio Exterior se determinan por la suma de los siguientes conceptos:

- Existe la obligación de presentar la declaración DUA ante la aduana para todas las operaciones de exportación e importación cuya procedencia o destino corresponda a países no comprendidos en la Unión Europea. Estas operaciones se contabilizan como se hacía antes de 1993.
- Para las introducciones y expediciones intracomunitarias (operaciones en que se compra o vende a otro país comunitario), se debe presentar la declaración correspondiente al mes anterior en los doce primeros días naturales de cada mes por vía telemática, mediante formulario electrónico en Internet o en soporte papel en las oficinas de INTRASTAT.

A partir del 1 de enero de 1999 quedó suprimida la clasificación según el tipo de obligación del operador. Se establece un único umbral de asimilación quedando suprimida la declaración simplificada, y subsistiendo única y exclusivamente la declaración detallada. Según la Orden del Ministerio de Hacienda de diciembre de 2000 se establece un solo umbral de asimilación que queda fijado en 100.000 Euros, por lo que están obligados a presentar declaración INTRASTAT en los ejercicios 2001 y 2002 los operadores que alcancen un importe facturado de introducciones o de expediciones igual o superior a 100.000 Euros en el ejercicio precedente.

C) Tipos de cambio

En esta sección se recogen datos sobre tipos de cambio proporcionados por el Banco de España y que el INE publica dentro de su Boletín Mensual de Estadística.

1.6.2.5. Información tributaria

A) Estadística sobre el mercado de trabajo y las pensiones

Las entidades, empresas y administraciones privadas y públicas, que realizan pagos por salarios, pensiones y servicios profesionales están obligadas a practicar retenciones e ingresar su importe en la Hacienda Pública por medio de autoliquidaciones mensuales o trimestrales o mediante formalización de crédito en el caso de organismos de la Administración del Estado. Al terminar cada año natural, dichas entidades están obligadas a presentar una Declaración resumen Anual de Retenciones sobre las rentas del Trabajo personal que se acompaña de una Relación de Perceptores en la que figuran los datos identificativos de cada perceptor, las retribuciones satisfechas y las retenciones practicadas. Igual obligación afecta a las entidades que operan en los territorios forales, respecto a las respectivas Haciendas Forales.

El Instituto de Estudios Fiscales elabora una estadística que presenta la información de forma agregada por provincias, sexo, sector institucional, dimensión de la entidad pagadora, actividad de la empresa, etc.

La información se publica en la Web del Instituto de Estudios Fiscales. (*http://www.ief.es/Investigacion/Estadistica/EstadBaseTrib.htm*).

B) Información tributaria

La página Web de la AEAT publica diversas estadísticas, entre las que destacamos las siguientes:

- Informe Anual de Recaudación Tributaria de la AEAT y Resúmenes informativos de Recaudación por Tributos Cedidos y Concertados de la Inspección General del Ministerio de Economía y Hacienda. Estas estadísticas ofrecen la información sobre la recaudación de tributos por las administraciones públicas.
- Resultados económicos y Tributarios en el IVA. Es una investigación de carácter censal basada en la información que suministran los agentes económicos a través del «Resumen Anual».
- La estadística Cuentas anuales en el Impuesto sobre Sociedades está basada en las declaraciones anuales del Impuesto sobre Sociedades que presentan a la Administración Tributaria las entidades con domicilio fiscal en el Territorio de Régimen Fiscal Común.
- Estadísticas de IRPF por partidas. Los datos del IRPF tienen su origen en la agregación de algunas de las variables tributarias que se recogen en los modelos de impresos en los que se presenta la declaración-autoliquidación del IRPF. No se incluye la información de las declaraciones presentadas ante las Haciendas Forales (C.F. de Navarra y País Vasco). Los datos proceden de la AEAT y de la base de datos BADESPE que elabora el Instituto de Estudios Fiscales, ambos pertenecientes al Ministerio de Economía y Hacienda.

1.6.3. Estadísticas a empresas

Dado el importante número de operaciones estadísticas que se recogen en este apartado, describimos a continuación las más importantes, desde el punto de vista del análisis económico de la actividad desarrollada por las empresas.

A) Encuesta sobre la Estructura de las Explotaciones Agrícolas

La Encuesta sobre la estructura de las explotaciones agrícolas se efectúa en todos los países miembros de la Unión Europea. España participó por primera vez en este programa comunitario con la encuesta por muestreo del año 1987. Con posterioridad se han realizado los censos en los años 1989 y 1999 y las encuestas por muestreo en los años 1993, 1995, 1997 y 2003. En esta encuesta participan organismos regionales de estadística, en concreto, el Instituto Vasco de Estadística (EUSTAT), el Instituto de Estadística de Cataluña (IDESCAT) y el Instituto Gallego de Estadística (IGE) en el ámbito territorial de sus comunidades, de acuerdo a los convenios firmados entre el INE y los respectivos Institutos de Estadística.

La encuesta de 2007 tiene como objetivos fundamentales los siguientes:

a) Evaluar la situación de la agricultura española y seguir la evolución estructural de las explotaciones agrícolas, así como obtener resultados comparables entre todos los Estados miembros de la Unión Europea.
b) Cumplir con la normativa legal fijada por la Unión Europea en los diferentes reglamentos del Consejo, así como atender a los requerimientos estadísticos nacionales y otras solicitudes internacionales de información estadística acerca del sector agrario.

Se analizan las producciones obtenidas, superficie de las explotaciones, régimen de tenencia y la mano de obra.

B) Estadística Minera de España

La Estadística Minera de España, elaborada por el Ministerio de Industria, Turismo y Comercio es la información más completa sobre la producción del sector extractivo en España.

El procedimiento de recogida de datos, depuración, gestión y publicación se realiza a través de Internet mediante certificado digital o DNI electrónico. Es de carácter anual.

Se ofrece información sobre:

— Número de explotaciones.
— Empleo: número de empleados remunerados, horas trabajadas, costes de personal.
— Consumo de materiales.

— Consumo de energía.
— Servicios empleados.
— Producción: producción vendible (los datos referentes a la producción y a los consumos de materiales y productos energéticos vienen expresados en cantidad y valor).
— Inversiones efectuadas durante el año de referencia, desglosadas por provincias y referidas al total de productos energéticos, minerales metálicos y minerales no metálicos, de acuerdo con la siguiente presentación:
— Subvenciones.
— Salarios pagados.
— Impuestos: licencia fiscal y otros impuestos y tasas.
— Variación de existencias de productos terminados y en curso de fabricación.

La información aparece desglosada según el tipo de producción, intervalos de empleo y provincias.

C) Estadísticas y balances energéticos

El Ministerio de Industria, Turismo y Comercio es también el responsable de la elaboración de las Estadísticas y balances energéticos. Las estadísticas energéticas elaboradas son las siguientes.

— Destilación de Carbón (Anual y Mensual).
— Estadísticas Eléctricas (Anual y Mensual).
— Estadísticas Gas GLP (Anual).
— Estadísticas Gas Natural (Anual).
— Estadísticas de Refinerías de Petróleo (Mensual).

Dentro de éstas, destaca la Estadística sobre la Industria de la Energía Eléctrica, investigación censal de carácter mensual y anual, que recoge datos físicos y económicos sobre el sector productor y distribuidor de energía eléctrica: número y tipo de centrales, potencia instalada, producción por tipo de central, costes de generación y costes de mantenimiento, intercambios y distribución de energía eléctrica. Esta estadística:

a) Es la principal fuente de información y la base imprescindible para realizar cualquier estudio con profundidad del sector energético.
b) Es fundamental para la realización de balances energéticos en el ámbito europeo, nacional o regional.

Los consumos de energía eléctrica se muestran con una sectorialización de 34 ramas de actividad. En el ámbito geográfico se distinguen las producciones por cuencas, provincias y Comunidades Autónomas. Respecto al tipo se distinguen las centrales de servicio público y las autoproductoras.

Respecto a los Balances energéticos, anualmente se elabora El libro de La Energía en España, que recoge la evolución del mercado energético en España durante el año corriente, con análisis detallado de los balances energéticos y precios así como de las nuevas disposiciones legales de ordenación del sector. También se elaboran informes de coyuntura trimestral y mensual.

D) Encuesta Industrial de Empresas y Encuesta Industrial de Productos

La principal estadística estructural que se realizaba en España sobre el sector industrial es la denominada «Encuesta Industrial», operación estadística iniciada por el INE en 1978 y que desde entonces ha venido ofreciendo información económica sobre el sector con periodicidad anual. La Encuesta Industrial se realizaba a una amplia muestra de establecimientos industriales que eran encuestados en su mayor parte por el INE, si bien establecimientos de determinados sectores (sectores delegados) eran investigados por los Ministerios de Industria y Energía y de Agricultura y Pesca. A los establecimientos se le solicitaba información de su estructura económica-financiera, y de detalle de su producción y compras.

En 1993 se produjo un importante cambio metodológico en esta Encuesta, que pasó a denominarse Sistema Integrado de Estadísticas Industriales, ya que incluye diferentes Encuestas, todas ellas ajustadas a unas definiciones, nomenclatura y criterios metodológicos comunes, que cubren hoy día las necesidades de información económica relativas al sector industrial español.

Estas investigaciones son:

— Encuesta Industrial anual de empresas.
— Encuesta Industrial anual de productos.
— Encuesta de I+D.
— Encuesta Industrial Anual de Empresas 1995 (Consumos e Inversiones).

La encuesta industrial anual de empresas, a diferencia de la antigua, adapta las principales variables del cuestionario a los criterios y normas del Plan General de Contabilidad. Este nuevo enfoque en la recogida de información del sector industrial, ha sido uno de los principales cambios metodológicos que se han incorporaron a la encuesta.

Con el objetivo de ajustar los datos demandados (y su mayor o menor especificación) a las características propias de cada unidad, se diseñan dos modelos diferentes de cuestionarios: un modelo dirigido a las empresas con 20 o más personas ocupadas, y otro, simplificado, con un número reducido de variables, para las empresas con menos de 20 ocupados.

Las variables que investiga la encuesta industrial de empresas son a grandes rasgos las siguientes:

— Personal no remunerado.
— Personal remunerado.
— Horas trabajadas.

— Ventas netas de productos.
— Ventas netas de mercaderías.
— Prestaciones de servicios.
— Trabajos realizados para el inmovilizado.
— Subvenciones a la explotación.
— Otros ingresos de explotación.
— Total de ingresos de explotación.
— Variación de existencias de productos.
— Consumo de materias primas.
— Consumo de otros aprovisionamientos.
— Consumo de mercaderías.
— Trabajos realizados por otras empresas.
— Sueldos y salarios.
— Indemnizaciones.
— Cargas sociales.
— Arrendamientos y cánones.
— Servicios de profesionales independientes.
— Suministros.
— Otros servicios exteriores.
— Dotaciones para amortización del inmovilizado.
— Total de gastos de explotación.
— Inversión realizada en activos materiales e inmateriales.

Desde el punto de vista geográfico, la encuesta industrial anual de empresas cubre el conjunto del territorio nacional, a excepción de Ceuta y Melilla. A los efectos de su explotación estadística, la encuesta fue diseñada para permitir ofrecer resultados con detalle territorial de Comunidad Autónoma, si bien hay que tener presente que la información relativa a las empresas multilocalizadas se basa en una regionalización de la actividad económica de la empresa utilizando para ello las siguientes variables para las que se pide detalle para cada UAE local:

— Número medio de personas ocupadas.
— Importe neto de la cifra de negocios.
— Sueldos y Salarios.
— Consumo de materias primas.
— Inversión en activos materiales.
— Inversión en activos inmateriales.

La Encuesta Industrial de Productos tiene como objetivo ofrecer detalle de la producción comercializada por la industria en España. A diferencia de la encuesta industrial de empresas la unidad de información es el establecimiento industrial. Se investigan los establecimientos de empresas industriales de más de 20 personas ocupadas en el DIRCE.

La producción se mide tanto en términos de cantidad como de valor, que se define como precio medio de venta neto, es decir, incluyendo los costes de envasado y excluyendo el IVA y otros impuestos sobre los productos, rappels y otros des-

cuentos realizados, y gastos de transporte facturados por separado. La producción se clasifica en código PRODCOM.

E) Encuesta sobre la Estructura de la Construcción

Es la estadística estructural sobre el sector más relevante en España. Dicha encuesta la elabora el Ministerio de Fomento pero fue diseñada con la colaboración del INE, por lo que el cuestionario en que se recogen los datos es bastante similar al de la Encuesta Industrial Anual.

La Encuesta de Estructura de la Construcción se viene realizando desde 1980, si bien fue en 1987 cuando se extendió el objeto de la investigación a los empresarios autónomos del sector, y en 1988 se extendió la encuesta a las empresas de instalación, montaje y acabado de edificios y obras que hasta entonces eran encuestadas por el INE.

Los principales objetivos de la Encuesta de la Estructura de la Construcción son:

a) Conocer las principales macromagnitudes del Sector de la Construcción, base esencial para la elaboración de las Cuentas Nacionales y Regionales.
b) Obtener un conjunto de información detallada, actualizada, fidedigna y completa del Sector de la Industria de la Construcción a nivel nacional y homogéneo a los diferentes sectores industriales que permita conocer las principales características estructurales de la industria en su conjunto.
c) Obtener series temporales homogéneas de resultados, con definiciones y criterios que permitan la comparación con datos de otros países.
d) Por último, servir como instrumento necesario para la ejecución de estudios y análisis relativos a los factores de producción utilizados y otros elementos que permitan medir la actividad, rendimiento y competitividad de las empresas, así como la estructura y la evolución de las mismas con el objeto de comparar su actividad y rendimiento con los de los competidores de su sector a escala regional, nacional e internacional.

La unidad estadística investigada es la «empresa» definida como la unidad productiva que ejerce exclusiva o principalmente la actividad de la construcción, ya que, esta unidad tipo es más adecuada que el «establecimiento» para el estudio del Sector de la Construcción.

Las variables contempladas son básicamente las mismas que la Encuesta Industrial de Empresas, es decir, personal ocupado, horas trabajadas, volumen de negocio, valor de producción, consumos intermedios, impuestos, costes de personal, inversión bruta, ventas de bienes de inversión, subvenciones a la explotación, y gastos en I+D.

La información aparece desagregada según tipo de actividad y estrato de número de empleados, así como también se ofrecen datos con menor detalle por comunidades autónomas, en concreto, sobre el número de empresas, la cifra de negocio, los salarios, el empleo y la inversión realizada.

F) Encuesta Anual de Servicios (EAS)

La Encuesta Anual de Servicios 2000 se configura como la primera operación en la que el INE ofrece una visión integrada del sector servicios, ya que hasta ahora los estudios realizados sobre este sector se han venido realizando de manera discontinua y dispersa.

Realizando una breve reseña sobre el sistema español de producción de estadísticas estructurales sobre el sector servicios podemos afirmar que es aún muy reciente, ya que exceptuando alguna investigación específica (encuesta de comercio interior de 1988), todas las estadísticas de servicios realizadas datan de los años 90. A comienzos de esta década el INE diseñó un plan estadístico para este sector que habría de ser desarrollado en plazos de cinco años, mediante encuestas de periodicidad anual y plurianual. El objetivo era que cada año se abordaban diferentes sectores de forma que en un periodo aproximado de cinco años se hubiera investigado todo el sector de servicios.

El plan no se ha cumplido en los términos previstos, si bien a lo largo de la década la mayor parte de las actividades de servicios se han investigado al menos una vez (nunca se han investigado las actividades de transporte por tubería, actividades inmobiliarias, intermediación financiera, y servicios de investigación y desarrollo). Del conjunto de encuestas realizadas hay que destacar su falta de unidad metodológica: objetivos, unidades de encuesta, ámbitos territoriales y cuestionarios diferentes; lo que se traduce en la imposibilidad actual de tener un conocimiento conjunto del sector de servicios en España.

Las encuestas a empresas de servicios realizadas por el INE son las siguientes, entre paréntesis aparece el año en que se realizó la encuesta:

— Servicios Audiovisuales (1992,1997).
— Servicios Técnicos (1992,1997).
— Servicios de Consultoría (1993).
— Servicios de Publicidad (1994).
— Servicios de Alquiler de vehículos, Maquinaria y equipo (1994).
— Servicios informáticos (1995).
— Servicios industriales de limpieza (1995).
— Servicios jurídicos (1996).
— Servicios de investigación y seguridad (1996).
— Servicios empresariales diversos (1996).
— Servicios personales (1997).
— Estadística de empresas de transporte de viajeros (anual desde 1990).
— Interurbano (regular y discrecional) por carretera.
— Transporte urbano de viajeros por autobús.
— Transporte metropolitano.
— Transporte aéreo.
— Transporte ferroviario.
— Transporte marítimo.
— Estadística de turismos de servicio público (1996).
— Estadística de empresas de turismo de servicio público (1998).

— Encuesta sobre la estructura de las empresas de mercancías por carretera 1992, 1993).
— Encuesta económica de empresas de actividades anexas a los transportes y las comunicaciones (1994).
— Encuesta de servicios postales y telecomunicaciones (1995, 1996).
— Encuesta sobre la estructura de las empresas de restauración (1989, 1994).
— Encuesta sobre la estructura de las empresas hoteleras (1991, 1992, 1996).
— Encuesta sobre la estructura de las empresas de agencias de viajes (1993, 1997).
— Encuesta de comercio interior (1994/95, 1988, 1992).

Con el propósito de suministrar la información sobre el sector de servicios a que obliga el reglamento del Consejo de la Unión Europea de 20 de diciembre de 1996 relativo a las estadísticas estructurales de las empresas, el INE puso en 1999 en marcha una nueva encuesta sobre el sector de servicios en España de carácter bianual. Dicha encuesta se realiza para las secciones H a K de la C.N.A.E. 1993 (excepto la sección de intermediación financiera).

Desde el año 2000 esta encuesta es de carácter anual y pretende cubrir los siguientes objetivos:

a) Estudio de las características estructurales y económicas de las empresas que componen el sector más importante de la economía española en términos del Producto Interior Bruto y creación de empleo.
b) Es un instrumento útil parara la Contabilidad Nacional y Regional.
c) Comparar la situación actual con anteriores años y, también, pretende ser comparable internacionalmente.
d) Servir de base para estudios estructurales y de actividad respecto a otros sectores económicos.
e) Evaluar la calidad de los directorios utilizados en las encuestas y la adecuación de los cuestionarios a la realidad.
f) La selección de la muestra se realiza teniendo en cuenta el sector de actividad de la empresa, el estrato de número de empleados y, a veces, la comunidad autónoma.

Las variables contempladas son básicamente las mismas que la Encuesta Industrial de Empresas y la de Estructura de la Construcción, donde la información aparece desagregada con los mismos criterios que en el caso anterior.

G) Las estadísticas sobre turismo del Instituto de Estudios Turísticos (IET)

Creado en 1962, el IET es, según lo establecido por el Real Decreto 561/2009, de 8 de abril del año 2009 (B.O.E. del 22 de abril de 2009), por el que se aprueba el Estatuto del Instituto de Turismo de España (Turespaña), un organismo público con rango de Subdirección General, dependiente directamente de la Presidencia de Turespa-

ña, que tiene como funciones la investigación de los factores que inciden sobre el turismo, así como la elaboración, recopilación y valoración de estadísticas, información y datos relativos al turismo. También le atribuye el mencionado Real Decreto, la creación y difusión del conocimiento y la inteligencia turística y la coordinación de la información sobre el sector turístico generada por las distintas unidades administrativas dependientes de la Secretaría de Estado de Turismo y del propio Turespaña.

El IET como productor de información, es el encargado de las operaciones estadísticas Movimientos Turísticos en Fronteras (Frontur), Encuesta de Gasto Turístico (Egatur) y Movimientos Turísticos de los Españoles (FAMILITUR), generando datos sobre las llegadas de visitantes extranjeros a nuestro país, sus peculiaridades, gastos que realizan y viajes realizados por los españoles y sus características.

El IET de igual modo explota información estadística procedente de fuentes externas, como son la evolución de llegadas de pasajeros extranjeros en compañías de bajo coste y el empleo en el sector turístico, información que divulga de manera periódica. Asimismo difunde otra información estadística de interés procedente de otros organismos.

Una gran parte de estos contenidos, así como la posibilidad de consultar la documentación turística existente en el Centro de Documentación Turística de España (CDTE), se encuentra disponible a través de esta Web (www.iet.tourspain.es).

Describimos a continuación las principales características de las tres estadísticas que produce el IET.

La Encuesta de Gasto Turístico en España (EGATUR)

La Encuesta de Gasto Turístico en España (EGATUR) es realizada por el Instituto de Estudios Turísticos con la colaboración del INE y del Banco de España; tiene como objetivo medir tanto el gasto realizado por los visitantes no residentes en España durante sus viajes a España, así como el gasto efectuado por los residentes españoles durante sus viajes al extranjero.

La Encuesta se inicia en el año 2000, tiene periodicidad mensual e investiga en la frontera a los visitantes no residentes en España cuando abandonan el territorio español y a los residentes españoles cuando regresan al territorio nacional. El trabajo de campo se lleva a cabo en los principales puestos fronterizos (carreteras que comunican con Portugal, Francia y Andorra, aeropuertos, puertos y ferrocarriles).

Esta estadística tiene representatividad nacional, aunque algunas Comunidades Autónomas realizan estimaciones propias mediante la regionalización de la información; la información que aporta tiene gran interés para los estudios de demanda de turismo extranjero, pero su interés fundamental es su contribución para la estimación de las rúbricas de ingresos y pagos por turismo de la Balanza de Pagos y del consumo de los no residentes en la Contabilidad Nacional de España.

Clasifica la información por tipo de entrada, duración de la estancia y nacionalidad de los visitantes, aportando detalles sobre aspectos tales como:

— El Motivo del viaje (turistas) o visitas (excursionistas).
— La frecuencia de las visitas.

— La duración de la estancia.
— El tipo de alojamiento.
— La forma de organización.
— El tamaño del grupo turístico.
— Las actividades realizadas.
— El grado de satisfacción del viaje.
— La fidelidad de las visitas a España.
— Los gastos realizados y su distribución, diferenciando el llevado a cabo en el país de origen (desglosados por conceptos como transporte, alojamiento, alquiler de vehículos, excursiones, espectáculos, etc.) y el realizado en el país de destino (también desglosado por los principales conceptos).
— La forma de pago (tarjetas, dinero en efectivo, cheques, etc.

Los Movimientos Turísticos de los Españoles (FAMILITUR)

Es una Encuesta mensual elaborada por la Secretaría de Estado de Comercio y Turismo y en concreto por el Instituto de Estudios Turísticos mediante entrevista personal directa a residentes mayores de 16 años en municipios de más de 5.000 habitantes; tiene como objetivo la estimación del número y características de los viajes realizados por la población residente en España.

Existe información disponible desde 1996 que cuantifica los flujos de viajeros españoles entre las distintas Comunidades Autónomas y desde España hacia el extranjero, caracterizando los mismos en función de aspectos como el motivo del viaje, la duración de la estancia, el tipo de alojamiento utilizado y el destino, etc.

La información se publica con un desfase de tres meses con respecto al cuatrimestre de referencia; en ella se incluye, con un nivel de desagregación autonómico, los resultados sobre número de viajes, su origen y destino, motivaciones, tipo de alojamiento, medio de transporte y gasto efectuado, duración y temporalidad, etc.

Las Encuestas en fronteras: Movimiento Turístico en Fronteras (FRONTUR)

La Encuesta sobre Movimiento Turístico en Fronteras (FRONTUR) se realiza de forma continuada por el Instituto de Estudios Turísticos en colaboración de la Dirección General de Tráfico, AENA, RENFE y Puertos del Estado. Tiene como fin estimar el número de entradas de viajeros por los distintos puntos de acceso, así como proporcionar información sobre el comportamiento turístico de los visitantes no residentes en el territorio nacional.

La información se obtiene mediante encuestas y aforos muestrales que complementan otras informaciones administrativas proporcionadas por los organismos responsables del tráfico de pasajeros en carreteras, aeropuertos, puertos y trenes.

Los principales objetivos de FRONTUR son: cuantificar mes a mes el número de visitantes que llegan a España por las distintas vías de acceso (carretera, aeropuerto, puerto marítimo y ferrocarril), caracterizarlos según su tipología de visitan-

te, diferenciando los turistas (que realizan al menos una pernoctación) de los excursionistas (que no pernoctan) y conocer el comportamiento turístico en sus desplazamientos dentro de España en relación con las principales características de sus viajes.

Se publica información mensual con 15 días de demora tras la finalización del mes de referencia sobre el número de visitantes, turistas y excursionistas, que llegan a España por las distintas fronteras, sus países de residencia, las Comunidades Autónomas de destino, el tipo de alojamiento utilizado, la duración de la estancia, la forma de organización del viaje (paquete turístico o no), el motivo del viaje y por último edad y sexo.

Anualmente y para las principales temporadas vacacionales se publica información adicional con un mayor desglose de estas variables (motivos y tipos de alojamiento, fidelidad de los turistas, actividades realizadas, grado de satisfacción, uso de Internet en la planificación de sus viajes, etc.).

1.6.4. Estadísticas económicas sobre las Administraciones Públicas

Entre estas estadísticas destacan las Cuentas del IGAE, que son elaboradas cada año por la Intervención General de la Administración del Estado (IGAE) con la información sobre las liquidaciones presupuestarias de cada uno de los organismos de la Administración Central, de las Comunidades Autónomas, de las Corporaciones Locales y de los Organismos Autónomos Administrativos de las Comunidades Autónomas.

Otras estadísticas que miden la actividad económica de las administraciones públicas, son la Estadística del Gasto Público en Educación y la Cuenta Satélite del Gasto Sanitario, elaboradas, respectivamente, por el Ministerio de Educación y por el Ministerio de Sanidad y Política Social.

1.6.5. Estadísticas sobre consumo y precios

A) La Encuesta de Presupuestos Familiares (EPF)

La EPF iniciada en enero de 2006, sustituye a la Encuesta Continua de Presupuestos Familiares (ECPF) base 1997, que con periodicidad trimestral se realizó desde 1997 hasta 2005. La nueva encuesta suministra información anual sobre la naturaleza y destino de los gastos de consumo, así como sobre diversas características relativas a las condiciones de vida de los hogares.

Los gastos de consumo se refieren tanto al flujo monetario que destina el hogar al pago de determinados bienes y servicios de consumo final, como al valor de los bienes percibidos en concepto de autoconsumo, salario en especie, comidas gratuitas o bonificadas y alquiler imputado a la vivienda en la que reside el hogar (cuando es propietario de la misma o la tiene cedida por otros hogares o instituciones). Los gastos se registran en el momento de adquisición, independientemente de que el pago sea al contado o a plazos.

El tamaño de muestra es de aproximadamente 24.000 hogares al año. Cada hogar permanece en la muestra dos años consecutivos, renovándose cada año la mitad de la muestra.

La encuesta proporciona estimaciones del gasto de consumo anual para el conjunto nacional y las comunidades autónomas y del consumo en cantidades físicas de determinados bienes alimenticios para el conjunto nacional.

B) Índice de Precios al Consumo (IPC)

El Índice de Precios de Consumo (IPC) mide la evolución mensual del nivel de precios de los bienes y servicios de consumo adquiridos por los hogares residentes en España. Es un índice de tipo Laspeyres, cuyo procedimiento de cálculo se estudiará más adelante.

En el Sistema Base 2006 se utiliza la definición de gasto de consumo de la EPF: «el gasto de consumo es el flujo monetario que destina el hogar y cada uno de sus miembros al pago de determinados bienes y servicios, con destino al propio hogar o para ser transferidos gratuitamente a otros hogares o instituciones».

Los distintos bienes y servicios de consumo se clasifican de acuerdo con la clasificación armonizada COICOP (Classification Of Individual Consumption by Purpose). La precisión con la que este indicador coyuntural mide la evolución del nivel de precios depende de dos cualidades que todo IPC debe tener: representatividad y comparabilidad temporal. El grado de representatividad del IPC viene determinado por la adaptación de este indicador a la realidad económica del momento; así, la tasa de variación calculada a partir del IPC se aproximará más a la evolución del conjunto de precios de la economía, cuanto más se adapten los elementos seleccionados para su medición a las pautas de comportamiento de los consumidores. Para conseguirlo, los artículos seleccionados que forman parte de la cesta de la compra deben ser los más consumidos por la mayoría de la población, los establecimientos de la muestra deben ser los más visitados, y la importancia relativa de cada artículo en la cesta de la compra debe responder a las tendencias de consumo de los hogares.

Por otro lado, el IPC es un indicador que sólo tiene sentido cuando se establecen comparaciones en el tiempo; de hecho, un número índice no tiene apenas significado si no se establece una comparación con índices de otros períodos, para obtener las tasas de variación correspondientes (puede ser un mes, un año, o cualquier otro período de tiempo). Por ello, la otra cualidad atribuible a un IPC es la comparabilidad temporal, es decir, la necesidad de que los elementos que definen el IPC permanezcan estables a lo largo del tiempo excepto, lógicamente, los precios recogidos mensualmente. De esta forma, se consigue que cualquier variación en el IPC sea sólo debida a cambios en los precios de los artículos seleccionados, y no a cualquier cambio en el contenido metodológico de este indicador.

Las aplicaciones del IPC son numerosas y de gran importancia en los ámbitos económico, jurídico y social. Entre ellas cabe destacar su utilización como medida de la inflación. También se aplica en la revisión de los contratos de arrendamiento de inmuebles, como referencia en la negociación salarial, en la fijación de las pen-

siones, en la actualización de las primas de seguros y otros tipos de contrato, y como deflactor en la Contabilidad Nacional.

C) Índice de Precios de Vivienda (IPV)

El IPV, base año 2007, tiene como objetivo la medición de la evolución de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano, a lo largo del tiempo. La fuente de información utilizada para el cálculo del IPV procede de las bases de datos sobre viviendas escrituradas que proporciona el Consejo General del Notariado, de donde se obtienen los precios de transacción de las viviendas, así como las ponderaciones que se asignan a cada conjunto de viviendas con características comunes.

La muestra utilizada para esta estadística comprende todas las viviendas escrituradas en el trimestre de referencia. El proceso de diseño y desarrollo del IPV se ha realizado de forma coordinada entre el INE y EUROSTAT.

D) Índices de Precios desde el punto de vista de la oferta

Respecto al sector agrario, destacamos los **Índices de Precios Percibidos y Pagados por los Agricultores**, que miden la evolución de los precios que percibe el agricultor y el ganadero por la venta de los productos agrarios y la de los precios que se pagan por las compras de materiales y aprovisionamientos. Estos índices son elaborados por el Ministerio de Medio Ambiente y Medio Rural y Marino.

El **Índice de Precios Industriales** (IPRI), elaborado por el INE, es un indicador coyuntural que 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. Para su obtención se realiza una encuesta continua de periodicidad mensual, que investiga todos los meses más de 8.000 establecimientos industriales.

Los **Índices de Precios de Materiales del Sector de la Construcción**, elaborados por el Ministerio de Fomento, miden la variación de los precios de los distintos materiales que se utilizan en las fórmulas tipo, para aplicar a las cláusulas de revisión de precios que figuran en los contratos de las Administraciones públicas. Se estudian los precios de venta, sin incluir el IVA facturado, de distintos materiales susceptibles de ser utilizados en la construcción (cemento, cerámica, maderas, energía, aluminio, acero, cobre y ligantes) para los índices de precios de materiales.

Los **Índices de Precios del Sector Servicios** miden la evolución de los precios, desde el lado de la oferta, de las actividades de las empresas que operan en los sectores de Transporte Marítimo de Mercancías, Transporte Aéreo Regular de Pasajeros, Manipulación de mercancías, Depósito y almacenamiento, Actividades Postales y de Correos, Telecomunicaciones, Programación y Consultoría Informática, Servicios de información, Asesoría Jurídica y Económica, Servicios y análisis téc-

nicos, Publicidad y Estudios de Mercado, Actividades relacionadas con el empleo, Actividades de Seguridad y Actividades de Limpieza. Los precios recogidos corresponden a los servicios suministrados a las empresas (segmento de negocios).

1.6.6. Estadísticas sobre el mercado laboral

A) La Encuesta de Población Activa (EPA)

La EPA es una macroencuesta elaborada mensualmente por el INE; obtiene información sobre la situación laboral de una muestra estadísticamente representativa de unas 200.000 personas; la situación laboral está clasificada de acuerdo con las definiciones de la Organización Internacional del Trabajo y de la Comisión Europea en Activos, Inactivos, Ocupados o Parados.

La situación de los ocupados se diferencia entre los asalariados y trabajadores por cuenta propia y dentro de ambos en distintas categorías; las cifras se desagregan por provincias y grandes sectores (agricultura, industria, construcción y servicios, etc.), aunque pueden pedirse explotaciones específicas al INE en las que se detallen datos más concretos para cualquiera de las variables de la Encuesta.

Los resultados generales se publican después de un mes y medio de finalizar el trabajo de campo y están disponibles en la página Web del INE.

B) La Encuesta de Coyuntura Laboral (Ministerio de Trabajo y Asuntos Sociales)

La Encuesta de Coyuntura Laboral (ECL) es una investigación por muestreo de periodicidad trimestral realizada por el Ministerio de Trabajo y Asuntos Sociales; está dirigida a empresas y proporciona información sobre los efectivos laborales, su composición según diversas características, movilidad laboral, altas, bajas y modificaciones de contrato, jornada laboral efectivamente realizada, horas no trabajadas y sus motivaciones, horas extraordinarias, opiniones de los empresarios respecto a la evolución de sus plantillas, etc.

Los resultados se publican en la página Web del Ministerio de Trabajo y Asuntos Sociales: http://www.mtas.es/estadisticas/ECL/Welcome.htm.

En el mismo sentido esta estadística proporciona información trimestral sobre otros aspectos tales como: tipo de contrato y jornada laboral, tamaño de los centros de trabajo, Comunidades Autónomas, efectivos laborales con remuneración igual al salario mínimo interprofesional, vacantes existentes el último día del trimestre de referencia, etc.

C) Afiliación de Trabajadores al Sistema de Seguridad Social (Ministerio de Trabajo y Asuntos Sociales).

La Secretaría de Estado para la Seguridad Social (Ministerio de Trabajo y Asuntos Sociales) publica mensualmente estadísticas sobre el número de Afiliados

dados de Alta en la Seguridad Social por Regímenes (general, autonómos y especiales) por Comunidades Autonomías y provincias; la información se publica en la Web de la Seguridad Social: http://www.seg-social.es/.

D) Estadística de Paro y Contratos Registrados

El Instituto Nacional de Empleo (INEM) publica mensualmente estadísticas sobre el paro registrado con desagregación de datos a nivel provincial, de edad y sexo de los demandantes de empleo y sectores económicos; en la publicación virtual sólo se proporciona información para grandes sectores (agricultura, Industria, construcción, servicios y sin empleo anterior), aunque existe información con mayor nivel de detalle.

En la misma línea se dispone de información sobre contrataciones por modalidades (indefinidos y temporales).

La información se publica en el apartado de estadísticas de la Web: http://www.inem.es/

E) Encuesta Anual de Coste Laboral (EACL)

La Encuesta Anual de Coste Laboral es una operación estadística de periodicidad anual que completa los resultados obtenidos en la Encuesta trimestral de Coste Laboral (ETCL) obteniendo una perspectiva anual de los mismos. Tiene como objetivo fundamental conocer los niveles anuales del Coste laboral medio por trabajador, detallando sus principales componentes: percepciones salariales, cotizaciones obligatorias a la Seguridad Social, cotizaciones voluntarias, prestaciones sociales directas, indemnizaciones, gastos en formación profesional y otros gastos por emplear mano de obra.

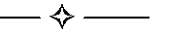
Desde el año de referencia 2008, la EACL obtiene estimaciones del coste laboral clasificado por Comunidades Autónomas, actividad económica según la clasificación CNAE-09 y tamaños de las unidades. Se investigan las cuentas de cotización cuya actividad económica esté encuadrada en los tres grandes sectores económicos: Industria, Construcción y Servicios, en concreto aquellos centros con actividades económicas comprendidas en las secciones de la B a la S de la CNAE-09.

F) Índices de Salarios Agrarios

Estos índices miden la evolución de los ingresos diarios brutos del trabajador y los complementos recibidos en especie valorados en euros. Las ponderaciones utilizadas para el cálculo de los salarios medios se corresponden con el número de trabajadores empleados a nivel nacional en cada categoría.

1.7. EJERCICIOS

Sobre conceptos estadísticos fundamentales y fuentes de datos estadísticos



Ejercicio 1.1. *Una empresa consultora quiere hacer un estudio con el fin de conocer ciertas opiniones sobre el comportamiento turístico de las mujeres de clase media residentes en la ciudad de Barcelona.*

Reflexione sobre cuál sería el universo a investigar.

Respuesta

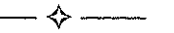
Tal como indica el enunciado, la población la integrarían las mujeres residentes en la ciudad de Barcelona; es necesario, sin embargo, precisar con más detalle las características de este universo.

En concreto, deberíamos decidir la condición de residencia de forma objetiva y concreta en el tiempo, definiendo como residentes, por ejemplo, a las personas inscritas en el Padrón Municipal de Habitantes de la ciudad de Barcelona en la fecha que interese al investigador.

Sería necesario, asimismo, fijar una edad mínima para que cumplan la condición exigida (mujer) y el nivel mínimo y máximo de ingresos que determinan la condición «clase media».

Los fines del estudio podrían aconsejar, por ejemplo, que el mismo estuviese referenciado a las mujeres residentes en Barcelona, con edad comprendida entre 18 y 70 años en el momento de realizar la toma de datos y con un ingreso familiar per cápita de entre 750 y 2000 euros mensuales.

Cualquier grupo de mujeres que cumpla con estos requisitos puede constituir una muestra.



Ejercicio 1.2. *Responda las siguientes preguntas:*

a) *En un determinado periódico, se ha publicado una noticia informándonos de que el 10% de los españoles que viajaron al extranjero durante el año 2009 prefirieron no contratar ningún seguro de viaje. ¿Se llegó a esta conclusión a partir de una muestra o de una población?*

b) *El 15% de las matriculaciones de automóviles realizadas en España durante los últimos cinco años fueron de una determinada marca comercial. ¿Se llegó a esta conclusión a partir de una muestra o de una población?*

Respuesta

a) Se llegó a esa conclusión a partir de una muestra, ya que no podríamos encuestar a todos los españoles que viajaron al extranjero durante 2009 porque, entre otras razones, además del elevado coste, no existe un registro en el que conste, de forma exhaustiva y detallada, esta información.

b) Todas las matriculaciones de automóviles, con las principales características del vehículo, quedan registradas en la Dirección General de Tráfico; sí la información tiene como fuente este registro, la conclusión indicada estará obtenida de una población.

Además, dado que existe este registro, y que es relativamente fácil su consulta, sería mucho más laborioso y caro, inferir esta información mediante la realización de una encuesta muestral.

Ejercicio 1.3. *Indique si las siguientes variables son cualitativas o cuantitativas.*

a) *Profesión de los empleados de cierta empresa.*

b) *Número de clientes de una empresa durante el año 2010.*

c) *Número de viajes de turismo realizados por una persona el último año.*

d) *Volumen de ventas, en miles de euros, de la industria del automóvil en España.*

e) *Grado de satisfacción de los clientes con un servicio dado.*

f) *Tiempo promedio de espera en la conexión entre 2 determinados vuelos en un aeropuerto durante el último mes.*

Respuesta

a) Variable Cualitativa (con modalidades del tipo: directivos, encargados de obra, administrativos, obreros de producción...).

b) Variable Cuantitativa (con valores 1, 2, 3...).

c) Variable Cuantitativa (con valores 1, 2, 3...).

d) Variable Cuantitativa (con valores 1000, 5000...).

e) Variable Cualitativa (con modalidades del tipo: muy satisfecho, poco satisfecho, ..).

f) Variable Cuantitativa (con valores 15, 20 minutos....).

Ejercicio 1.4. *Decida sobre qué población seleccionaría una muestra para analizar.*

a) *El sueldo promedio de los empleados de una gran empresa durante un determinado año.*

b) *El tiempo promedio en la gestión de una determinada subvención a la explotación por las empresas dedicadas a la agricultura.*

c) *El número de personas que viajan en avión entre Barcelona y Madrid de lunes a viernes durante el último año.*

d) *La proporción de personas que eligen cierta marca de detergente en un supermercado.*

Respuesta

a) Se debería elegir un grupo de empleados representativo de la compañía y del período analizado; en este grupo deberían estar adecuadamente representadas todas las categorías laborales de la compañía, ya que, muy probablemente, nuestra variable de interés (sueldo) esté estrechamente ligada a la categoría de los empleados.

b) Se debe consignar el tiempo utilizado por todas las empresas agrarias que han pedido la citada subvención en el período de referencia.

c) Se deberían elegir al azar un número significativo de los aviones que hubiesen volado de lunes a viernes entre Barcelona y Madrid durante el último año.

d) Se debería elegir aleatoriamente un grupo de clientes de dicho supermercado.

Ejercicio 1.5. *Clasifique, por tipos, la siguiente lista de variables:*

— *Peso máximo autorizado de un vehículo de transporte de mercancías.*

— *Edad en años cumplidos del sustentador principal del hogar.*

— *Cantidad de pasajes vendidos durante este mes.*

— *Nivel de educación.*

— *Razón social de una empresa.*

— *Estado civil.*

— *Cantidad de reservas hechas por minuto.*

— *Contenido de un disquete.*

— *Volumen de agua registrado en un embalse.*

Respuesta

— Peso máximo autorizado de un vehículo: variable cuantitativa continua.

— *Edad en años cumplidos* del sustentador principal del hogar: variable cuantitativa discreta.

— Cantidad de pasajes vendidos durante este mes: variable cuantitativa discreta.

— Nivel de educación: variable cualitativa.

— Razón social de una empresa: variable cualitativa.

— Estado civil: variable cualitativa.

— Cantidad de reservas hechas por minuto: variable cuantitativa discreta.

— Contenido de un disquete: variable cualitativa.

— Volumen de agua registrado en un embalse: variable cuantitativa continua.

Nótese que sólo las variables cuantitativas tienen la calificación de continuas o discretas, ya que las variables cualitativas son categóricas (toman atributos o categorías).

Ejercicio 1.6. *Una gran empresa extranjera de fabricación de componentes electrónicos ha decidido expandir su producción instalando una nueva fábrica. De cara a su ubicación, se barajan 3 países, entre los cuales uno de ellos es España. Analice que fuentes estadísticas habría de consultar en España de cara a tomar la decisión sobre donde instalarse (supongamos que esta misma información sea accesible en los otros dos países.*

Respuesta

Para simplificar, supongamos que la decisión final sobre la ubicación depende de dos variables básicas: la situación del mercado y los costes de fabricación existentes en cada país.

Respecto a la situación del mercado, se podrían consultar en primer lugar la Encuesta Industrial de Empresas y la Encuesta Industrial de Productos, obteniendo información sobre los ingresos y gastos de las empresas del sector en España y sobre el volumen de producción.

Otra estadística a analizar son los Resultados del Comercio Exterior, que nos informará del volumen de importaciones necesario para satisfacer la demanda, y del volumen de producción que se exporta.

Otras fuentes alternativas serían las de origen tributario, en concreto los Resultados Económicos y Tributarios en el IVA o las Cuentas Anuales en el Impuesto sobre Sociedades en el sector.

Podrían también consultarse la Contabilidad Nacional de España (CNE), donde podremos analizar la demanda de este tipo de productos según el tipo de consumidor, empresas, hogares, administraciones públicas e instituciones sin fin de lucro, distinguiendo la parte que se produce en España de las importaciones y el Directorio Central de Empresas (DIRCE), para saber el colectivo de empresas y locales que operan en España.

También podría ser interesante analizar los datos de la Estadística del Procedimiento Concursal y el Mercado Bursátil.

En lo relativo a los costes de fabricación, a partir de la Encuesta Anual de Coste Laboral, podremos obtener información sobre los salarios medios que se pagan.

Por último, deberían analizarse los precios medios y la evolución de los mismos. A partir del Índice de Precios Industriales (IPRI), podemos evaluar la evolución de los precios de venta. Conjugando éste con el Índice de Precios del Sector Servicios, tendremos información de la evolución de los correspondientes precios de las materias primas, aprovisionamientos y servicios consumidos.

En definitiva, con todos estos datos, podríamos valorar la situación de las empresas del sector y del mercado en nuestro país. Si obtuviéramos información similar en los otros dos países tendríamos información suficiente para decidir donde instalar la nueva fábrica.

Capítulo 2

DISTRIBUCIONES UNIDIMENSIONALES

En este capítulo se abordan las técnicas de Estadística Descriptiva dedicadas al estudio de las distribuciones de frecuencias de una sola variable y a sus representaciones gráficas.

2.1. DISTRIBUCIÓN O DISTRIBUCIÓN DE FRECUENCIAS

Se denomina distribución o distribución de frecuencias al ***conjunto de valores que toma una variable, adecuadamente ordenados — de mayor a menor o de menor a mayor—, y acompañado de sus frecuencias absolutas,*** es decir, de las veces en las que aparece cada valor.

La notación más habitual que se emplea es la siguiente:

- X (normalmente en mayúscula): La variable o característica objeto de estudio.
- x_i (normalmente en minúscula): El valor que toma la variable o característica X *para* el individuo i.
- r: El número de valores distintos que toma una variable; es común también utilizar el símbolo «k» para esta notación.
- n_i: El número de veces o frecuencia con la que aparece un determinado valor x_i.
- N: El número de unidades en las cuales efectuamos la medición o disponemos de datos.
- n: A veces suele utilizarse la «n» minúscula para hacer referencia al total de datos; lo más habitual, sin embargo, es notar con «N» mayúscula el total de datos referido a una población y «n» minúscula para el total de datos referidos a una muestra de dicha población.

Ejemplo 2.1. Se tira un dado 100 veces y se obtienen:

10 veces la cara 1,
15 veces la cara 2,
25 veces la cara 3,
30 veces la 4,
15 veces la cara 5 y
5 veces la cara 6.

Diremos que la variable X (resultados obtenidos con el lanzamiento de un dado), toma k posibles valores (las caras del dado: 1, 2, 3, 4, 5, 6), de forma que el subíndice i varía entre 1 y 6; los posibles valores que toma X (los x_i) serán:

$$x_1 = 1; \quad x_2 = 2; \quad x_3 = 3; \quad x_4 = 4; \quad x_5 = 5; \quad \text{y} \quad x_6 = 6$$

El valor x_i *aparece* n_i veces, por ejemplo el valor x_1 *es decir*, la cara 1 del dado, aparece n_1 veces, en nuestro experimento, 10 veces y N, o sea, el número total de unidades en las cuales efectuamos la medición, es 100.

Las distribuciones de frecuencias pueden ser *unidimensionales, bidimensionales o multidimensionales*, según observemos una, dos o múltiples características de la población.

2.2. DEFINICIONES

En una distribución de frecuencias unidimensional se definen los siguientes conceptos elementales:

Frecuencia absoluta

Es el ***número de veces que se presenta un valor*** (sí se trata de una variable) ***o un carácter o modalidad*** (sí se trata de un atributo) en la población analizada. Para una variable X, se representa como n_i.

En el Ejemplo 2.1, la frecuencia absoluta sería:

$$n_1 = 10; \quad n_2 = 15; \quad n_3 = 25; \quad n_4 = 30; \quad n_5 = 15; \quad \text{y} \quad n_6 = 5$$

Es decir, en nuestro experimento, la frecuencia absoluta de la cara 1 del dado, es decir el número de veces que ha salido la cara 1, es 10.

Frecuencia total o total de datos

Es la ***suma de todas las frecuencias absolutas*** o lo que es lo mismo el número de datos que tenemos en la distribución. Se denota con la letra N; en el Ejemplo 2.1 sería 100 tiradas.

Frecuencia relativa

Es el ***cociente entre la frecuencia absoluta con la que se presenta un valor o una modalidad y la frecuencia total de datos*** (n_i / N); suele expresarse en tantos por 100 o en tantos por uno y se denota como f_i.

En general la frecuencia absoluta sólo aporta información respecto al experimento en cuestión (en el caso referido $n_1 = 10$ sólo indica que la cara 1 ha salido 10 veces); la frecuencia relativa facilita la generalización de esta información; en nuestro caso diremos que la frecuencia relativa con la que ha salido la cara 1 es $f_1 = 10/100 = 0{,}10 = 10\%$; es decir que la cara 1 ha salido un 10% de las veces; en términos de probabilidad se dice, si el dado no está trucado, a medida que se aumenta el número de lanzamientos todas las caras tienden a salir el mismo número de veces y diremos que la probabilidad de salida de un cara es 1/6, es decir un 16,66%.

En el *Ejemplo 2.1*, la frecuencia relativa de los valores de la variable, expresada en porcentajes, será:

$$f_1 = 10\%; \quad f_2 = 15\%; \quad f_3 = 25\%; \quad f_4 = 30\%; \quad f_5 = 15\%; \quad \text{y} \quad f_6 = 5\%$$

Como puede comprobarse la suma de las frecuencias relativas siempre es 1 (100%).

Frecuencia absoluta acumulada N_i

La frecuencia absoluta acumulada de un determinado valor de la variable (X_i) es la frecuencia absoluta de dicho valor más la ***suma de las frecuencias absolutas de todos los valores anteriores***. Se denota como N_i; para obtener adecuadamente esta frecuencia es necesario que la distribución esté previamente ordenada.

En el Ejemplo 2.1, la frecuencia absoluta acumulada de los lanzamientos sería:

$$N_1 = 10; \quad N_2 = 25; \quad N_3 = 50; \quad N_4 = 80; \quad N_5 = 95; \quad \text{y} \quad N_6 = 100$$

$N_2 = 25$, indica que la cara 1 o la cara 2 han salido 25 veces; $N_5 = 95$ indica que las caras 1, 2, 3, 4 6 5 han salido 95 veces.

Frecuencia relativa acumulada

Llamamos frecuencia relativa acumulada de un determinado valor ordenado X_i a la ***suma de las frecuencias relativas*** f_i de dicho valor y de los valores inferiores a él. Se denota como F_i.

En el ejemplo 2.1, tendríamos:

$$F_1 = 10\%; \quad F_2 = 25\%; \quad F_3 = 50\%; \quad F_4 = 80\%; \quad F_5 = 95\%; \quad \text{y} \quad F_6 = 100\%$$

Esta presentación se denomina también, a veces, como porcentaje acumulado.

2.3. TIPOS DE DISTRIBUCIONES DE FRECUENCIAS

Los datos estadísticos suelen presentarse en tres situaciones diferentes:

a) Los valores no se repiten en ningún caso. Son las denominadas ***Distribuciones de Tipo I*** o ***Distribuciones Unitarias***; se representan como una sucesión del tipo:

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_n$$

Ejemplo 2.2. Los pesos, en Kg. de 3 personas son:

55, 68, 79

---

b) Cada valor de la característica medida se repite un determinado número de veces. Son las denominadas ***Distribuciones de Tipo II***.

Ejemplo 2.3. Las puntuaciones, de 0 a 10, otorgadas por 30 clientes de un hotel sobre la percepción que han tenido de la limpieza general del mismo son las siguientes:

0 4 5 2 9 8 9 0 2 4 9 6 8 2 6 7 4 6 4 3 9 9 8 0 8 8 8 9 3 7

---

c) Cuando trabajamos con variables continuas o con variables discretas que presentan una gran cantidad de valores, resulta conveniente agrupar estos valores en intervalos o clases; cuando las observaciones están clasificadas en intervalos trabajamos con ***Distribuciones de Tipo III***.

Ejemplo 2.4. Una muestra de 500 clientes de un determinado paquete turístico, se distribuye, por edades, de acuerdo con los siguientes intervalos:

0 a 10 años	50 clientes
11 a 18 años	70 clientes
19 a 45 años	20 clientes
46 a 65 años	160 clientes
Más de 65 años	200 clientes

La presentación de datos por intervalos presenta bastantes *ventajas* (en el caso de referencia hemos evitado tener que indicar las 500 edades una tras otra o preguntar a cada uno de los entrevistados por su edad exacta) y algunos *in-*

convenientes (con una presentación de este tipo tenemos una información menos precisa sobre el fenómeno investigado).

2.4. ELABORACIÓN DE TABLAS DE FRECUENCIAS EN DISTRIBUCIONES DE TIPO II

Para la construcción de tablas de frecuencias en Distribuciones de Tipo II se opera, con carácter general, de la siguiente forma:

— Ordenación de los datos:
De menor a mayor o de mayor a menor en el caso de ser valores y en el orden alfabético o el que se considere más apropiado en el caso de tratarse de modalidades (excelente, bueno, regular, malo y pésimo, por ejemplo).
— Recuento de las frecuencias absolutas.
— Construcción de una tabla estadística de frecuencias: n_i, N_i, f_i, F_i.

Ejemplo 2.5. Las valoraciones del servicio recibido por una muestra de 40 usuarios de un paquete turístico son las siguientes:

9 9 3 9 4 9 4 9 2 9 8 9 9 2 1 5 6 7 8 1 6 7 4 6 4 3 9 9 8 0 8 8 8 9 3 7 9 9 9 9

Construir una tabla de frecuencias.

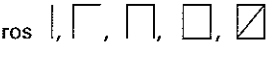
Procederemos de la siguiente forma:

— Ordenamos los valores para facilitar el recuento:

0 1 1 2 2 3 3 3 4 4 4 4 5 6 6 6 7 7 7 8 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

— Construimos la tabla de frecuencias, indicando en la primera columna los distintos valores que ha tomado la variable (de 0 a 9) y en la segunda el número de veces que ha aparecido cada valor n_i; para ello podemos contar directamente o agrupar previamente los datos. Algunas técnicas que nos facilitan el recuento son:

Por los tradicionales «palitos» 

Formando cuadros 

En la tercera columna obtenemos las frecuencias absolutas acumuladas N_i, sumando, a cada n_i, la n_i propia y las n_i anteriores; en la cuarta se obtienen las frecuencias relativas f_i y en la quinta columna las frecuencias relativas acumuladas, F_i.

La tabla de frecuencias quedaría en la siguiente forma:

x_i	n_i	N_i	f_i	F_i
0	1	1	1/40 ⇒ 2,5%	1/40 = 2,5%
1	2	(2+1) = 3	2/40 ⇒ 5,0%	3/40 = 7,5%
2	2	(3+2) = 5	2/40 ⇒ 5,0%	5/40 = 12,5%
3	3	(5+3) = 8	3/40 ⇒ 7,5%	8/40 = 20,0%
4	4	12	4/40 ⇒ 10,0%	12/40 = 30,0%
5	1	13	1/40 ⇒ 2,5%	13/40 = 32,5%
6	3	16	3/40 ⇒ 7,5%	16/40 = 40,0%
7	3	19	3/40 ⇒ 7,5%	19/40 = 47,5%
8	6	25	6/40 ⇒ 15,0%	25/40 = 62,5%
9	15	40	15/40 ⇒ 37,5%	40/40 = 100,0%
	40		40/40 = 1 ⇒ 100%	

2.5. ELABORACIÓN DE TABLAS DE FRECUENCIAS NO UNITARIAS EN DISTRIBUCIONES DE FRECUENCIAS UNIDIMENSIONALES CON DATOS AGRUPADOS EN INTERVALOS

Como indicábamos con anterioridad, cuando trabajamos con variables continuas o con variables discretas que presentan una gran cantidad de valores, resulta conveniente agrupar estos valores en intervalos o clases.

En este tipo de presentaciones es necesario manejar dos conceptos: amplitud del intervalo y marca de clase.

Se denomina **amplitud del intervalo** a la diferencia entre los dos extremos del intervalo.

Según la distribución con la que estemos trabajando y los fines para los que se diseñe la investigación, pueden construirse más o menos intervalos y diseñarse con amplitud (c_i) constante o variable, es decir, con igual o de diferente amplitud cada intervalo.

La notación, más habitual que se emplea para los intervalos es la siguiente:

L_i : Límite superior del intervalo.
L_{i-1} : Límite inferior del intervalo.
c_i : Amplitud de un intervalo $= L_i - L_{i-1}$.

El punto medio de cada intervalo se llama **marca de clase** del intervalo y hace la misma función que el valor de la variable; le designamos por ello, indistintamente, con x_i o m_i.

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

También puede obtenerse mediante la siguiente expresión equivalente:

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

Para evitar confusiones en la clasificación de los individuos es importante definir con claridad los límites de cada intervalo. En el ejemplo que sigue se ha tomado la opción de definir los límites inferiores de cada intervalo con un minuto de diferencia con respecto al límite superior del intervalo anterior, de forma que no quede duda de en que intervalo incluir un cliente que entró, por ejemplo a las 12 horas en punto.

Ejemplo 2.6. Supongamos que queremos planificar los horarios de la plantilla de trabajadores de un establecimiento comercial y que disponemos para ello de información detallada sobre la afluencia de clientes en las distintas horas del día; supongamos que la información se refiere a la hora de entrada de una muestra representativa de 1.000 clientes y que es del siguiente tipo:

Cliente	1	2	3	4	5	6	7	8	...	1.000
Hora	10,01	10,25	17,36	19,21	10,3	10,31	10,32	10,35	...	19,01

Como esta serie resulta excesivamente larga y muy difícil de interpretar, al menos visualmente, conviene agruparla en intervalos; se piensa que para los efectos de nuestra planificación la agrupación más conveniente es en intervalos de dos horas; el centro comercial abre de 10 a 20 horas; los datos agrupados, con su correspondiente tabla de frecuencias podrían quedar de la siguiente forma:

$L_{i-1} - L_i$	n_i	N_i	f_i	f_i (en %)	Fi	Fi (en %)
(20 - 10 h.)	0	0	0/1.000	0%	0/1.000	0%
(10,01 - 12 h.)	550	(550 + 0 =) 550	550/1.000	55%	550/1.000	55%
(12,01 - 14 h.)	100	(550 + 100 =) 650	100/1.000	10%	650/1.000	65%
(14,01 - 16 h.)	50	(650 + 50 =) 700	50/1.000	5%	700/1.000	70%
(16,01 - 18 h.)	100	(700 + 100 =) 800	100/1.000	10%	800/1.000	80%
(18,01 - 20 h.)	200	(800 + 200 =) 1.000	200/1.000	20%	1.000/1.000	100%
	1.000		1	100%		

En la primera columna se observan los intervalos indicados en el párrafo anterior, constándose que la entrada en el establecimiento comercial tiene lugar entre las 10 y las 20 horas y que el resto de los intervalos son de dos horas; en la segunda columna se refleja la frecuencia absoluta de visitantes en cada intervalo; en la tercera se indica la frecuencia absoluta acumulada (entre paréntesis se muestra la forma en la que se va obteniendo la columna, sumando a la N_i anterior la n_i de cada intervalo); en la cuarta

y quinta columnas se obtiene la frecuencia relativa f_i (división de la n_i entre N = 1000) y su expresión en porcentajes y en las dos últimas columnas la frecuencia absoluta relativa acumulada F_i (N_i dividido por N y expresado en porcentajes).

La quinta columna nos indica, por ejemplo, que entre las 10 y las 12 horas entran un 55% de los clientes del establecimiento comercial; en principio parece que esta desagregación de la información es más que suficiente para planificar los horarios del personal, pero cualquier otra, más agrupada o desagrupada, podría valer para trabajar estadísticamente con esta muestra.

Debemos agrupar de acuerdo con los fines que persiga nuestro trabajo y teniendo en cuenta que, probablemente, una vez que hagamos la agrupación y dispongamos de la tabla agrupada, será esta tabla la única información de que dispongamos y no la tabla original con los datos individualizados; así, por ejemplo, sí con la información original hubiésemos generado la tabla anterior, tendríamos información de las entradas de clientes entre las 10,01 y las 12 horas, pero no de los habidas entre las 10 y las 11 horas y sería costoso o incluso imposible volver a desagregar esta información.

Los intervalos pueden ser abiertos y cerrados por alguno de sus extremos o por ambos:

Las notaciones empleadas al efecto son las siguientes:

Tipo de intervalo	Ejemplo	Significado
Intervalo abierto por ambos extremos.	(8 h. - 10 h.) $8 < x < 10$	En el intervalo están excluidas las 8 y las 10 horas en punto
Intervalo cerrado por ambos extremos.	[8 h. - 10 h.] $8 \leq x \leq 10$	En el intervalo están incluidas las 8 y las 10 horas en punto.
Intervalo abierto por el extremo inferior y cerrado por el extremo superior.	(8 h. - 10 h.] $8 < x \leq 10$	En el intervalo están incluidas las 10 horas en punto, pero no las 8 horas en punto, que se incluirá en el intervalo anterior.
Intervalo abierto por el extremo superior y cerrado por el extremo inferior.	[8 h. - 10 h.) $8 \leq x < 10$	En el intervalo están incluidas las 8 horas en punto, pero no las 10 horas en punto, que se incluirá en el siguiente intervalo.

La configuración de los intervalos debe permitir incluir inequívocamente a cada uno de los elementos de la población ***en uno y sólo en uno*** de los intervalos diseñados.

Con este criterio, una agrupación de datos de una variable como la distancia kilométrica entre 100 municipios, podría quedar de la siguiente forma:

$L_{i-1} - L_i$	n_i
(0 - 5 Km.]	5
(5 - 10 Km.]	10
(10 - 25 Km.]	5
(25 - 50 Km.]	50
(50 - 100 Km.]	30
(100 - 500 Km.]	105

Así, la distancia 0 Km. no estaría contemplada en esta distribución, dado que los intervalos son abiertos por el extremo inferior, mientras que al ser cerrados por el extremo superior, la distancia 5 Km. exactos estaría incluida en el primer intervalo y el siguiente intervalo empezaría para los municipios cuya distancia fuese ligeramente superior.

2.6. REPRESENTACIÓN GRÁFICA DE LAS DISTRIBUCIONES

Es conveniente representar gráficamente las distribuciones a partir de sus diversas frecuencias; las figuras más empleadas para este fin son: los diagramas de barras, polígonos de frecuencias, los diagramas acumulativos de frecuencias, los polígonos acumulados de frecuencias, los diagramas de sectores, los cartogramas, pirámides de población, pictogramas, etc.

Describimos a continuación los más utilizados:

A) Gráficos de barras

En general, se emplean para variables discretas en distribuciones de frecuencias con datos sin agrupar; representan los valores de las variables en el eje de abscisas y en el de ordenadas se levanta, para cada punto, una barra con un valor igual a la frecuencia absoluta o relativa.

También se conocen como diagramas de rectángulos; todas las barras o rectángulos tienen la misma base y sus áreas son proporcionales a las frecuencias absolutas n_i.

El alumno debe saber manejar cualquier programa informático de gráficos para permitir la generación de los mismos; aconsejamos a tal fin la hoja de cálculo Excel, con la que se pueden generar buena parte de los gráficos que explicamos a continuación[2].

Ejemplo 2.7. Representar en un diagrama de barras los siguientes datos sobre el número clientes de una empresa durante los diversos meses del año:

Meses	N.º de clientes	Meses	N.º de clientes
Enero	45	Julio	125
Febrero	35	Agosto	110
Marzo	121	Septiembre	120
Abril	18	Octubre	75
Mayo	85	Noviembre	67
Junio	100	Diciembre	99
Total anual		1.000	

[2] Otros programas o aplicaciones informáticas de interés son Minitab, SPSS, PSPP, TeeChart Office, StatGraphics, etc.

La representación gráfica sería la siguiente:

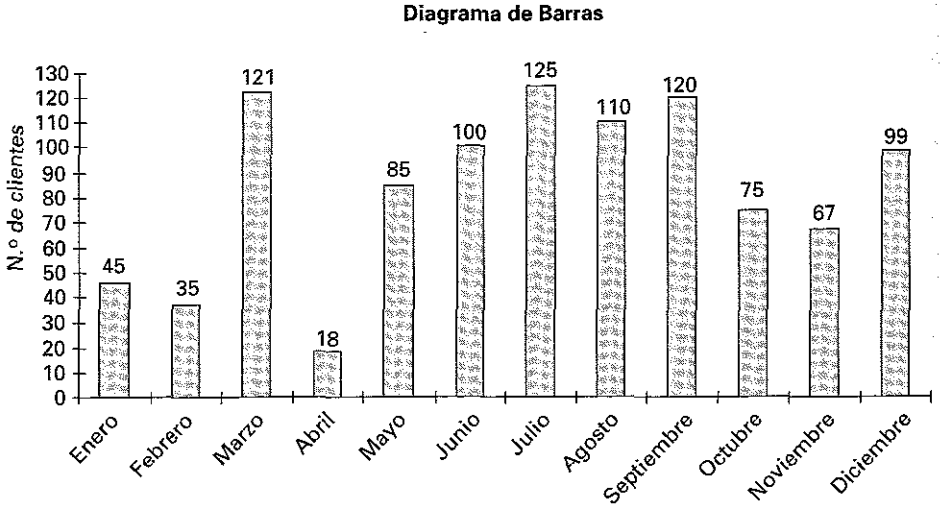


Gráfico 2.1. Ejemplo de diagrama de barras.

Este tipo de gráficos permite comparar varios fenómenos estudiados; supongamos dos empresas A y B, con la siguiente información sobre la distribución de su clientela:

Meses	N.º de clientes de la empresa A	N.º de clientes de la empresa B
Enero	45	25
Febrero	35	38
Marzo	121	139
Abril	18	54
Mayo	85	95
Junio	100	99
Julio	125	115
Agosto	110	135
Septiembre	120	80
Octubre	75	75
Noviembre	67	65
Diciembre	99	80
Total	1.000	1.000

El Gráfico comparativo sería el siguiente:

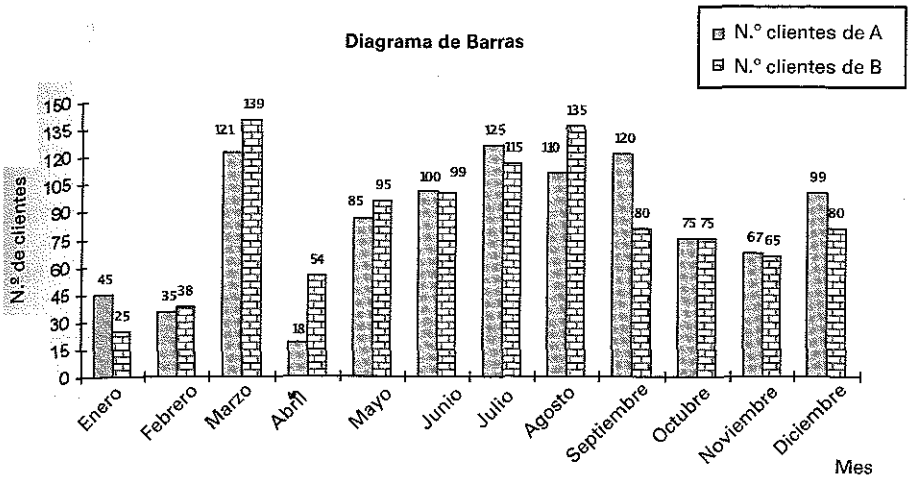


Gráfico 2.2. Ejemplo de diagrama de barras comparativo de dos variables.

B) Histogramas

Los histogramas son representaciones similares a los gráficos de barras; son, de hecho, un tipo especial de gráfico de barras que suelen utilizarse para ver los datos agrupados; en este caso la representación es un conjunto de rectángulos donde cada uno representa una clase; la base de los rectángulos sería igual a la amplitud del intervalo y la altura se determinaría de forma que el área del rectángulo sea proporcional a la frecuencia de cada clase.

Ejemplo 2.8. Representar un histograma con los siguientes valores y frecuencias:

Valores	Frecuencia
500 a 1.000	2
1.000 a 3000	10
3.000 a 4000	5
4.000 a 5000	4

El gráfico resultante sería el siguiente:

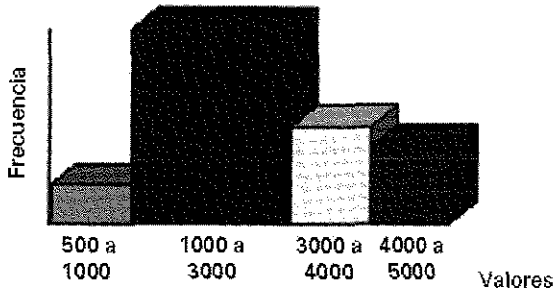


Gráfico 2.3. Ejemplo de histograma.

Como puede verse, la columna del intervalo 1000 a 3000 tiene el doble de anchura que la del resto de los intervalos.

C) Polígono de frecuencias

El polígono de frecuencias es una representación gráfica alternativa al histograma. Se forma uniendo los puntos que representan las frecuencias mediante segmentos, de tal manera que el punto con mayor altura representa la mayor frecuencia; el área bajo la curva representará al 100% de los datos.

Suelen utilizarse para representar tablas de frecuencia asociadas a distribuciones de datos cuantitativos de tipo II, si bien pueden utilizarse con cualquier tipo de datos. En el caso de las distribuciones de tipo III, los puntos en el eje de abcisas corresponderán a las marcas de clase de los intervalos.

Ejemplo 2.9. Representar el polígono de frecuencias asociado al ejemplo anterior.

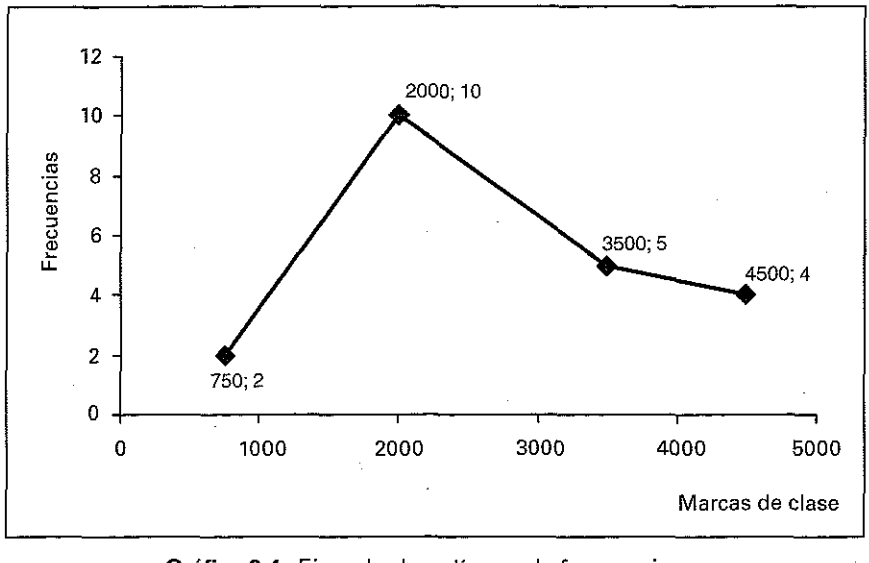


Gráfico 2.4. Ejemplo de polígono de frecuencias.

D) Gráficos de sectores

Estos gráficos se utilizan para mostrar las contribuciones relativas de cada punto de los datos al total de la serie. En un gráfico de sectores sólo se representa una serie.

En los gráficos de sectores el área de cada sector es proporcional a las frecuencias absolutas n_i de cada modalidad.

Ejemplo 2.10. Durante un determinado período la distribución de clientes de una empresa según su nacionalidad fue la siguiente:

Españoles	250
Franceses	150
Italianos	90
Otros	10

Representarlos gráficamente.

Para este tipo de representaciones son muy adecuados los gráficos por sectores; estos gráficos ofrecen gran variedad (quesos, círculos, anillos, etc.); cómo puede verse en el primer gráfico, a veces puede ser aconsejable destacar levemente los sectores para mejorar la visión del gráfico.

Cualquier programa de gráficos, como la propia hoja de cálculo Excel, permite generar con bastante facilidad este tipo de gráficos; los datos que facilitan su interpretación pueden indicarse en cifras absolutas (primer gráfico) o en términos porcentuales (segundo gráfico).

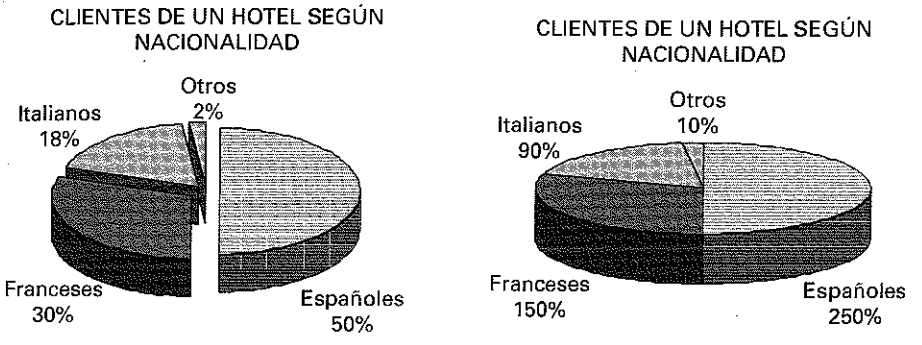


Gráfico 2.5. Ejemplos de gráfico de sectores.

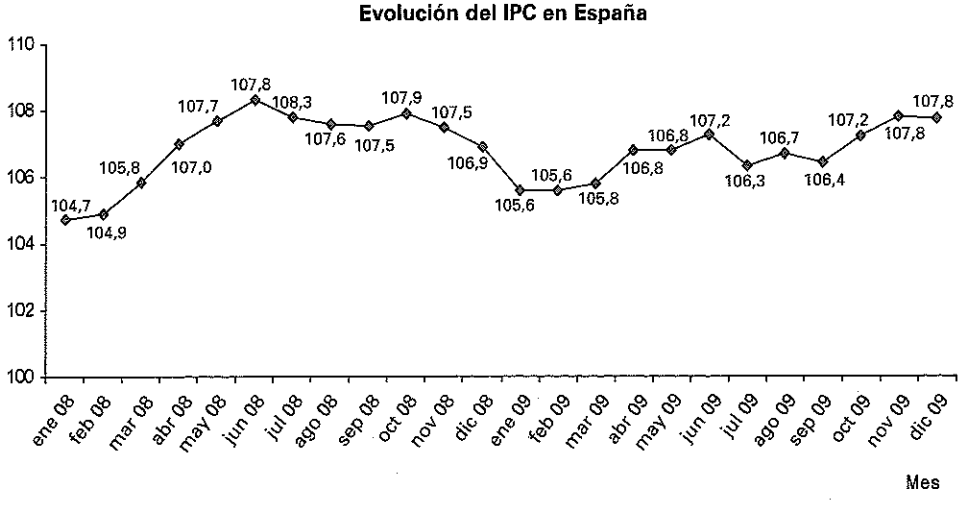
E) Gráficos de series temporales

Una secuencia de valores a intervalos regulares de tiempo constituye una serie temporal. En los gráficos de series temporales se representan los valores ordenados según la secuencia temporal, la cual figura en el eje de abcisas, en tanto que los valores obtenidos se representan en el eje de ordenadas.

Ejemplo 2.11. Según la información aportada por el INE, la evolución del IPC en España (Base 2006 = 100) en distintos meses del año viene reflejada en la siguiente serie.

Meses	IPC	Meses	IPC
Ene-08	104,747	Ene-09	105,592
Feb-08	104,91	Feb-09	105,603
Mar-08	105,841	Mar-09	105,776
Abr-08	106,98	Abr-09	106,809
May-08	107,702	May-09	106,772
Jun-08	108,322	Jun-09	107,242
Jul-08	107,802	Jul-09	106,327
Ago-08	107,571	Ago-09	106,698
Sep-08	107,549	Sep-09	106,446
Oct-08	107,918	Oct-09	107,205
Nov-08	107,46	Nov-09	107,786
Dic-08	106,909	Dic-09	107,758

Elaborar un gráfico representativo de la serie.
Un gráfico que representa esta evolución es el siguiente:



FUENTE: Instituto Nacional de Estadística. .

Gráfico 2.6. Ejemplo de serie temporal.

f) Diagramas de Pareto

Este diagrama está basado en el principio de Pareto, aplicado a la mala distribución de la riqueza (muchos tienen poco y pocos tienen mucho) y que puede generalizarse indicando que en la mayor parte de las situaciones se presentan muchos hechos

o problemas sin importancia frente a unos pocos especialmente importantes; el objetivo de la representación gráfica es permitir visualizar de forma rápida los problemas o los datos más significativos, que quedan a la izquierda de la gráfica, mientras que a medida que se avanza hacia la derecha del gráfico se van situando los datos de menor interés.

Tiene gran utilidad en economía, ya que este principio paretiano se ajusta a bastantes situaciones reales (pocos clientes suelen acumular muchas ventas mientras un gran número de pequeños clientes suman un pequeño porcentaje del total, pocas empresas contratan a muchos trabajadores mientras la mayoría sólo tiene pocos empleados, algunos errores del control de calidad de un producto o servicio son los más habituales mientras otros muchos se producen sólo de vez en cuando, pero cuando se producen son muy importantes, etc.).

Su construcción se realiza en dos pasos:

— Se ordenan las clases o categorías según la frecuencia relativa de su aparición.
— Cada clase se representa por un rectángulo con una altura igual a la frecuencia relativa.

El diagrama de Pareto representa los valores de las variables en el eje de abscisas y las frecuencias absolutas y relativas acumuladas en el eje de ordenadas.

Ejemplo 2.12. Construir un diagrama de Pareto ajustado a la siguiente distribución de frecuencias correspondiente a las ventas de una empresa a 100 clientes:

Ventas anuales en euros	N.º de clientes
Menos de 1.000 €	3
De 1.000 a 2.000 €	10
De 2.000 a 3.000 €	7
De 3.000 a 4.000 €	8
De 4.000 a 5.000 €	4
De 6.000 a 7.000 €	21
De 7.000 a 8.000 €	9
De 8.000 a 9.000 €	32
De 9.000 a 10.000 €	2
De 10.000 a 20.000 €	3
Más de 20.000 €	1
Total	100

Procedemos a la ordenación por clases o categorías (de mayor a menor en la importancia de sus frecuencias) y a obtener las frecuencias relativas y las frecuencias relativas acumuladas obteniendo la siguiente tabla:

Ventas en miles de euros	Frecuencia	Frecuencia %	% acumulado
De 8.000 a 9.000 €	32	32,00%	32,00%
De 6.000 a 7.000 €	21	21,00%	53,00%
De 1.000 a 2.000 €	10	10,00%	63,00%
De 7.000 a 8.000 €	9	9,00%	72,00%
De 3.000 a 4.000 €	8	8,00%	80,00%
De 2.000 a 3.000 €	7	7,00%	87,00%
De 4.000 a 5.000 €	4	4,00%	91,00%
Menos de 1.000 €	3	3,00%	94,00%
De 10.000 a 20.000 €	3	3,00%	97,00%
De 9.000 a 10.000 €	2	2,00%	99,00%
Más de 20.000 €	1	1,00%	100,00%
Total	100		

A partir de esta información construimos el siguiente gráfico o diagrama de Pareto, que nos indica, con un golpe de vista que más de la mitad de los clientes compran entre 6.000 y 9.000 euros anuales:

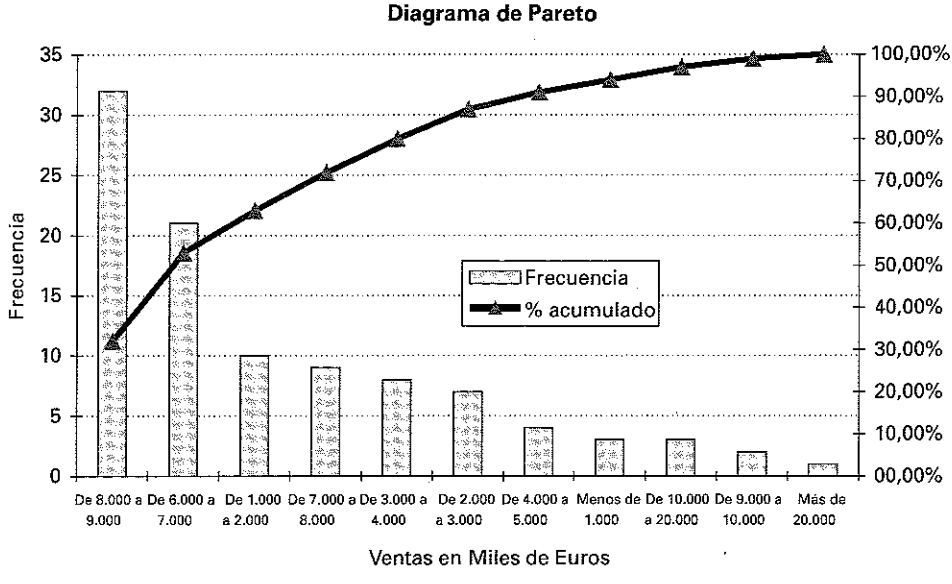


Gráfico 2.7. Ejemplo de diagrama de Pareto.

G) Diagramas de Tallos y Hojas

Haremos mención, finalmente, a la técnica recuento y ordenación de datos denominada **Diagramas de tallos y hojas**; esta técnica, similar a los histogramas, es bastante fácil y, según los casos, aporta más información que los propios histogramas. Su utilización será adecuada si el número de datos a representar no es muy elevado.

Ejemplo 2.13. Representar la siguiente distribución de frecuencias en un diagrama de tallos y hojas

33 24 36 29 39 20 36 45 31 31 39 24 29 23 41 40 33 24 34 41

Comenzamos seleccionando los **tallos** que en nuestro caso son las cifras de decenas, es decir 3, 2, 4, que reordenadas son 2, 3 y 4.

A continuación efectuamos un recuento y vamos «añadiendo» cada **hoja** a su **tallo.**

Tallos	Hojas
2	4 9 0 4 9 3 4
3	3 6 9 6 1 1 9 3 4
4	5 1 0 1

Por último reordenamos las **hojas** y hemos terminado el diagrama

Tallos	Hojas
2	0 3 4 4 4 9 9
3	1 1 3 3 4 6 6 9 9
4	0 1 1 5

H) Otras representaciones gráficas

Dado que el verdadero interés de los gráficos es describir la información de los datos, la naturaleza de las variables nos puede sugerir otras representaciones distintas a las anteriores.

A continuación se han seleccionado dos ejemplos de pictogramas; en el segundo se ha optado por aumentar el tamaño de las figuras haciéndolo proporcional a las frecuencias de cada modalidad, mientras que en el primero se opta por aumentar el número de objetos representados.

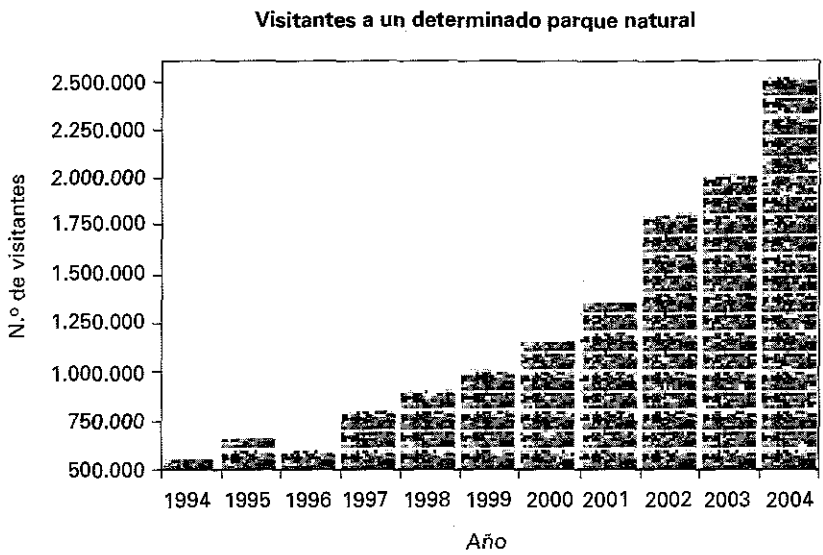


Gráfico 2.8. Ejemplo de pictograma.

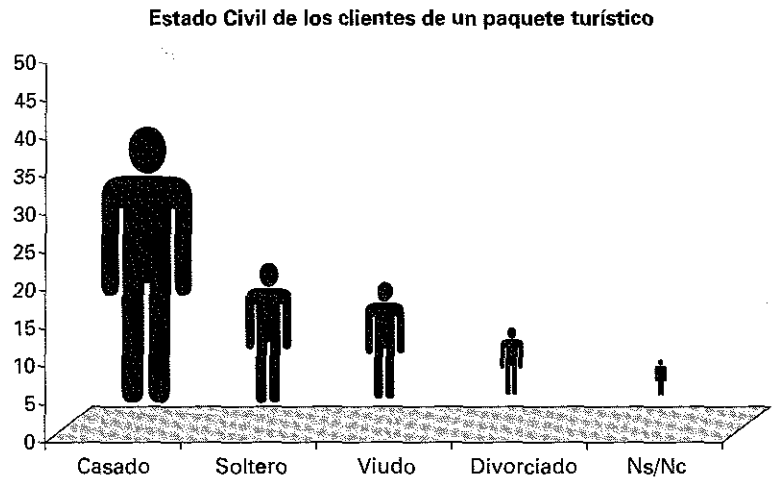


Gráfico 2.9. Ejemplo de pictograma.

La mayor parte de las distribuciones permiten diversas representaciones alternativas.

Ejemplo 2.14. Veamos un ejemplo de representación gráfica sobre datos referidos a una distribución de los clientes de un hotel según su nacionalidad.

Veamos un ejemplo de representación gráfica sobre datos referidos a una distribución de los clientes de un hotel según su nacionalidad.

Datos obtenidos:

Clientes de un hotel por nacionalidad	
Franceses	350
Belgas	150
Suecos	420
Noruegos	185
Ingleses	215
Otros	145

Algunas de las posibles representaciones gráficas que pueden establecerse son:

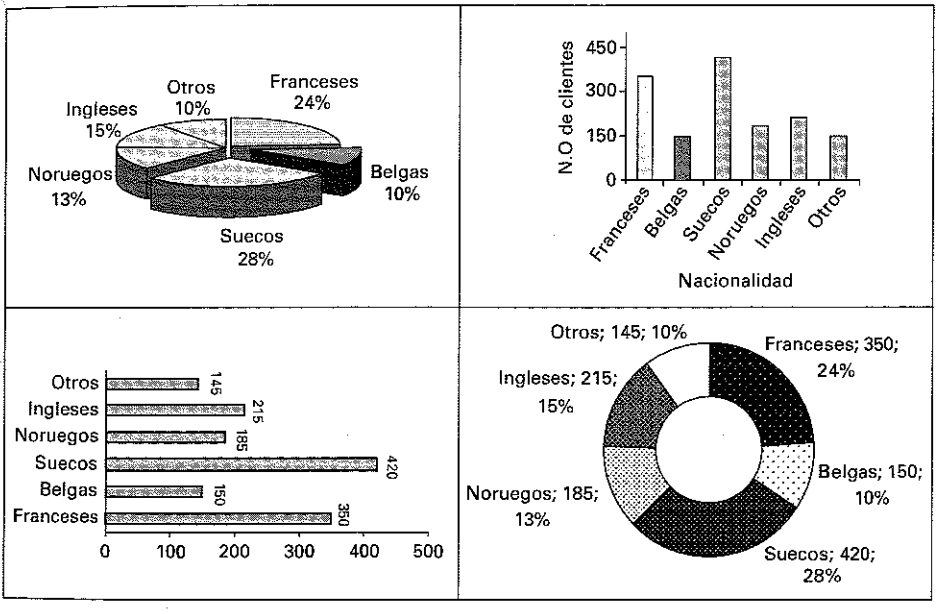


Figura 2.1. Distintas representaciones gráficas sobre unos mismos datos: nacionalidad de los clientes del hotel.

2.7. GENERACIÓN DE GRÁFICOS CON LA HOJA DE CÁLCULO EXCEL Y SPSS

2.7.1. La generación de gráficos con Excel

Para generar gráficos pueden utilizarse múltiples aplicaciones informáticas (SPSS, TeeChart Office, StatGraphics, etc...); en este epígrafe daremos algunas instrucciones básicas para trabajar con la hoja de cálculo Excel, que permite crear una gran variedad de gráficos. Existen varios gráficos de tipo estándar (barras, columnas, líneas, circular, anillos, radial, dispersión, etc..) y otros personalizados (apilado, áreas en tres dimensiones, circular, etc..), cada uno de ellos con diversas variaciones.

Los cuatro pasos que se requieren para realizar una representación gráfica en Excel son:

1. Seleccionar el tipo de gráfico que van a ser representado. Esta selección incluye también la posibilidad de seleccionar subtipos de gráficos, tal como se indica en la Figura 2.1. Para dar acceso a este cuadro de diálogo ha de hacerse click en el siguiente botón de comandos, o bién seleccionar el menú *Insertar –> Gráfico*.

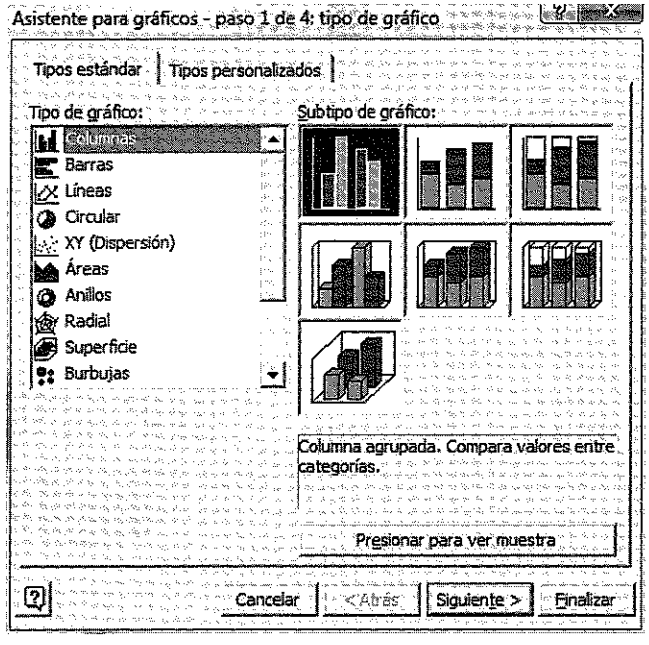


Figura 2.2. Asistente para Gráficos de Excel. Paso 1 de 4.

2. Selección de los datos de origen. Está opción, tal como se indica en la Figura 2.2, permite colocar los datos en filas o en columnas o incluir varias series de datos en el mismo gráfico.

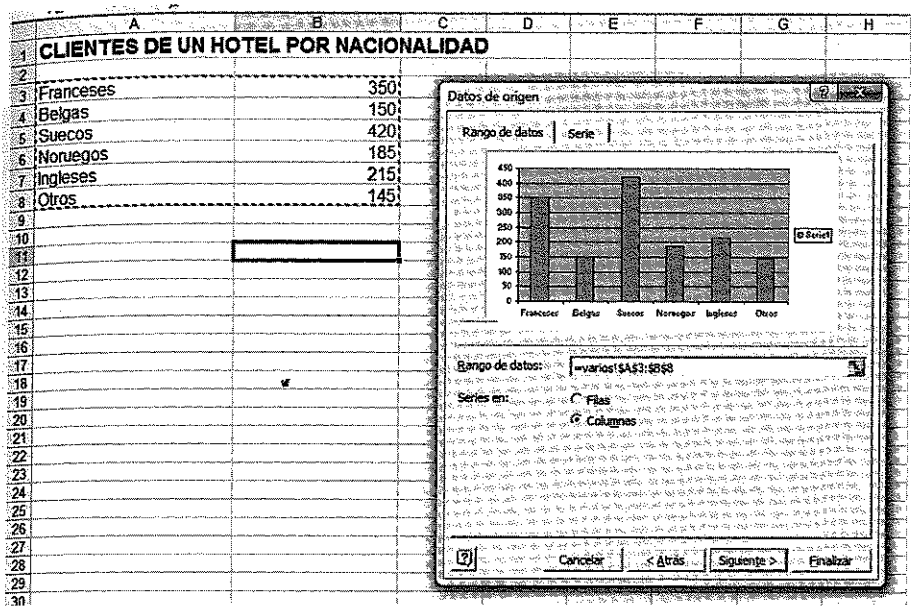


Figura 2.3. Asistente para Gráficos de Excel. Paso 2 de 4.

3. En el tercer paso se agregan las leyendas, el título del gráfico, los títulos de los ejes, las opciones de rótulos de datos o se especifica si adjuntar o no la tabla de datos, tal como puede verse en la Figura 2.3.

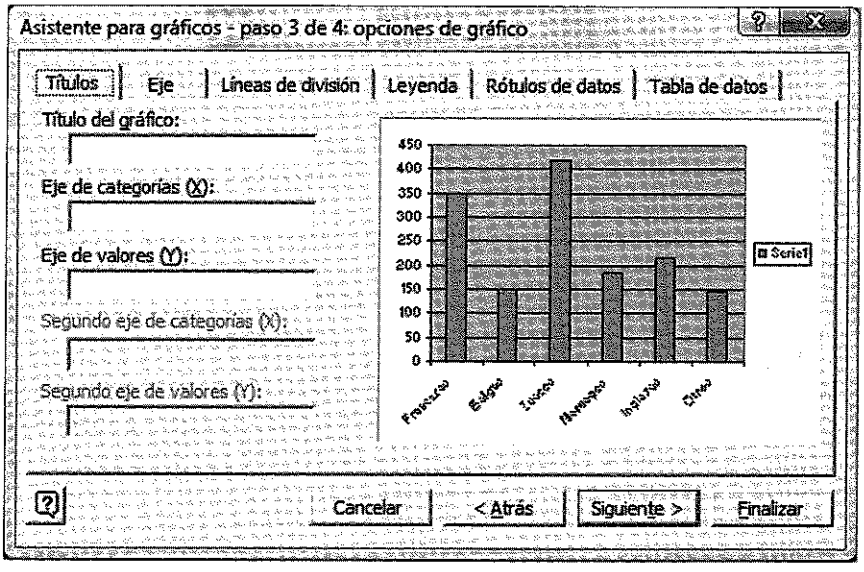


Figura 2.4. Asistente para Gráficos de Excel. Paso 3 de 4.

4. Finalmente el último paso nos pide que le indiquemos el lugar en el que Excel construirá el Gráfico, pudiendo elegir entre una casilla de la hoja de cálculo en la que trabajamos o como una hoja nueva que Excel generará automáticamente.

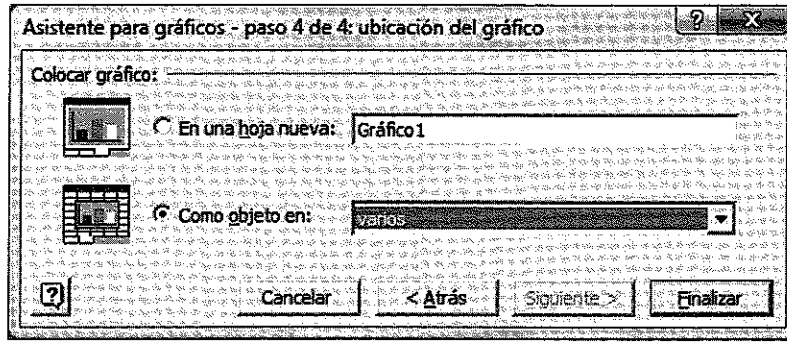


Figura 2.5. Asistente para Gráficos de Excel. Paso 4 de 4.

2.7.2. La generación de gráficos con SPSS

El programa estadístico SPSS es un sistema global para el análisis de datos. SPSS puede adquirir datos de casi cualquier tipo de archivo y utilizarlos para generar informes tabulares, gráficos y diagramas de distribuciones, múltiples estadísticos descriptivos y análisis estadísticos complejos.

En particular, y en lo referente a este apartado permite generar gráficos de alta resolución, que incluyen gráficos de sectores, gráficos de barras, histogramas, diagramas de dispersión y gráficos 3-D de alta resolución y a todo color, entre muchos otros.

Una de las ventajas de SPSS en su función de editor de gráficos es que permite modificar los gráficos y diagramas en la propia ventana a la que incorpora los gráficos generados; es posible cambiar los colores, seleccionar diferentes tipos de fuentes y tamaños, intercambiar los ejes horizontal y vertical, rotar diagramas de dispersión 3-D e incluso cambiar el tipo de gráfico.

La operativa del programa es muy simple y la iremos desarrollando a lo largo de los siguientes capítulos:

El alumno debe consultar un manual de SPSS para trabajar adecuadamente con el programa; veamos, no obstante, aquí algunas nociones básicas:

Existen diversos tipos de ventanas en SPSS:

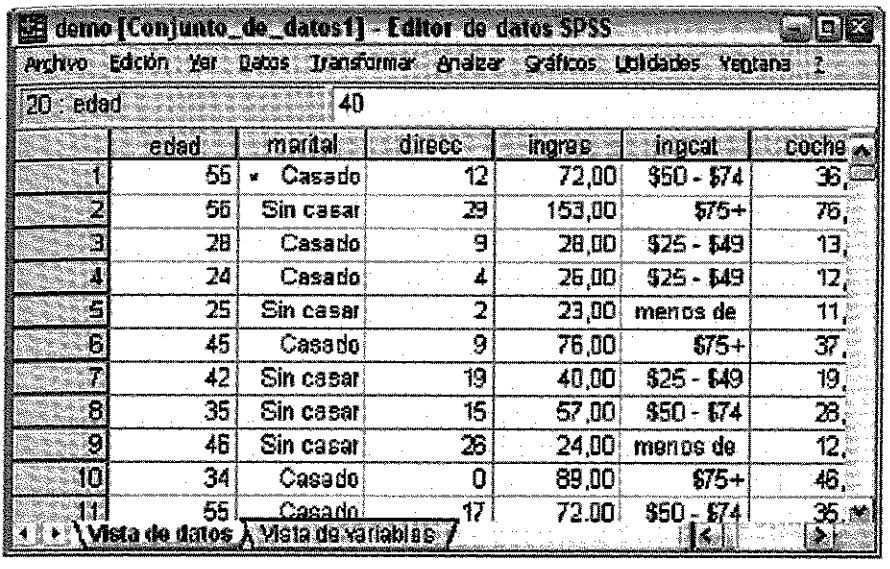
Editor de datos. El Editor de datos muestra el contenido del archivo de datos. Puede crear nuevos archivos de datos o modificar los existentes con el Editor de datos. Si tiene más de un archivo de datos abierto, habrá una ventana Editor de datos independiente para cada archivo.

El Editor de datos proporciona dos vistas de los datos.

— **Vista de datos.** Esta vista muestra los valores de datos reales o las etiquetas de valor definidas.

— **Vista de variables.** Esta vista muestra la información de definición de las variables que incluye las etiquetas de la variable definida y de valor, tipo de dato (por ejemplo, cadena, fecha y numérico), nivel de medida (nominal, ordinal o de escala) y los valores perdidos definidos por el usuario.

En ambas vistas, se puede añadir, modificar y eliminar la información contenida en el archivo de datos.


demo [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

20 : edad 40

	edad	marital	direcc	ingres	ingcat	coche
1	55	Casado	12	72,00	$50 - $74	36,
2	56	Sin casar	29	153,00	$75+	76,
3	28	Casado	9	28,00	$25 - $49	13,
4	24	Casado	4	26,00	$25 - $49	12,
5	25	Sin casar	2	23,00	menos de	11,
6	45	Casado	9	76,00	$75+	37,
7	42	Sin casar	19	40,00	$25 - $49	19,
8	35	Sin casar	15	57,00	$50 - $74	28,
9	46	Sin casar	26	24,00	menos de	12,
10	34	Casado	0	89,00	$75+	46,
11	55	Casado	17	72,00	$50 - $74	35,

Vista de datos Vista de variables

Figura 2.6. Editor de datos de SPSS.

Muchas de las funciones de la Vista de datos son similares a las que se encuentran en aplicaciones de hojas de cálculo. Sin embargo, existen varias diferencias importantes: Las filas son casos. Cada fila representa un caso o una observación. Por ejemplo, cada individuo que responde a un cuestionario es un caso.

— Las columnas son variables. Cada columna representa una variable o una característica que se mide.
— Las casillas contienen valores. Cada casilla contiene un valor único de una variable para cada caso.

Puede especificarse el nivel de medida como Escala (datos numéricos de una escala de intervalo o de razón), Ordinal o Nominal. Los datos nominales y ordinales pueden ser de cadena (alfanuméricos) o numéricos.

— **Nominal.** Una variable puede ser tratada como nominal cuando sus valores representan categorías que no obedecen a una ordenación intrínseca. Por ejemplo, el departamento de la compañía en el que trabaja un empleado. Son ejemplos de variables nominales: la región, el código postal o la confesión religiosa.

— **Ordinal.** Una variable puede ser tratada como ordinal cuando sus valores representan categorías con alguna ordenación intrínseca. Por ejemplo los niveles de satisfacción con un servicio, que vayan desde muy insatisfecho hasta muy satisfecho. Son ejemplos de variables ordinales: las valoraciones del nivel de satisfacción, confianza o preferencia.
— **Escala.** Una variable puede ser tratada como de escala cuando sus valores representan categorías ordenadas con una métrica con significado, por lo que son adecuadas las comparaciones de distancia entre valores. Son ejemplos de variables de escala: la edad en años y los ingresos en dólares.

Visor. Todas las tablas, los gráficos y los resultados estadísticos se muestran en el Visor. Puede editar los resultados y guardarlos para utilizarlos posteriormente. La ventana del Visor se abre automáticamente la primera vez que se ejecuta un procedimiento que genera resultados.

Visor de borrador. Los resultados pueden mostrarse como texto simple (en lugar de como tablas pivote interactivas) en el Visor de borrador.

Editor de tablas pivote. Con el Editor de tablas pivote es posible modificar los resultados mostrados en este tipo de tablas de diversas maneras. Puede editar el texto, intercambiar los datos de las filas y las columnas, añadir colores, crear tablas multidimensionales y ocultar y mostrar los resultados de manera selectiva.

Editor de gráficos. Puede modificar los gráficos y diagramas de alta resolución en las ventanas de los gráficos. Es posible cambiar los colores, seleccionar diferentes tipos de fuentes y tamaños, intercambiar los ejes horizontal y vertical, rotar diagramas de dispersión 3-D e incluso cambiar el tipo de gráfico.

Editor de resultados de texto. Los resultados de texto que no aparecen en las tablas pivote pueden modificarse con el Editor de resultados de texto. Puede editar los resultados y cambiar las características de las fuentes (tipo, estilo, color y tamaño).

Editor de sintaxis. Puede pegar las selecciones del cuadro de diálogo en una ventana de sintaxis, donde aparecerán en forma de sintaxis de comandos. A continuación puede editar esta sintaxis para utilizar las funciones especiales de SPSS que no se encuentran disponibles en los cuadros de diálogo. También puede guardar los comandos en un archivo para utilizarlos en sesiones de SPSS posteriores.

Veamos en este capítulo el uso del **Editor de Gráficos:**

Antes de crear un gráfico es necesario tener los datos en el Editor de datos. Es posible introducir los datos directamente en el Editor de datos; abrir un archivo de datos previamente guardado o leer una hoja de cálculo, un archivo de datos de texto delimitado por tabuladores o un archivo de base de datos. Si selecciona Tutorial en el menú Ayuda podrá ver ejemplos en pantalla de creación y modificación de gráficos. Además, el sistema de ayuda en pantalla incluye información sobre cómo crear y modificar cualquier tipo de gráfico.

Una vez que los datos se encuentran en el Editor de datos, seleccione Generador de gráficos en el menú Gráficos.

Se abrirá el cuadro de diálogo Generador de gráficos.

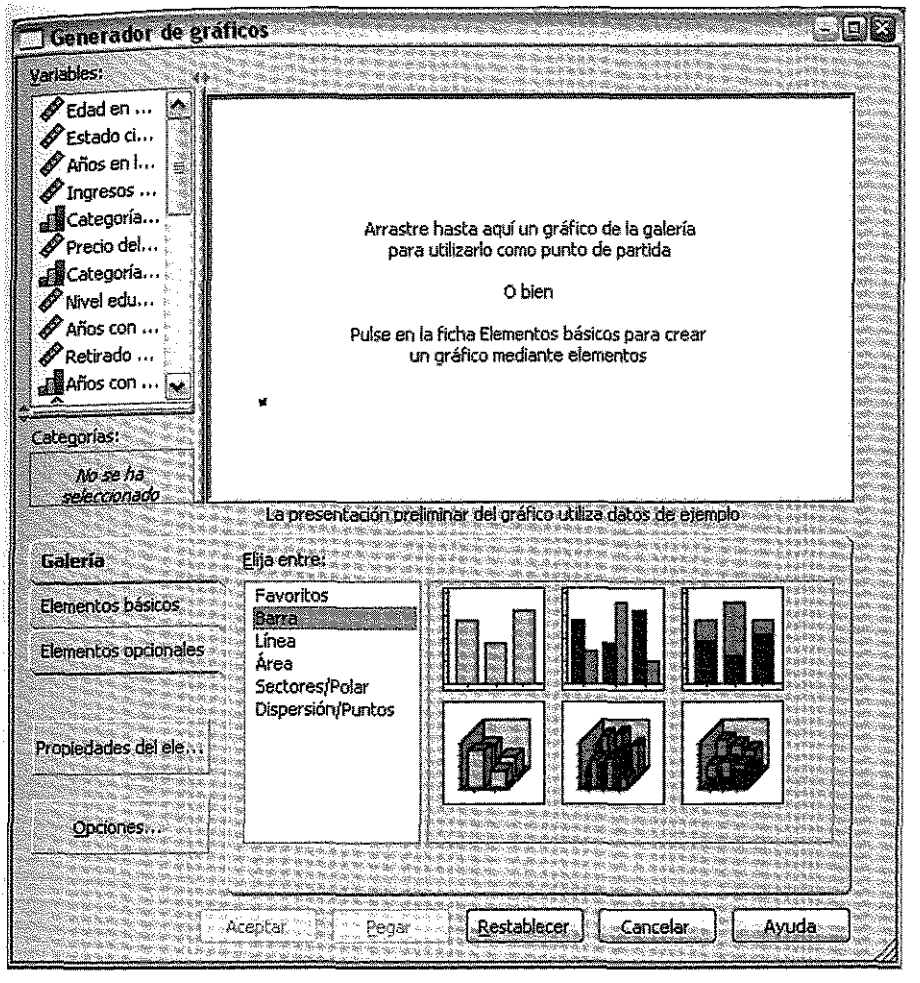


Figura 2.7. Generador de Gráficos de SPSS. Paso 1/2.

El cuadro de diálogo Generador de gráficos se utiliza para los tipos de gráficos más habituales, que aparecen en la pestaña Galería. Si debe crear un gráfico que no está disponible en el generador de gráficos, también puede seleccionar un tipo de gráfico específico en el menú Gráficos.

En el generador de gráficos, arrastre el icono correspondiente al gráfico al «lienzo», que es la zona grande que hay encima de la galería.

Arrastre las variables desde la lista Variables a las zonas de arrastre del eje. (Si desea obtener más información acerca del generador de gráficos, pulse en Ayuda.) Cuando se haya terminado de definir el gráfico, el resultado será similar al siguiente gráfico.

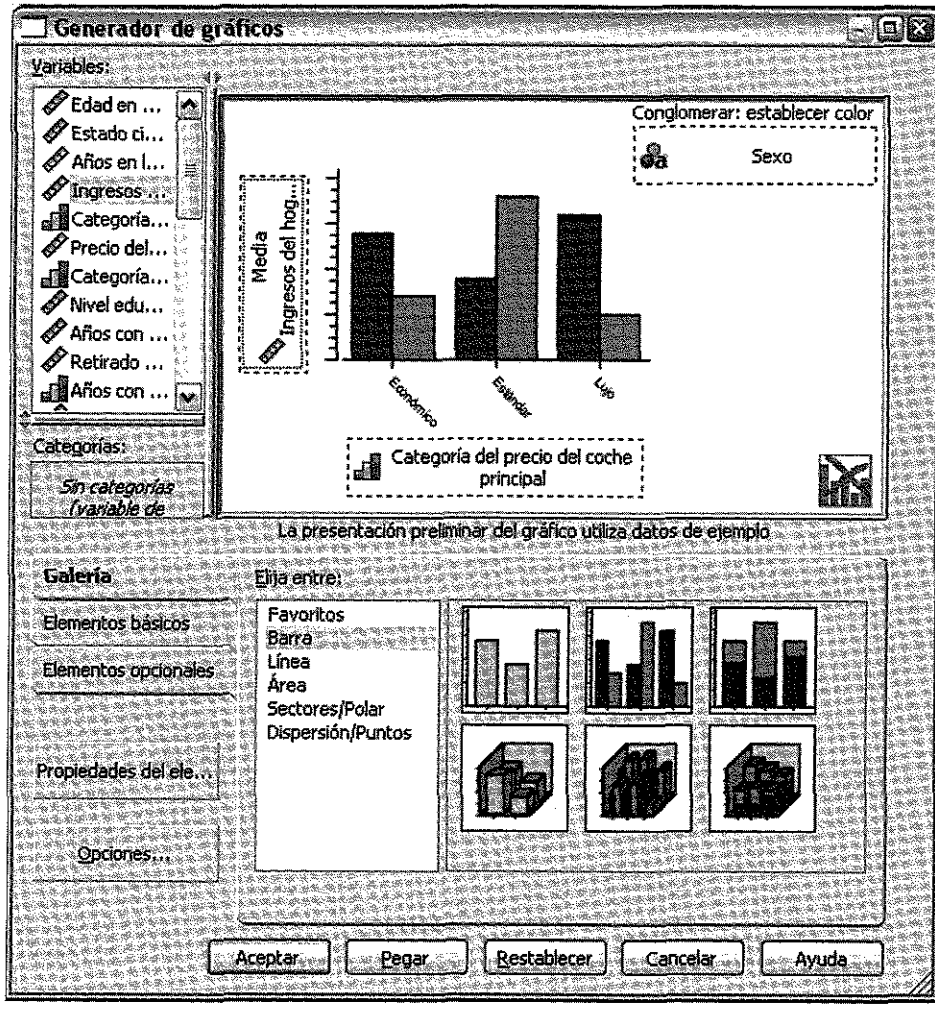


Figura 2.8. Generador de Gráficos de SPSS. Paso 2/2.

Si necesita cambiar los estadísticos o modificar los atributos de los ejes o las leyendas, pulse en Propiedades del elemento.

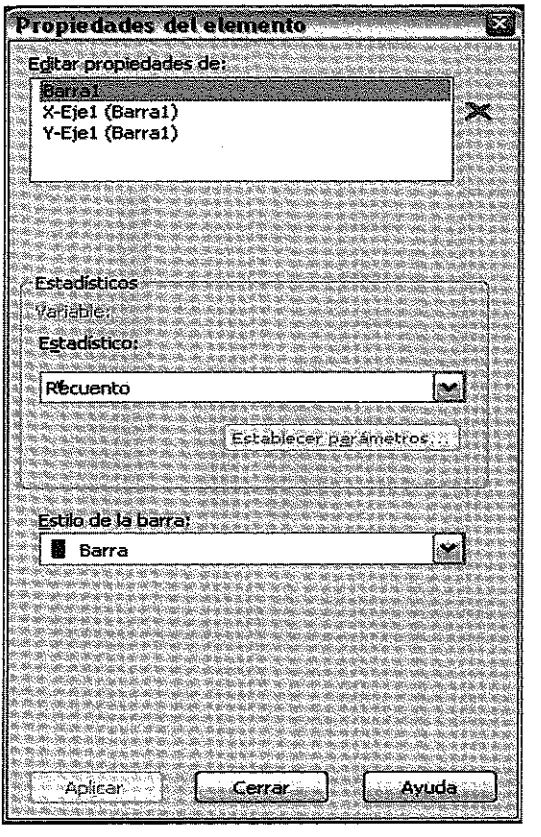
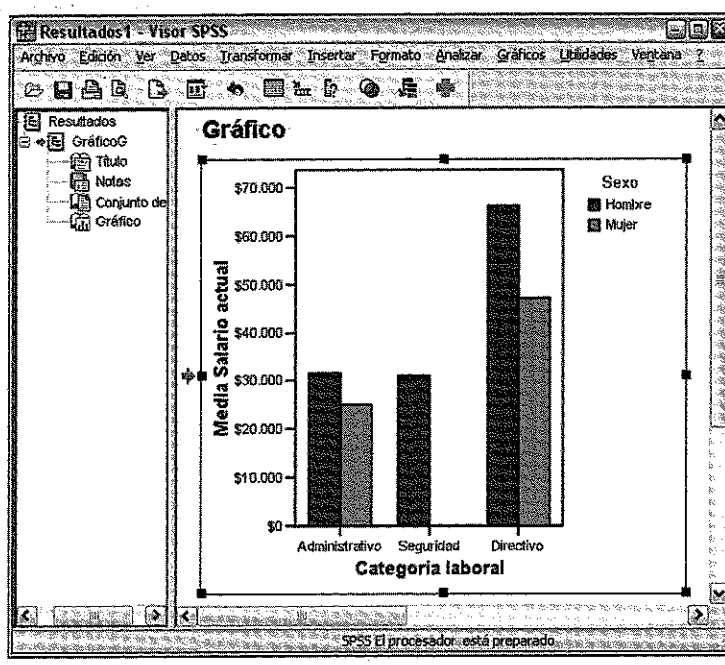


Figura 2.9. Selección de Propiedades de los Gráficos de SPSS.

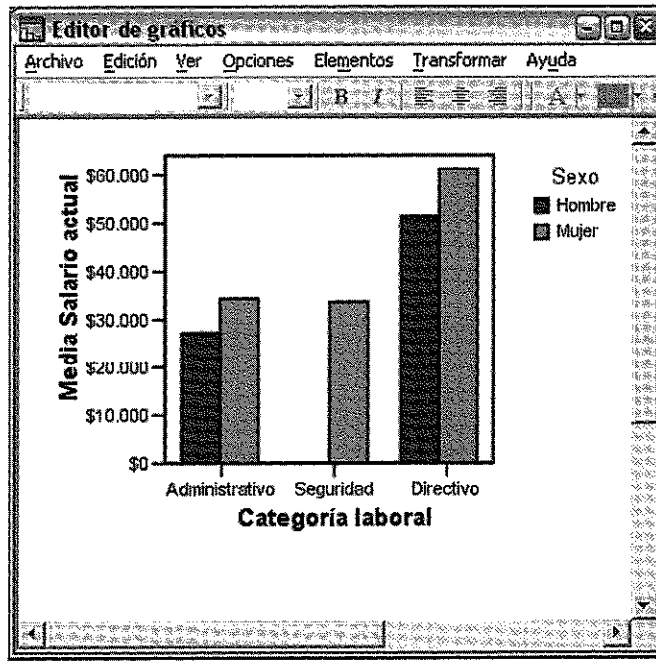
En la lista Editar propiedades de, seleccione el elemento que desea cambiar. (Si desea obtener información acerca de propiedades específicas, pulse en Ayuda.)

Pulse en Aceptar en el cuadro de diálogo Generador de gráficos para crear el gráfico.

Aparecerá el gráfico en el Visor.
Ejemplo:



Para modificar un gráfico, pulse dos veces en cualquier parte del gráfico que aparece en el Visor. Al hacer esto, aparecerá el gráfico en el Editor de gráficos.



Puede modificar cualquier parte del gráfico o cambiar a otro tipo de gráfico que represente los mismos datos. Puede añadir elementos, así como ocultarlos o mostrarlos utilizando los menús del Editor de gráficos.

Para modificar un elemento de un gráfico

Seleccione el elemento que desea modificar.
Elija en los menús:

Edición
Propiedades...

Se abrirá la ventana Propiedades. Las pestañas que aparecen en la ventana Propiedades dependen de la selección realizada. La ayuda en pantalla explica cómo hacer que aparezcan las pestañas que necesita.

Tipos de gráficos

SPSS puede generar múltiples tipos de gráficos (gráfico de áreas, barras, cajas, tallos y hojas, histogramas, gráficos de líneas, puntos y bandas, sectores, diagramas de dispersión, etc.; para un mejor conocimiento y aplicabilidad de la Estadística el alumno debe practicar con este programa la elaboración de gráficos explicada en este capítulo.

2.8. EJERCICIOS

Sobre distribuciones de frecuencias unidimensionales

Ejercicio 2.1. *Una empresa tiene un total de 20 empleados que trabajan en cuatro oficinas. Las oficinas están numeradas del 1 al 4. Los registros de la empresa indican las oficinas en las que se ubican los 20 empleados listados en orden alfabético en la siguiente forma:*

Empleados	Número de Oficina	Empleados	Número de Oficina
1	1	11	4
2	4	12	4
3	1	13	2
4	3	14	2
5	3	15	1
6	2	16	1
7	1	17	2
8	1	18	4
9	1	19	4
10	3	20	1

Sintetice los datos y muestre, en una tabla, las frecuencias asociadas con los valores 1, 2, 3 y 4.

Respuesta

Procedemos a contar las frecuencias con las que aparecen los empleados en cada una de las oficinas; por ejemplo, en la oficina 1 contaremos 8 empleados (los enumerados en las posiciones 1, 3, 7, 8, 9 15, 16 y 20).

Nos quedaría, para una ***Frecuencia total o total de datos:*** $N = 20$, la siguiente tabla:

Variable (x_i) (oficinas)	Frecuencia absoluta (n_i) (número de empleados)	Frecuencia acumulada (N_i)
1	8	8
2	4	(8+ 4 =) 12
3	3	(12+ 3 =) 15
4	5	(15+ 5 =) 20

Ejercicio 2.2. *Convierta las frecuencias del Ejercicio 2.1 en frecuencias relativas y en frecuencias relativas acumuladas y en porcentaje acumulado.*

Respuesta

En el Ejercicio 2.1 el número total de observaciones fue 20; para obtener las frecuencias relativas deben dividirse las frecuencias absolutas por el número total de observaciones.

La información demandada sería:

Variable (X_i)	1	2	3	4
Frecuencia absoluta (n_i)	8	4	3	5
Frecuencia relativa $f_i = \frac{n_i}{N}$	$\frac{8}{20} = 0{,}4$	$\frac{4}{20} = 0{,}2$	$\frac{3}{20} = 0{,}15$	$\frac{5}{20} = 0{,}25$
Frecuencia relativa acumulada	0,40	0,60	0,75	1
Porcentaje acumulado	40%	60%	75%	100%

La Frecuencia relativa también podemos expresarla en porcentajes: en la oficina 1 trabajan un 40% de los empleados, en la 2 un 20%, en la 3 un 15% y en la 4 el 25% restante.

La Frecuencia relativa acumulada indica que en las 2 primeras oficinas trabajan conjuntamente un 60% de empleados y entre las 3 primeras oficinas agrupan un 75%.

Ejercicio 2.3. *Suponga que en cierto mes los 20 empleados del ejercicio anterior reciben, en euros, los siguientes salarios:*

850	1.265	**895**	575	2.410	470	660	1.820	1.510	**1.100**
620	425	750	**965**	840	1.505	1.375	695	1.125	1.475

Utilice estos datos para construir una tabla de frecuencias con intervalos de clase adecuados.

Respuesta

Con solo 20 observaciones no debemos formar excesivas clases o intervalos; construyamos, por ejemplo, 5 intervalos del mismo tamaño. El ingreso varía entre 425 y 2.410 euros, una elección adecuada de clases, en este caso, podría consistir en dividir los ingresos en grupos de unos 500 euros de amplitud, en la siguiente forma:

Intervalo de Clase $(L_{i-1} - L_i)$	Frecuencia absoluta (n_i)	Valores incluidos en el intervalo
(250 - 750]	7	575, 470, 660, 620, 425, 750 y 695
(750 - 1.250)	6	850, 895, 1.100, 965, 840 y 1.125
[1.250 - 1.750)	5	1.265, 1.850, 1.510, 1.375 y 1.475
[1.750 - 2.250)	1	1.820
[2.250 - 2.750)	1	2.410

Una norma práctica es ir tachando los valores de la distribución a medida que vamos incluyéndolos en un intervalo; de esta forma comprobáremos que los hemos incluido todos y evitaremos errores.

Nótese que:

— Cada observación cae exactamente dentro y sólo dentro de una clase o intervalo; a tal fin hemos construido los intervalos abiertos por la izquierda (extremo inferior) y cerrados por la derecha (extremo superior), de forma que, por ejemplo, el empleado que gana 750 ⇔al mes ha quedado incluido en el primer intervalo (250 - 750].

— No hay espacios entre los intervalos, de forma que todo el rango o recorrido de los datos está incluido en los extremos de las clases.

— Todas las clases tienen la misma amplitud (500 euros).

— Las marcas de clase de los intervalos $x_i = \frac{L_{i-1} + L_i}{2}$ serían, redondeando: 500, 1000, 1500, 2000 y 2500.

Con esta información operamos para construir la siguiente tabla:

Intervalo de Clase ($L_{i-1} - L_i$)	Frecuencia absoluta (n_i)	Frecuencia acumulada (N_i)	Frecuencia relativa (f_i)	Frecuencia relativa acumulada (F_i)
(250 - 750)	7	7	(7/20 · 100 =) 35%	35%
[750 - 1.250)	6	13	(6/20 · 100 =) 30%	65%
[1.250 - 1.750)	5	18	(5/20 · 100 =) 25%	90%
[1.750 - 2.250)	1	19	(1/20 · 100 =) 5%	95%
[2.250 - 2.750)	1	20	(1/20 · 100 =) 5%	100%
Total	20		100%	

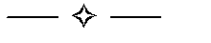
No hay una regla general que nos determine el número de intervalos o los límites que debemos emplear para sintetizar una información estadística en intervalos; tenemos que pensar fundamentalmente en el interés que puede aportar una síntesis de este tipo en los análisis que necesitemos realizar con los datos; en el ejemplo anterior parece que el número y los límites elegidos podrían ser adecuados para conocer la distribución de salarios en la empresa, pero según el detalle con el que nos interese la síntesis podríamos, por ejemplo, haber "unido" los dos últimos intervalos, de forma que, sí sólo tuviéramos acceso a la nueva información sintetizada, sólo sabríamos que un 10% de los empleados ganan entre 1.750 y 2.750 € al mes.

No obstante, una formulación que puede utilizarse para decidir el número de intervalos es la denominada ***Regla de Sturges***, que tiene la siguiente formulación:

$$\text{Número de clases o intervalos} = 1 + 3.332 \log_{10} N$$

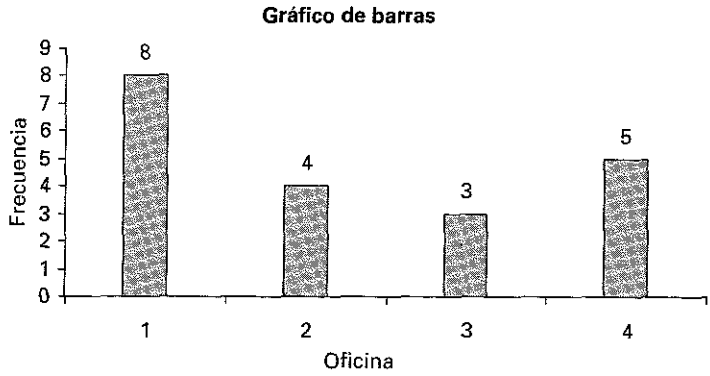
$$\text{Longitud del intervalo} = \text{Rango} / \text{Número de clases}$$

Con este criterio, en nuestro problema se obtendrían $1 + 3.332 \log_{10} 20 = 5{,}3 \Leftrightarrow 5$ intervalos; longitud del intervalo $= \dfrac{2.750 - 250}{5} = 500$.

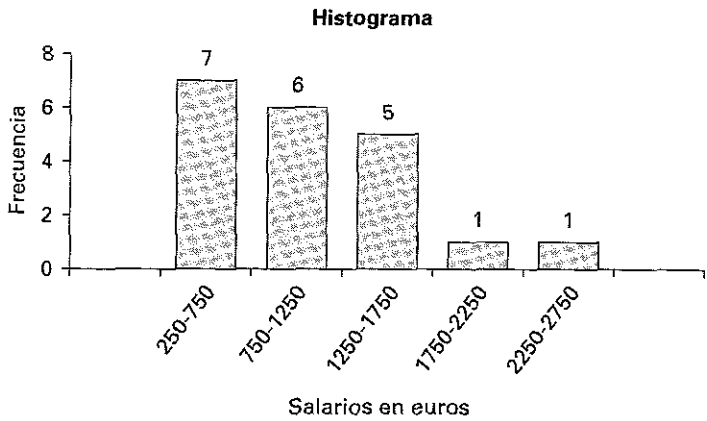


Ejercicio 2.4. *A continuación, realice una gráfica de barras con los datos del Ejercicio 2.1 y un histograma con los datos del Ejercicio 2.3.*

Respuesta

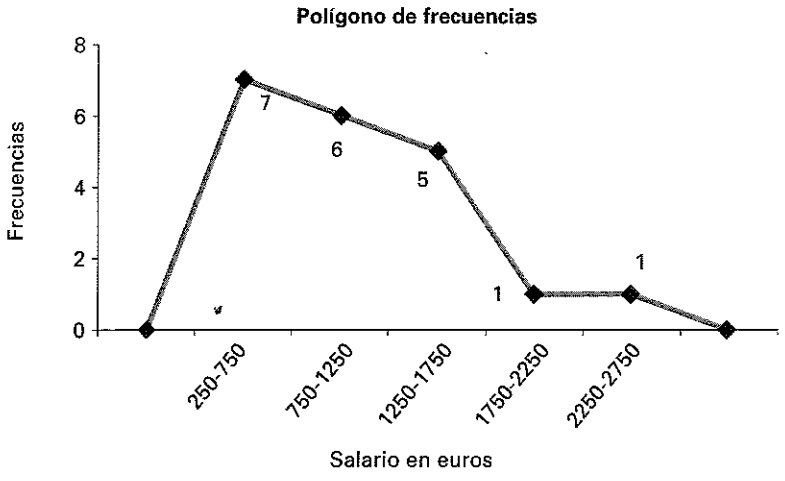


Dado que todos los intervalos tienen la misma amplitud, el histograma quedaría del siguiente modo:



Ejercicio 2.5. *Construya un polígono de frecuencias con los datos del Ejercicio 2.3.*

Respuesta



Ejercicio 2.6. *Suponga que las calificaciones de 40 aspirantes a un empleo de en una prueba de aptitud son las siguientes:*

42	21	46	69	87	29	34	59	81	97
64	60	87	81	69	77	75	47	73	82
91	74	70	65	86	87	67	69	49	57
55	68	74	66	81	90	75	82	37	94

Realice un diagrama de tallos y hojas.

Respuesta

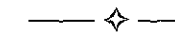
Si ordenamos los valores veremos que las puntuaciones varían entre 21 y 97. Los primeros dígitos, del 2 al 9, se colocan en una columna en la parte izquierda del diagrama, y los segundos dígitos se registran en la fila correspondiente al primer dígito. Por ejemplo, ubiquemos los puntos 21, 42 y 46.

2	1
3	
4	2 6

A continuación se puede apreciar el diagrama de tallos y hojas completo.

2	1	9							
3	4	7							
4	2	6	7	9					
5	9	7	5						
6	9	4	0	9	5	7	9	8	6
7	7	5	3	4	0	4	5		
8	7	1	7	1	2	6	7	1	2
9	7	1	0	4					

NOTA: La forma más práctica de construir el diagrama es ir tachando los valores de la tabla inicial, a medida de que se van incorporando al diagrama.



Ejercicio 2.7. *Una compañía de alquiler de automóviles registra el número de automóviles alquilados en cada turno de ocho horas. Los datos correspondientes a los últimos 28 turnos son:*

366	390	324	385	380	375	384
383	375	339	360	386	387	384
379	386	374	366	377	385	381
359	363	371	379	385	367	364

a) *Realice una tabla de frecuencias, frecuencias acumuladas, frecuencias relativas y frecuencias relativas acumuladas utilizando intervalos de clase adecuados.*
b) *Construya un histograma.*
c) *Construya un polígono de frecuencias.*
d) *Construya un diagrama de tallos y hojas.*

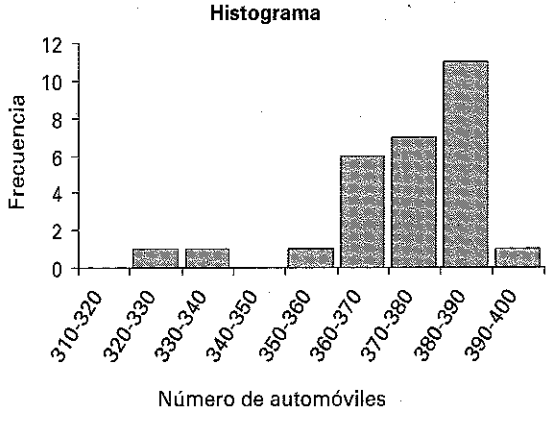
Respuesta

a) La variable oscila entre 324 y 390; dadas las características de los datos utilizaremos intervalos de clase cuya longitud sea 10; la tabla de frecuencias, frecuencias acumuladas, frecuencias relativas y frecuencias relativas acumuladas, quedaría:

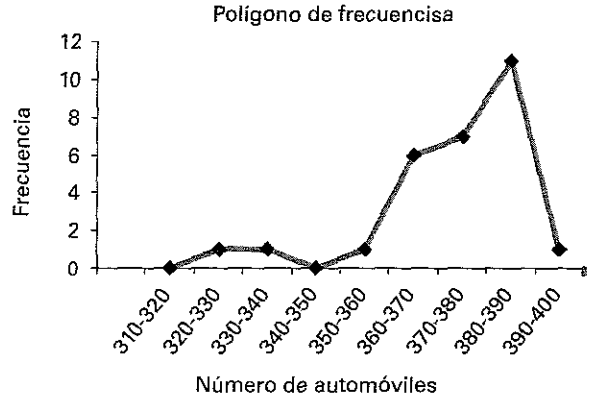
Intervalo de Clase $(L_{i-1} - L_i)$	Frecuencia absoluta (n_i)	Frecuencia acumulada (N_i)	Frecuencia relativa (f_i)	Frecuencia relativa acumulada (F_i)
$320 \leq x < 330$	1	1	1/28 = 0,036	0,036
$330 \leq x < 340$	1	2	0,036	0,072
$340 \leq x < 350$	0	2	0	0,072
$350 \leq x < 360$	1	3	0,036	0,108
$360 \leq x < 370$	6	9	0,214	0,322
$370 \leq x < 380$	7	16	0,250	0,572
$380 \leq x < 390$	11	27	0,392	0,964
$390 \leq x < 400$	1	28	0,036	1

Nótese que hemos elegido intervalos cerrados por el extremo inferior y abiertos por el extremo superior, del tipo [320 – 330), en el que está incluido el valor 320, pero no el 330, que se incluirá en el siguiente intervalo. Esta forma de señalar los límites de los intervalos es similar, aunque menos utilizada, que la indicada hasta ahora: [320 – 330), [330 – 340), etc.

b) Dado que todos los intervalos tienen la misma amplitud, el histograma será:



c) El polígono de frecuencias sería:



d) El diagrama de tallos y hojas, quedaría

32	4										
33	9										
34											
35	9										
36	6	0	6	3	7	4					
37	5	5	9	4	7	1	9				
38	5	0	4	3	6	7	4	6	5	1	5
39	0										

Ejercicio 2.8. *En una empresa se observaron en una muestra aleatoria, los siguientes valores de las ventas de cada uno de sus 16 empleados (datos de ventas un determinado día, en miles de euros):*

2 2,5 3 2 2,3 2,1 2,5 2
2 1,5 2 2 1,7 1,5 2,5 3

a) *Agrupe los datos en una tabla, indicando las frecuencias absolutas y las acumuladas, y calcule las frecuencias relativas y relativas acumuladas.*
b) *Represente la información en un gráfico de barras.*

Respuesta

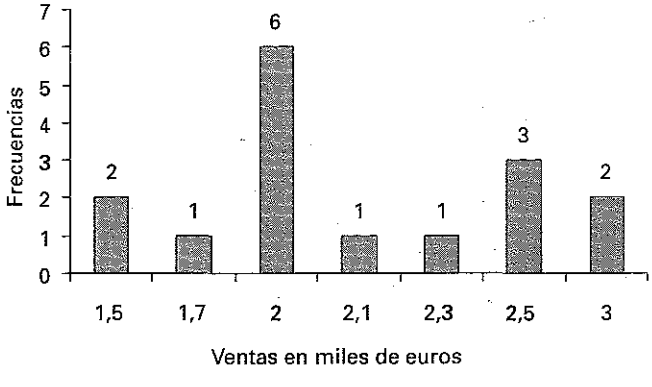
a) Lo mejor es comenzar ordenando los datos; de menor a mayor quedarían en la siguiente forma:

1,5 1,5 1,7 2 2 2 2 2
2 2,1 2,3 2,5 2,5 2,5 3 3

b) Obtenemos las frecuencias absolutas y acumuladas, las frecuencias relativas y las relativas acumuladas.

Valores (x_i)	Frecuencia absoluta (n_i)	Frecuencia acumulada (N_i)	Frecuencia relativa (f_i)	Frecuencia relativa (f_i) en porcentajes	Frecuencia relativa acumulada (F_i)
1,5	2	2	2/16	12,50%	2/16
1,7	1	3	1/16	6,25%	3/16
2	6	9	6/16	37,50%	9/16
2,1	1	10	1/16	6,25%	10/16
2,3	1	11	1/16	6,25%	11/16
2,5	3	14	3/16	18,75%	14/16
3	2	16	2/16	12,50%	1

c) La representación gráfica sería la siguiente:



Ejercicio 2.9. *Los siguientes valores corresponden a las ventas en miles de euros realizadas por una empresa sus últimas 15 operaciones:*

20 25 22 20 25 20 21 22 22 24 23 20 23 20 25

a) *Construya una tabla de frecuencias y frecuencias acumuladas.*
b) *Agrupe los datos en intervalos de amplitud dos.*
c) *Dibuje el histograma.*

Respuesta

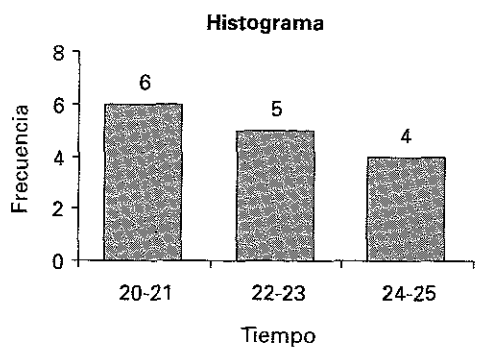
a) Tabla de frecuencias y de frecuencias acumuladas.
Después de ordenar los valores y contar las frecuencias obtenemos el siguiente resultado:

Valores de X (x_i)	Frecuencia (n_i)	Frecuencia acumulada (N_i)
20	5	5
21	1	6
22	3	9
23	2	11
24	1	12
25	3	15

b) Como la información disponible varía entre 20 y 25 y debemos agruparla en intervalos de amplitud 2 (2.000 €), tendremos:

Intervalo de Clase ($L_{i-1} - L_i$)	Frecuencia (n_i)	Frecuencia acumulada (N_i)
20-21	6	6
22-23	5	11
24-25	4	15

c) Histograma.
Dado que todos los intervalos tienen la misma amplitud, los rectángulos tienen también la misma anchura.



Ejercicio 2.10. *Se dispone de la siguiente distribución de frecuencias.*

52	50	38	42	48	55	52	51	36	47	52	53	57	56
35	36	58	49	46	50	49	52	38	41	55	48	59	49

a) *Sintetice los datos en una tabla, agrupando los datos en intervalos de amplitud 5.*

b) *Calcule la frecuencia relativa de cada intervalo.*

c) *Calcule las frecuencias acumuladas y acumuladas relativas.*

d) *Determine cuántos valores son inferiores y superiores a 50*

Respuesta:

a) Ordenados los datos de menor a mayor

35	36	36	38	38	41	42	46	47	48	48	49	49	49
50	50	51	52	52	52	52	53	55	55	56	57	58	59

Se observa que varían entre 35 y 59; nuestro primer intervalo podría ser el 35-39 y el último el 55-59; en este caso, la tabla de frecuencias quedaría:

Intervalo de Clase $(L_{i-1} - L_i)$	Frecuencia absoluta (n_i)
35-39	5
40-44	2
45-49	7
50-54	8
55-59	6

b) Las frecuencias relativas serían las siguientes

Intervalo de Clase $(L_{i-1} - L_i)$	Frecuencia relativa (f_i)
35-39	5/28
40-44	2/28
45-49	7/28
50-54	8/28
55-59	6/28

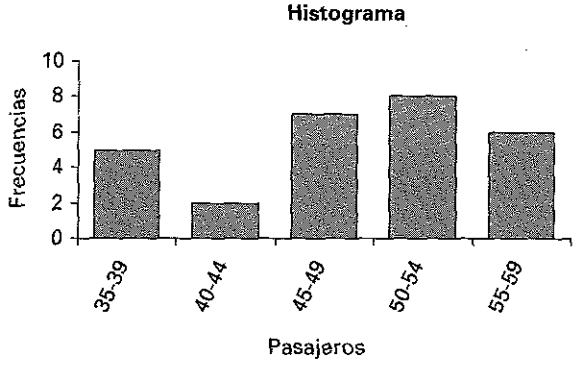
c) Calculamos las frecuencias acumuladas y acumulas relativas.

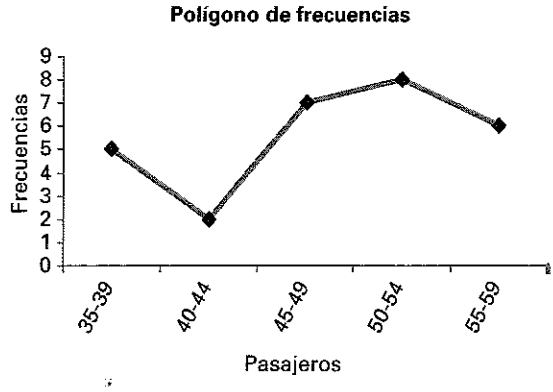
Intervalo de Clase $(L_{i-1} - L_i)$	Frecuencia acumulada (N_i)	Frecuencia relativa acumulada (F_i)	Frecuencia relativa acumulada (F_i) en porcentaje
35-39	5	5/28	17,9%
40-44	7	7/28	25,0%
45-49	14	14/28	50,0%
50-54	22	22/28	78,6%
55-59	28	28/28	100,0%

d) Dado los límites de los intervalos tenemos fácil determinar el número de valores inferiores y superiores a 50; en el primer caso están las frecuencias acumuladas en los tres primeros intervalos (5, 2 y 7 (5 + 2 + 7 = 14) y en el segundo los acumulados en los dos siguientes (8 + 6 = 14); en el capítulo siguiente veremos que este valor se denomina la Mediana de la distribución.

Ejercicio 2.11. *Para el ejercicio anterior, represente el histograma y el polígono de frecuencias.*

Respuesta





Ejercicio 2.12. *La base de datos de una compañía de seguros ofrece la siguiente información diaria sobre el número de pólizas realizadas durante un determinado mes con 26 días laborables:*

62 65 65 57 55 50 65 77 73 30 62 54 48
79 60 63 45 51 68 79 73 33 41 49 55 65

a) *Agrupe estas cifras en una tabla utilizando intervalos de clase adecuados y calcule las frecuencias absolutas, acumuladas, relativas y relativas acumuladas.*

b) *Construya el histograma.*

c) *Construya el polígono de frecuencias.*

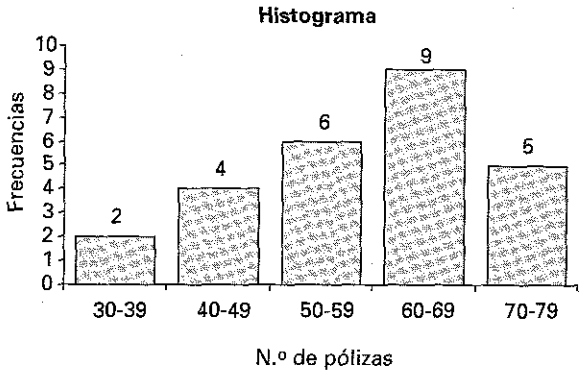
d) *Construya el diagrama de tallos y hojas.*

Respuesta:

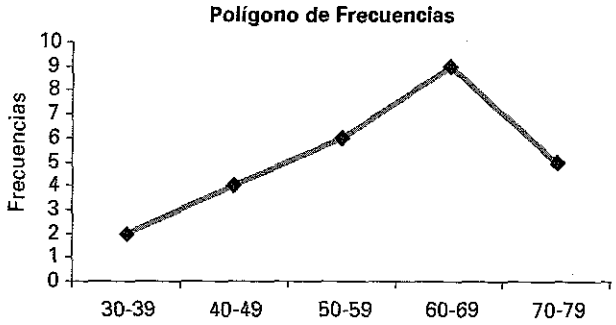
a) Considerando intervalos con amplitud 10, la tabla de frecuencias queda como sigue:

Intervalo de Clase (L_{i-1}-L_i)	Frecuencia absoluta (n_i)	Frecuencia acumulada (N_i)	Frecuencia relativa (f_i)	Frecuencia relativa acumulada (F_i)
30-39	2	2	7,7%	7,7%
40-49	4	6	15,4%	23,1%
50-59	6	12	23,1%	46,2%
60-69	9	21	34,6%	80,8%
70-79	5	26	19,2%	100,0%
Total	26		100,0%	

b) El histograma asociado a la anterior tabla sería:



b) Polígono de frecuencias.



d) Diagrama de tallos y hojas.

3	0	3							
4	8	5	1	9					
5	7	5	0	4	1	5			
6	2	5	5	5	2	5	0	3	8
7	7	3	9	9	3				

Ejercicio 2.13. *Con los datos del ejercicio anterior responda:*

a) *¿Cuántos días se superaron las 50 pólizas de seguro?*

b) *¿Cuántos días se hicieron menos de 70 pólizas?*

c) *¿Qué porcentaje de días se efectuaron entre 30 y 49 pólizas?*

Respuesta

a) 20 días (la suma de las frecuencias de los tres últimos intervalos).

b) 21 días (la suma de las frecuencias de los cuatro primeros intervalos o, lo que es lo mismo, *la frecuencia acumulada* del cuarto intervalo).

c) 6/26 = 3/13 = 23% (la frecuencia relativa del tercer intervalo).

Ejercicio 2.14. *Los trabajadores de una empresa efectuaron el siguiente número de horas extras en una muestra aleatoria de 20 semanas:*

154	204	218	185	168	403	259	233	171	335
329	501	306	331	261	183	188	218	115	112

Construya un gráfico de tallos y hojas con los rótulos de los tallos 1, 2, 3, 4 y 5 (por lo tanto con hojas de dos cifras).

Respuesta

1	54	85	68	71	83	88	15	12
2	04	18	59	33	61	18		
3	35	29	06	31				
4	03							
5	01							

Ejercicio 2.15. *Un hotel conserva un registro de motivos por los que sus huéspedes solicitaron cambios de habitación. Las frecuencias son las siguientes:*

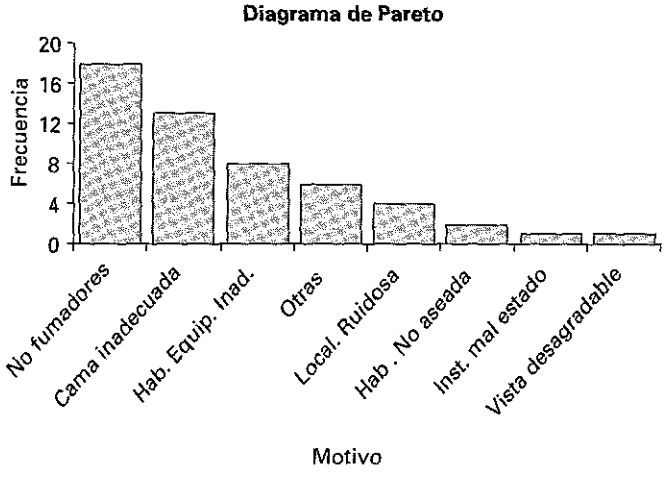
Motivo	Frecuencia
Habitación no aseada	2
Instalación sanitaria en mal estado	1
Cama inadecuada	13
Localización ruidosa	4
Cambio al área no fumadores	18
Vista desagradable	1
Habitación equipada inadecuadamente	8
Otras (no registradas)	6

a) *Construya un diagrama de Pareto.*

a) *El gerente del hotel creía que los principales problemas estaban relacionados con el mantenimiento y la limpieza. ¿Confirma el diagrama que esas fueron las razones principales de las quejas?*

Respuesta

a) Diagrama de Pareto.



b) No, los principales problemas están relacionados con el registro de huéspedes, no con el mantenimiento, ya que la asignación de una habitación con cama diferente a la deseada o con ubicación en el área de fumadores debería haberse resuelto en recepción.

Ejercicio 2.16. *Los salarios en euros de los empleados de una empresa son:*

620	850	1.275	890	730	950	1.200	1.150	690	875
1.270	880	745	950	850	1.230	990	755	770	1.150
630	850	1.270	880	730	960	1.250	1.200	660	850

a) *Construir una tabla de frecuencias agrupadas y obtener la marca de clase de los intervalos.*

b) *Realizar el histograma correspondiente.*

c) *Diseñar la tabla de frecuencias acumuladas absolutas y relativas.*

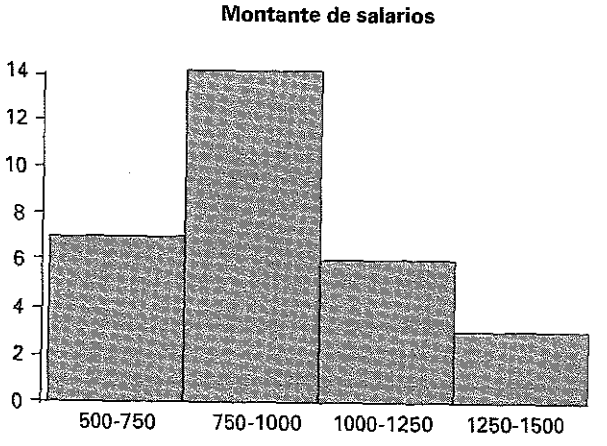
d) *Realizar el gráfico de frecuencias acumuladas.*

Respuesta

a) La tabla de frecuencias asociada es:

$L_{i-1} - L_i$	Marca de clase m_i	Frecuencia n_i	Frecuencia relativa f_i
500 a 750	625	7	23,3%
750 a 1.000	875	14	46,7%
1.000 a 1.250	1.125	6	20,0%
1.250 a 1.500	1.375	3	10,0%
		30	100,0%

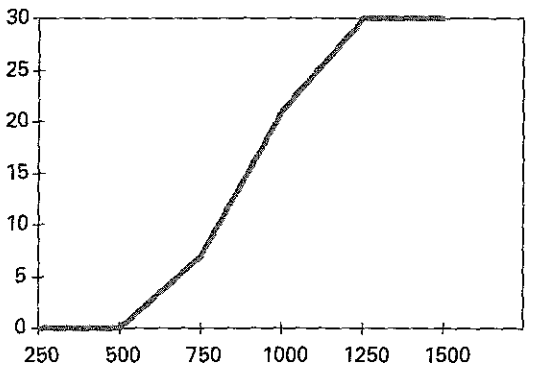
b) El histograma correspondiente sería:



c) La tabla correspondiente a la frecuencia acumulada absoluta y relativa es:

$L_{i-1}-L_i$	**Frecuencia absoluta acumulada N_i**	**Frecuencia relativa acumulada F_i**
500 a 750	7	23,3%
750 a 1000	21	70,0%
1000 a 1250	27	90,0%
1250 a 1500	30	100,0%

d) La gráfica con los valores acumulados obtenidos es:



Ejercicio 2.17. *El gerente de una empresa desea conocer el estado civil de sus trabajadores para asignarles cierta compensación familiar. Una vez recogida la información pertinente y realizado un resumen de los datos, obtuvo la siguiente distribución de frecuencias:*

Categorías	Frecuencias absolutas (n_i)
Casados	52
Divorciados con hijos a cargo	32
Divorciados sin hijos a cargo	41
Solteros	99
Total	224

a) *Identifique la variable en estudio, el tipo a que pertenece y su escala de medición.*

b) *Complete el cuadro calculando las frecuencias relativas correspondientes a cada una de las categorías establecidas.*

c) *Represente las frecuencias calculadas por medio de un gráfico de barras.*

Respuesta

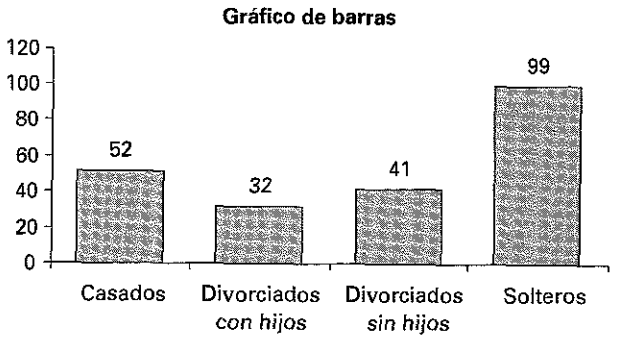
a) El tipo de variable que estamos analizando sería:

Variable en estudio: Estado civil
Tipo de variable: Cualitativa

b) Tabla de frecuencias

Categorías	Frecuencias absolutas n_i	Frecuencias relativas f_i
Casados	52	52/224 = 0,232 = 23,2%
Divorciados con hijos a cargo	32	32/224 = 0,143 = 14,3%
Divorciados sin hijos a cargo	41	41/224 = 0,183 = 18,3%
Solteros	99	99/224 = 0,442 = 44,2%
Total	224	224/224 = 1,000

c) El gráfico de barras queda como sigue:



Ejercicio 2.18. *Con el objetivo de estudiar la clientela de una empresa se llevó a cabo una Encuesta entre 13 de sus clientes elegidos de forma aleatoria y estadísticamente representativa; se les preguntó por el número de veces que se habían interesado en sus servicios Los resultados fueron resumidos en la siguiente tabla de frecuencias:*

N.º de cliente (x_i)	Frecuencia (veces por las que se ha interesado en los servicios) n_i	Frecuencia (F_i)	Frecuencia acumulada	Frecuencia relativa acumulada
1	1	.....	.....	.....
2	1	.....	.....	.....
3	3	.....	.....	.....
4	7	.....	.....	.....
5	17	.....	.....	.....
6	18	.....	.....	.....
7	15	.....	.....	.....
8	28	.....	.....	.....
9	13	.....	.....	.....
10	14	.....	.....	.....
11	1	.....	.....	.....
12	1	.....	.....	.....
13	1	.....	.....	.....

a) *Identifique la variable en estudio, a qué tipo pertenece y la escala de medición.*

b) *Complete la tabla calculando las frecuencias relativas, absolutas acumuladas y relativas acumuladas.*

c) *Represente gráficamente las frecuencias absolutas y absolutas acumuladas.*

Respuesta

a) La variable a analizar será la siguiente.

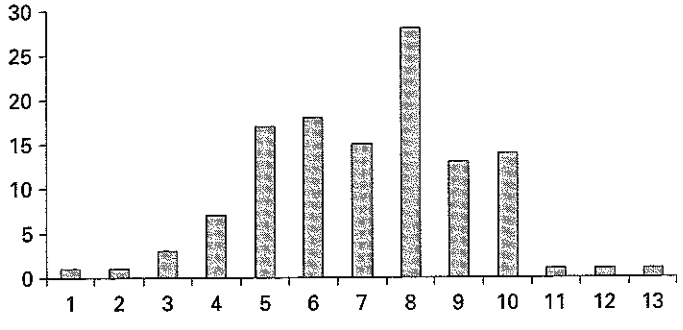
Variable en estudio: Número de veces que se interesa por los servicios de la empresa.

Tipo de variable: Cuantitativa discreta.

b) Tabla de frecuencias.

N.º de cliente (x_i)	Frecuencia (veces por las que se ha interesado en los servicios (n_i)	Frecuencia (F_i)	Frecuencia acumulada	Frecuencia relativa acumulada
1	1	0,8%	1	0,8%
2	1	0,8%	2	1,7%
3	3	2,5%	5	4,2%
4	7	5,8%	12	10,0%
5	17	14,2%	29	24,2%
6	18	15,0%	47	39,2%
7	15	12,5%	62	51,7%
8	28	23,3%	90	75,0%
9	13	10,8%	103	85,8%
10	14	11,7%	117	97,5%
11	1	0,8%	118	98,3%
12	1	0,8%	119	99,2%
13	1	0,8%	120	100,0%
	120	100,0%		

Representación gráfica de la Frecuencia relativa acumulada



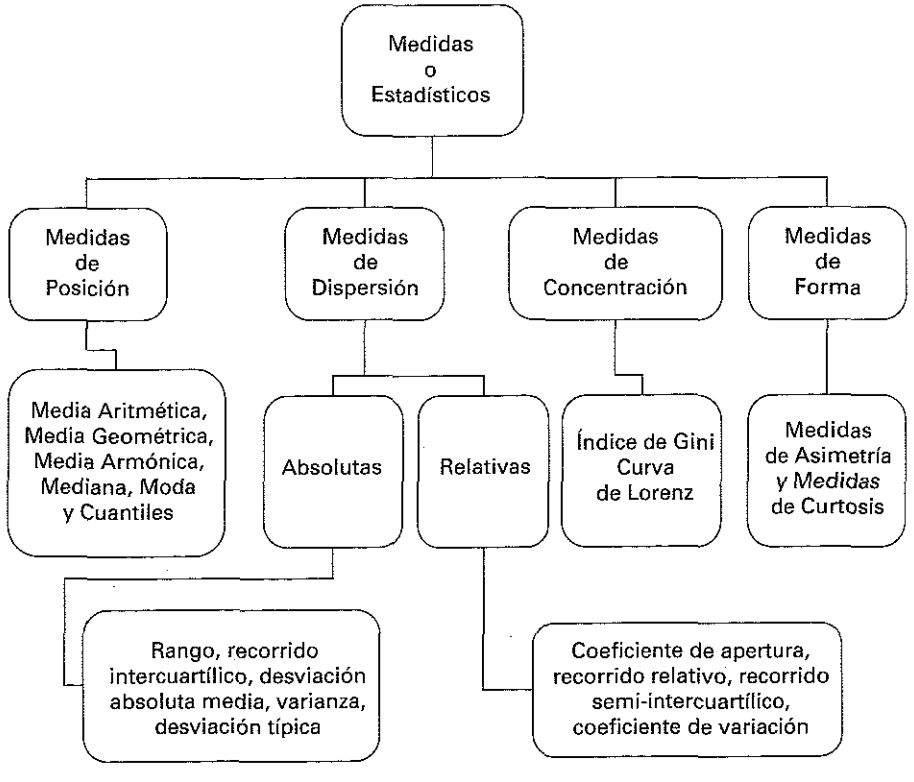
Capítulo 3

LAS MEDIDAS DE POSICIÓN EN DISTRIBUCIONES UNIDIMENSIONALES

3.1. *INTRODUCCIÓN*

Las distribuciones de frecuencias de una variable estadística pueden estudiarse a través de unas medidas, que se conocen con el nombre genérico de ***estadísticos*** y que, analizadas conjuntamente, nos dan un panorama sobre las características de la distribución.

Los estadísticos más habituales de una distribución de frecuencias se agrupan y resumen en el siguiente esquema gráfico:



En este capítulo explicamos las principales medidas de posición y en el siguiente estudiaremos las medidas de dispersión, forma y concentración.

3.2. LA MEDIA ARITMÉTICA

3.2.1. La media aritmética simple

La media aritmética de una variable se define como la suma de todos los valores de la variable dividida por el número total de observaciones; suele denotarse por $\bar{x}$. En las distribuciones de tipo I se obtiene mediante la siguiente formulación:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{i=n} x_i}{N} \qquad [3.2.1]$$

El símbolo $\sum$ significa «Sumatorio»; la expresión $\sum_{i=1}^{i=n} x_i$, que utilizaremos habitualmente en este libro, se lee como «el sumatorio de todos los valores de x desde el primero ($i = 1$) hasta el último ($i = n$)».

Ejemplo 3.1. Los años de antigüedad en una empresa de 5 trabajadores son 6, 5, 4, 3, 2; Obtener la media aritmética simple de estas valoraciones.

Aplicando la expresión [3.2.1], el estadístico pedido será:

$$\bar{x} = \frac{6+5+4+3+2}{5} = \frac{20}{5} = 4$$

3.2.2. La media aritmética ponderada por las frecuencias

En las distribuciones de tipo II o de tipo III es necesario utilizar las frecuencias para obtener la media aritmética; para ello se emplea la siguiente formulación:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_r n_r}{n_1 + n_2 + n_3 + \ldots + n_r} = \frac{\sum_{i=1}^{i=r} x_i n_i}{N} \qquad [3.2.2]$$

Que indica que cada valor ha sido «ponderado» o multiplicado por la frecuencia con la que aparece.

Ejemplo 3.2. Obtener la media aritmética de las valoraciones de 0 a 10, otorgadas por 20 clientes sobre la percepción del servicio de atención al cliente de una determinada empresas; los valores otorgados por dichos clientes son:

4 3 3 5 2 3 0 2 1 5 6 7 8 1 6 7 4 6 4 3

Podríamos operar aplicando la expresión [3.2.1], pero también y dado que algunas de las valoraciones se repiten varias veces, podemos agruparlas obteniendo una tabla de frecuencias de tipo II; esta agrupación es inevitable cuando disponemos de un gran número de observaciones (imaginemos una encuesta sobre valoraciones realizada a 1.000 clientes de la empresa).

Construimos una tabla de frecuencias de tipo II, es decir, con frecuencias agrupadas, que quedaría de la siguiente forma:

x_i	n_i	$x_i \cdot n_i$
0	1	0
1	2	2
2	2	4
3	4	12
4	3	12
5	2	10
6	3	18
7	2	14
8	1	8
9	0	0
10	0	0
Suma	20	80

Aplicando la expresión [3.2.2] tendremos:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_r n_r}{n_1 + n_2 + n_3 + \ldots + n_r} = \frac{\sum_{i=1}^{i=r} x_i n_i}{N} = \frac{80}{20} = 4$$

En el caso de que los datos estén agrupados en clases, es decir, en las distribuciones de tipo III, se opera igual que en los casos anteriores, tomando la marca de clase m_i como x_i.

El alumno puede comprobar que el resultado obtenido por este método es similar al que se hubiese obtenido aplicando la fórmula [3.2.1]

Ejemplo 3.3. Obtener la media aritmética de la siguiente distribución de distancias entre ciudades:

L_{i-1} - L_i	n_i
(0 - 5 Km.]	4
(5 - 10 Km.]	10
(10 - 25 Km.]	6
(25 - 50 Km.]	40
(50 - 100 Km.]	5
(100 - 500 Km.]	35

Aproximamos las marcas de clase mediante la expresión $m_i = x_i = \frac{L_{i-1} + L_i}{2}$, de dónde queda:

L_{i-1} - L_i	$m_i = x_i$	n_i	$x_i \cdot n_i$
(0 - 5 Km.]	2,5	4	10
(5 - 10 Km.]	7,5	10	75
(10 - 25 Km.]	17,5	6	105
(25 - 50 Km.]	37,5	40	1.500
(50 - 100 Km.]	75	5	375
(100 - 500 Km.]	300	35	10.500
Suma		100	12.565

La media aritmética, aplicando la expresión [3.2.2] quedaría:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_r n_r}{n_1 + n_2 + n_3 + \ldots + n_r} = \frac{\sum_{i=1}^{i=r} x_i n_i}{N} = \frac{12.565}{100} = 125{,}65 \text{ Km.}$$

No siempre puede calcularse la media aritmética, a veces incluso, tiene un sentido limitado o no útil.

Particularmente, la media aritmética *no es posible calcularla cuando los datos son cualitativos* (no tiene mucho sentido obtener una media aritmética entre los colores o

tonalidades de pelo o para la valoración de algún ratio en el que las opciones o atributos sean «excelente, buena y mala», salvo que le adjudiquemos, en este caso, unas determinadas puntuaciones de, por ejemplo, excelente equivale a 9, buena a 6 y mala a 4); también existe alguna dificultad para el cálculo y *sólo puede ser aproximada cuando los datos están agrupados en intervalos* abiertos.

Nótese que en cierto modo este es el caso de nuestro ejemplo anterior, en el que hemos calculado las marcas de clase aproximando los extremos de los intervalos, haciendo, por ejemplo, (0+5]/2 = 2,5 Km. para el primer intervalo, lo que en realidad es una aproximación; este cálculo hubiese sido más complicado si hubiésemos trabajado, por ejemplo, con la renta anual media de nuestros clientes y la información disponible estuviese en intervalos abiertos como el del Ejemplo 3.4.

Ejemplo 3.4. Mediante una encuesta efectuada a 1.000 compradores de un determinado producto, se ha obtenido que su renta anual, en euros, es la siguiente:

Renta anual de los clientes en euros	n_i
Menos de 10.000 €	50
De 10.001 a 20.000 €	100
De 20.001 a 50.000 €	600
De 50.001 a 100.000 €	150
Más de 100.000 €	100
Suma	1.000

Con esta información, calcúlese la media aritmética muestral.

En este caso sólo podríamos aproximar la media aritmética haciendo una hipótesis sobre la marca de clase de los dos intervalos abiertos, es decir, del primero (menos de 10.000 € y del último (más de 100.000 €).

Así, por ejemplo, podríamos suponer que la media de los que ganan menos de 10.000 € es 5.000 € (media entre 0 y 10.000 €) y la media de los que ingresan «Más de 100.000 euros anuales» es 150.000 €.

Con esta hipótesis construimos la marca de clase x_i en la siguiente forma:

Renta anual de los clientes en euros	n_i	$m_i = x_i$	$x_i \cdot n_i$
Menos de 10.000 €	50	5.000	250.000
De 10.001 a 20.000 €	100	15.000	1.500.000
De 20.001 a 50.000 €	600	35.000	21.000.000
De 50.001 a 100.000 €	150	75.000	11.250.000
Más de 100.000 €	100	150.000	15.000.000
Suma	1.000		49.000.000

Operando con la expresión [3.2.2] tendremos:

$$\bar{x} = \frac{\sum_{i=1}^{i=5} x_i n_i}{N} = \frac{49.000.000}{1.000} = 49.000$$

Que nos da una media aritmética de 49000 € de renta.

Pero cualquier otra hipótesis sobre las marcas de clase, que podría ser tan válida como la primera, nos daría otro resultado. Así, podíamos hacer la hipótesis de que el primer intervalo tiene una media de 6.000 (suponiendo que nadie tiene 0 y no es realista la hipótesis anterior) y que hay muchos clientes con alta renta, que aumentan la marca de clase del último intervalo hasta 250.000 €; con esta segunda hipótesis tendríamos:

Renta anual de los clientes en euros	n_i	$m_i = x_i$	$x_i \cdot n_i$
Menos de 10.000 €	50	6.000	300.000
De 10.001 a 20.000 €	100	15.000	1.500.000
De 20.001 a 50.000 €	600	35.000	21.000.000
De 50.001 a 100.000 €	150	75.000	11.250.000
Más de 100.000 €	100	250.000	25.000.000
Suma	1.000		59.050.000

$$\bar{x} = \frac{\sum_{i=1}^{i=5} x_i n_i}{N} = \frac{59.050.000}{1.000} = 59.050$$

Con lo que la media de la distribución sería 59.050 €.

3.2.3. La media aritmética ponderada por coeficientes

En ocasiones resulta conveniente introducir un coeficiente de ponderación que de mayor peso a algunos valores de la variable. Estos coeficientes o pesos suelen denominarse w_i, dando lugar a la siguiente expresión de la media aritmética ponderada para distribuciones de tipo I:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \qquad [3.2.3]$$

Para las distribuciones de tipo *II* y *III* sería:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i w_i}{\sum_{i=1}^{r} n_i w_i} \qquad [3.2.4]$$

Para ilustrarnos sobre la importancia de este tipo de ponderaciones veamos los siguientes ejemplos:

Ejemplo 3.5. Supongamos que queremos hacer una selección de personas para cubrir un puesto comercial en una empresa en donde se considera muy importante el conocimiento de Inglés, dado que se tiene un importante número de clientes extranjeros, y menos importante del de otras materias como la Estadística o el Marketing; en el currículum vitae de un candidato tenemos información sobre las notas medias obtenidas en distintos bloques de conocimientos; se considera que la calificación de Inglés debe ponderarse el doble que la del resto de materias:

Materias	Calificaciones		Coeficientes de ponderación
	Candidato 1	Candidato 2	
Inglés	5	7	2
Marketing	6	5	1
Estadística	10	8	1

En este caso tenemos una distribución de tipo I (frecuencias unitarias). Utilizando la expresión [3.2.3], tendríamos para ambos alumnos la siguiente nota media ponderada:

$$\bar{x}_1 = \frac{5\cdot 2 + 6\cdot 1 + 10\cdot 1}{4} = \frac{26}{4} = 6{,}5 \qquad \bar{x}_2 = \frac{7\cdot 2 + 5\cdot 1 + 8\cdot 1}{4} = \frac{27}{4} = 6{,}75$$

Si hubiésemos operado con la media aritmética simple daríamos una nota media de 7 (21 entre 3) para el primer candidato y una media de 6,66 (20 partido de 3) para el segundo; elegiríamos, en consecuencia, al primero; al introducir los coeficientes de ponderación, la elección recaería en el segundo candidato.

Ejemplo 3.6. Supongamos que queremos hacer un seguimiento de la valoración que dan los clientes de una empresa con los servicios recibidos; realizamos para ello una encuesta a una muestra representativa de de dichos clientes en la que les pedimos que valoren de 0 a 10 su grado de satisfacción con el citado servicio.

Teniendo en cuenta la importancia de las ventas de cada estrato, obtener la media aritmética ponderada según el número de clientes y la importancia del montante facturado a cada cliente.

Los resultados obtenidos son los siguientes:

Estratos de ventas anuales	**N.º de clientes n_i**	**% de ventas realizado w_i**	**Valoración media otorgada (puntuación de 0 a 10) x_i**
Más de 500.000 €	5	10%	6
De 100.000 a 500.000 €	25	15%	4
De 10.000 a 100.000 €	80	45%	9
Menos de 10.000 €	390	30%	3
Total	500	100%	

Media aritmética ponderada por las frecuencias:

$$\overline{x}_1 = \frac{\sum_{i=1}^{r} x_i n_i}{\sum_{i=1}^{r} n_i} = \frac{5 \cdot 6 + 25 \cdot 4 + 80 \cdot 9 + 390 \cdot 3}{500} = \frac{2.020}{500} = 4,04$$

En este caso se ha ponderado por el número de clientes; o lo que es lo mismo, a la opinión de cada cliente se le ha dado el mismo valor; la media obtenida 4,04 está más próxima a la expresada por el grupo en el que hay más clientes (los 390 clientes con ventas inferiores a 10.000 € valoraron el servicio con un 3).

$$\overline{x}_2 = \frac{\sum_{i=1}^{r} x_i w_i}{\sum_{i=1}^{r} w_i} = \frac{10 \cdot 6 + 15 \cdot 4 + 45 \cdot 9 + 30 \cdot 3}{100} = 6,15$$

En este caso se ha ponderado por la importancia de las ventas; cómo el grupo de clientes cuyas ventas se sitúan entre 10.000 y 100.000 € valora muy bien los servicios

prestados (un 9) y representan un alto porcentaje de las ventas (45%), la media de valoración de la empresa ha subido notablemente hasta un 6,15; de ambos hechos debe concluirse que aunque nuestra empresa no atiende bien con carácter general a sus clientes (obtiene un suspenso de media), sí lo hace adecuadamente con el grupo de sus clientes medianos.

REFLEXIONES

La estadística descriptiva debe adaptarse a los fines del análisis que estemos realizando, de forma que según la variable con la que estemos trabajando y según el motivo para el que realicemos dicho análisis, puede interesar prescindir de un cálculo exacto de la media aritmética en aras de obtener una mejor información de conjunto; así, por ejemplo, en el Ejemplo 3.4, sí estamos programando una política de marketing para nuestra empresa, puede interesarnos más trabajar con la distribución de nuestros clientes por estratos de nivel de renta anual, que con la media aritmética de su renta; el estadístico renta media anual apenas tendría un significado, mientras que saber que un 60% de nuestros clientes tienen una renta comprendida entre 20.001 y 50.000 € anuales, nos permite perfilar determinadas técnicas de marketing que podemos suponer adecuadas para las familias con este estrato de renta.

El ejemplo 3.6. pone también de manifiesto el carácter instrumental de la Estadística, mostrando una situación en la que, según se emplee uno u otro indicador, se obtienen resultados distintos que ponen en todo caso de manifiesto el trato discriminatorio dado a pequeños y medianos o grandes clientes de la empresa.

Otro aspecto a tener en cuenta, en función también de los objetivos perseguidos por nuestro trabajo estadístico, es la exactitud de los datos disponibles; alguien podría pensar que siempre es preferible trabajar con datos agrupados mediante distribuciones de tipo II en vez de con datos por intervalos mediante distribuciones de tipo III, ya que es evidente que al agruparlos en intervalos se pierde exactitud y precisión en la información; en realidad no siempre es así; la información sobre la renta manejada en el ejemplo anterior, probablemente tuvo que ser obtenida mediante una pregunta en una Encuesta a una muestra representativa de nuestros clientes; sí hubiésemos querido obtener información más precisa, preguntando sobre la renta exacta, ¿lo habríamos conseguido?; seguramente no, porque habría habido más clientes que se hubiesen negado a contestar la pregunta con ese detalle, otros que ni siquiera habrían podido ser precisos en el momento de realizar la Encuesta, etc..; en cambio, sí recabamos una respuesta por intervalos relativamente imprecisos y amplios, la pregunta puede ser contestada con menor reticencia y por un mayor número de entrevistados, lo que nos aportará mayor exactitud en las estimaciones.

Llegado a este punto, es importante resaltar que para un usuario de la estadística, como puede ser en nuestro caso un administrador o director de una empresa, por encima del conocimiento con detalle de los cálculos matemáticos de un estadístico, están los planteamientos lógicos; cualquier programa informático-estadístico o una simple calculadora puede darnos la media aritmética de unos datos, pero, generalmente, tiene más importancia saber sí tiene sentido trabajar con este estadístico cal-

culado, su interés para el tema de trabajo y sus ventajas, limitaciones, e inconvenientes, que los propios detalles del cálculo.

3.2.4. Propiedades de la media aritmética

La media aritmética tiene, entre otras, las siguientes **propiedades:**

1. La suma de las desviaciones de todos los valores respecto a su media aritmética es cero, es decir:

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

Ejemplo 3.7.

x_i	n_i	$x_i \cdot n_i$
8	10	80
3	15	45
15	25	375
Total	50	500

La media aritmética será:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{\sum_{i=1}^{r} n_i} = \frac{500}{50} = 10$$

Obteniendo la columna diferencia y multiplicando por la de frecuencias n_i, se comprueba que la suma

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

$x_i - \bar{x}$	$(x_i - \bar{x})n_i$
–2	–20
–7	–105
5	125
	$\sum_{i=1}^{r}(x_i - \bar{x})n_i = 0$

2. Si multiplicamos o dividimos todas las observaciones por un mismo número, lo que se conoce como *cambio de escala*, la media queda multiplicada o dividida por dicho numero.
3. Si le sumamos a todas las observaciones un mismo número, lo que se conoce como *cambio de origen*, la media aumentará en dicha cantidad[3].

Como consecuencia de estas dos últimas propiedades, *si a la variable estadística* x_i *la sometemos al mismo tiempo a un cambio de origen* O_t *y a un cambio de escala* C *mediante la transformación:* $y_j = \frac{x_j - O_t}{C}$ *(siendo* O_t *y* C *constantes), resulta que:* $\overline{x} = C_{\overline{y}} + O_t$.

Esta propiedad es bastante utilizada para la simplificación de los cálculos cuando los valores observados son muy elevados y tienen un máximo común divisor; veamos un ejemplo:

Ejemplo 3.8. Tenemos la siguiente distribución de frecuencias

x_i	n_i
10.000	2
15.000	5
25.000	3
30.000	6
35.000	4

Podemos comprobar cómo se cumple la primera propiedad:

x_i	n_i	$x_i \cdot n_i$
10.000	2	20.000
15.000	5	75.000
25.000	3	75.000
30.000	6	180.000
35.000	4	140.000
Suma	20	490.000

[3] La demostración de estas propiedades puede verse en textos de Estadística más avanzada; por ejemplo en: Casas Sánchez y Santos Peñas, «Introducción a la Estadística para Administración y Dirección de Empresas». Editorial CEURA. 2002.

De dónde se deduce que la media aritmética es 24.500:

$$\bar{x} = \frac{\sum x_i n_i}{N} = \frac{490.000}{20} = 24.500$$

Obteniendo la columna de diferencias se observa que:

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

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})n_i$
10.000	2	–14.500	–29.000
15.000	5	–9.500	–47.500
25.000	3	500	1.500
30.000	6	5.500	33.000
35.000	4	10.500	42.000
Suma	20		0

Los cambios de escala nos pueden facilitar los cálculos; sí realizamos un cambio de escala para poder operar mejor: (C = 1000), nos queda la siguiente distribución para la nueva variable y_i.

x_i	$y_i = \frac{x_i}{1.000}$	n_i
10.000	10	2
15.000	15	5
25.000	25	3
30.000	30	6
35.000	35	4

Obtenemos la media aritmética para la nueva variable y_i

$$\bar{y} = \frac{1}{20}(10 \cdot 2 + 15 \cdot 5 + 25 \cdot 3 + 30 \cdot 6 + 35 \cdot 4) = \frac{490}{20} = 24{,}5$$

Y, aplicando la propiedad estudiada, obtenemos la media aritmética para la variable original x_i

$$\bar{x} = C\bar{y} = 1.000 \cdot 24{,}5 = 24.500$$

Para facilitar los cálculos, también podríamos haber efectuado un cambio de origen del tipo: $O_i = 25.000$, de forma que

x_i	$y_i = x_i - 25.000$	n_i
10.000	–15.000	2
15.000	–10.000	5
25.000	0	3
30.000	5.000	6
35.000	10.000	4

$$\bar{y} = \frac{1}{20}\left(-15.000 \cdot 2 - 10.000 \cdot 5 + 0 \cdot 3 + 5.000 \cdot 6 + 10.000 \cdot 4\right) =$$

$$= \frac{-80.000 + 70.000}{20} = -500$$

$$\bar{x} = \bar{y} + O_i = -500 + 25.000 = 24.500$$

O, podríamos realizar conjuntamente las dos transformaciones:

x_i	$y_i = \dfrac{x_i - 25.000}{1.000}$	n_i
10.000	–15	2
15.000	–10	5
25.000	0	3
30.000	5	6
35.000	10	4

$$\bar{y} = \frac{1}{20}\left(-15 \cdot 2 - 10 \cdot 5 + 0 \cdot 3 + 5 \cdot 6 + 10 \cdot 4\right) = \frac{-80 + 70}{20} = -0{,}5$$

Transformando nuevamente los valores a nuestra variable inicial X, se tiene:

$$\bar{x} = C\bar{y} + O_i = 1.000 \cdot (-0{,}5) + 25.000 = -500 + 25.000 = 24.500$$

Según los datos disponibles, algunas transformaciones de este tipo pueden facilitar sustancialmente los cálculos de la media aritmética.

3.2.5. Ventajas e inconvenientes de la media aritmética

Las principales **ventajas** de la media aritmética son las siguientes:

— Se trata de un concepto familiar para la mayoría de las personas y es intuitivamente claro.
— Es calculable en todas las variables, es decir siempre que nuestras observaciones sean cuantitativas.
— Para su cálculo se utilizan todos los valores de la distribución.
— Es única para cada distribución de frecuencias
— Tiene un claro significado, ya que al ser el *centro de gravedad* de la distribución representa todos los valores observados.
— Es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.

Los principales **inconvenientes** son:

— Que *es un valor muy sensible a los valores extremos*, con lo que en las distribuciones con gran dispersión de datos puede llegar a peder totalmente su significado. Recordemos aquí la famosa anécdota del pollo, si una persona se come un pollo y otra no come pollo, como media, entre las dos se habrán comido medio pollo cada una.
— Que *no es calculable cuando los parámetros son cualitativos*.
— Podemos tener *dificultades para su cálculo en distribuciones de tipo III con intervalos abiertos*; en estos casos es necesario estimar una marca de clase para poder calcular la media y ésta nos varía sí cambiamos la marca de clase.

3.3. MEDIA GEOMÉTRICA

La media geométrica de una distribución de frecuencias con N observaciones es ***la raíz de índice N del producto de todas las observaciones elevado a sus respectivas frecuencias***.

La representaremos por *G* y adopta la siguiente formulación:

— En distribuciones unitarias o distribuciones de tipo I:

$$G = \sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n} = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

— En distribuciones no unitarias:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \ldots \cdot x_r^{n_r}} = \sqrt[N]{\prod_{i=1}^{r} x_i^{n_i}}$$

El signo $\sqrt[N]{\prod_{i=1}^{r} x_i^{n_i}}$ indica que se trata del producto de todos los valores de la variable x, desde el primero ($i = 1$) hasta el último ($i = r$).

El estadístico media geométrica *sólo se puede calcular si no hay observaciones nulas*, ya que si algún valor es cero se anulan los productos y la media toma valor cero.

También *puede no tener sentido su cálculo cuando algún valor es negativo*, ya que en este caso, podemos obtener números irracionales que desnaturalicen el valor del estadístico.

Es una medida estadística que debe emplearse cuando los valores de la variable no son de naturaleza aditiva (*tasas, tipos de interés, porcentajes, números índices*, etc.).

Las principales **ventajas** de la media geométrica son las siguientes:

— En su determinación intervienen todos los valores de la distribución.
— Es menos sensible que la media aritmética cuando la distribución tiene valores extremos.
— Por su formulación *es más representativa que la media aritmética cuando la distribución evoluciona de forma acumulativa o con efectos multiplicativos.*
— Cuando existe, es decir, cuando la distribución no tiene valores negativos, y cuando está definida, es decir cuando la distribución no tiene valores nulos, su valor está definido de forma objetiva y es único.

Los principales **inconvenientes** son:

— Su significado estadístico es menos intuitivo que la media aritmética.
— La mayor complicación de los cálculos
— Su indefinición (da números con naturaleza imaginaria) cuando tiene valores negativos y su valor nulo cuando una observación toma este valor.

Dada su formulación, el cálculo de la media geométrica exigirá normalmente la utilización de logaritmos o de programas informáticos.

En este sentido, recordemos que por la propia formulación de la media geométrica se cumple la siguiente propiedad:

$$\log G = \frac{1}{N}\sum_{i=1}^{r} n_i \log x_i \quad ; \text{ en neperianos} \qquad \ln G = \frac{1}{N}\sum_{i=1}^{r} n_i \ln x_i$$

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

Ejemplo 3.9. Una empresa de fabricación de automóviles que ha iniciado su actividad en 2007 experimenta durante 2008 un aumento de las ventas de sus vehículos de un 10%; en 2009 sus ventas aumentan un 20% más y en 2010 vende un 25% más que en 2009. ¿Cuál es la tasa media de incremento durante el último trienio?

Dado que la variable presenta una evolución multiplicativa es preferible utilizar la media geométrica para el cálculo.

Considerando un valor 100 para 2006, tendríamos:

2007: 100
2008: 110 (es decir, un 10% más que 100)
2009: 132 (es decir, un 20% más que 110)
2010: 165 (es decir, un 25% más que 132)

$$G = \sqrt[N]{x_1^1 \cdot x_2^1 \cdot x_3^1} = \sqrt[N]{\prod_{i=1}^{i=3} x_i} = \sqrt[3]{110 \cdot 120 \cdot 125} \approx 118{,}17$$

Lo que indica que durante el período las ventas han aumentado con una media del 18,17% anual. Así, si el primer año son 100, el segundo año serán $100 \times 1{,}1817 \approx 118{,}17$, el tercer año: $118{,}17 \times 1{,}1817 \approx 139{,}63$ y el cuarto año: $139{,}63 \times 1{,}1817 \approx 165$.

La media aritmética sería:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{110 + 120 + 125}{3} = 118{,}33$$

Es decir, aritméticamente se obtiene un incremento anual del 18,33% y mediante la media geométrica un incremento algo inferior del 18,17%.

Esta relación siempre se cumple, de forma que $G \leq \bar{X}$.

Ejemplo 3.10. En los años 2007, 2008 y 2009 los 20 establecimientos de una empresa han experimentado el siguiente crecimiento en su número de clientes:

Año	Crecimiento	N.º de establecimientos
2007	Crecimiento del 10%	4
2008	Crecimiento del 15%	6
2009	Crecimiento del 20%	10

Si suponemos que comenzó a operar en el año 2006 con 100 clientes. ¿Cuál fue la tasa media de incremento durante los últimos 3 años?

En este caso se trata de una distribución de tipo II, en la que la media geométrica viene dada por la expresión:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot x_3^{n_3} \cdot \ldots \cdot x_r^{n_r}} = \sqrt[N]{\prod_{i=1}^{3} x_i^{n_i}} = \sqrt[20]{1{,}10^4 \cdot 1{,}15^6 \cdot 1{,}20^{10}} =$$

$$= \sqrt[20]{1{,}464 \cdot 2{,}313 \cdot 6{,}192} = \sqrt[20]{20{,}969} = 1{,}1643$$

donde como puede observarse hemos decido, dado que las cifras en tanto por cien resultarían muy elevadas, efectuar el cálculo a partir de crecimientos unitarios, obteniendo por tanto un crecimiento medio anual del 16,43%.

Si trabajamos con logaritmos neperianos, aplicando la expresión

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

y trabajando en tantos por cien, obtendríamos:

$$\ln G = \frac{1}{20}(4 \ln 110 + 6 \ln 115 + 10 \ln 120) =$$

$$= \frac{1}{20}(4 \cdot 4{,}700 + 6 \cdot 4{,}745 + 10 \cdot 4{,}787) = \frac{95{,}146}{20} = 4{,}76$$

G = antilogaritmo neperiano (exponencial) de 4,76 = 116,43; el mismo resultado que hemos obtenido con el método general.

La media de crecimiento sería, pues, del 16,43%, ligeramente inferior a 16,50% que se obtiene como media aritmética:

$$\bar{x} = \frac{\sum_{i=1}^{i=r} x_i n_i}{N} = \frac{(10 \cdot 4) + (15 \cdot 6) + (20 \cdot 10)}{20} = \frac{40 + 90 + 200}{20} = 16{,}5$$

Cualquiera de los dos promedios (el aritmético o el geométrico) puede ser válido para la obtención de un ratio medio, sin embargo, dado que «la variable evoluciona de forma acumulativa y con efectos multiplicativos», la media geométrica refleja más adecuadamente el promedio de crecimiento y debería utilizarse como mejor estadístico descriptivo del promedio.

Cómo puede compararse en el ejemplo, el cálculo de la media geométrica en distribuciones de tipo II y III es relativamente complicado y exige la utilización de programas informáticos o de tablas logarítmicas.

3.4. MEDIA ARMÓNICA

La media armónica de N observaciones es la inversa de la media de las inversas de las observaciones; suele denotarse con la letra H.

En distribuciones unitarias o de tipo I, viene dada por la siguiente formulación:

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

En distribuciones de tipo II, vendrá definida por:

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

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

Las principales **ventajas** de la media armónica son las siguientes:

— Está definida de forma objetiva y es única.

— Para su cálculo tiene en cuenta todos los valores de la distribución.

— Es más representativa que otras medidas en los casos de obtener *promedios de velocidades, rendimientos, productividades*, etc.

— Los valores extremos tienen una menor influencia que en la media aritmética.

Los principales **inconvenientes** son:

— Matemáticamente *sólo se puede calcular si no hay observaciones iguales a cero*, ya que en este caso nos aparecería un cociente indeterminado del tipo

$$\frac{n_i}{0} = \infty.$$

— *Cuando la variable toma algunos valores muy pequeños puede carecer de significado*; en estos casos sus inversos pueden aumentar casi hasta el infinito, eliminando el efecto del resto de los valores.

Ejemplo 3.11. Cinco trabajadores de una empresa de catering han producido 200, 250, 300, 350 y 400 raciones de un determinado preparado alimenticio con unos rendimientos, respectivamente, de 10, 15, 17, 19 y 20 unidades por trabajador y hora. Calcular el rendimiento medio de los trabajadores de la empresa.

En este ejemplo el ritmo de producción es el rendimiento obtenido por trabajador y la producción son las raciones alimenticias obtenidas por cada trabajador. Dado que es la distribución de *una variable cociente*, es adecuado emplear la media armónica.

La distribución de frecuencias sería:

x_i	n_i
10	200
15	250
17	289
19	361
20	400
Suma	1.500

Y la media armónica vendría dada por:

$$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_r}{x_r}} = \frac{1.500}{\frac{200}{10} + \frac{250}{15} + \frac{289}{17} + \frac{361}{19} + \frac{400}{20}} =$$

$$= \frac{1.500}{20 + 18 + 17 + 19 + 20} = 16{,}17$$

El alumno puede comprobar que la media aritmética es 17,00 y la media geométrica es 16,62.

3.5. RELACIÓN ENTRE LAS MEDIAS ARMÓNICA, GEOMÉTRICA Y ARITMÉTICA

Puede demostrarse que en todas las distribuciones de frecuencias la media armónica es, si existe, igual o inferior a la media geométrica y que ésta es a su vez (también, sí existe), igual o inferior a la media aritmética.

Se cumple, pues, siempre que existan los distintos estadísticos, la siguiente relación:

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

En el ejemplo anterior se tiene que:

$$16{,}17 < 16{,}62 < 17{,}00$$

3.6. LA MEDIANA

Las medias estudiadas hasta ahora son medidas que tratan de equilibrar los valores de una distribución compensando los más grandes con los más pequeños para buscar su centro de gravedad o posicionamiento central; estos estadísticos tienen en general dos problemas:

— Son muy sensibles a los valores extremos de las distribuciones de forma que cuando existe mucha dispersión los hacen poco representativos, y

— No es posible calcularlos en las distribuciones cualitativas (por ejemplo, no tiene mucho sentido calcular la media aritmética, armónica o geométrica, de los colores del pelo de un grupo de individuos).

Para resolver el primero de estos problemas se emplea **la mediana**, que es un estadístico que en vez de equilibrar los valores de la variable equilibra las frecuencias y para resolver el segundo inconveniente se utiliza **la moda**, que es un estadístico que toma en consideración los valores más repetidos.

La mediana de una distribución de frecuencias, previamente ordenada en orden creciente o decreciente, se define como el valor central de la variable que divide la distribución en dos partes iguales; es decir, es el valor que deja el mismo número de observaciones o de frecuencias a su izquierda que a su derecha.

A) Cálculo de la mediana en el caso de distribuciones de tipo I

Se presentan dos casos diferentes:

1. **Que el número de observaciones, *N*, sea impar:** En este caso hay un término central, el término $x_{\frac{N+1}{2}}$ que será el valor de la mediana.

Ejemplo 3.12. Calcular la mediana de una distribución estadística con los siguientes valores, indicativos de los años de antigüedad en una empresa de 5 trabajadores:

5, 9, 3, 8, 11

Lo primero que debe hacerse es ordenar dichos valores; los ordenamos en sentido creciente y tendremos:

3, 5, 8, 9, 11

Como son 5 valores (impar), la mediana vendrá dada por el valor central 8, que deja a su izquierda la misma masa (2 valores, es decir dos trabajadores) que a su derecha (otros 2 trabajadores).

Obsérvese el muy diferente significado que tiene la mediana respecto de la media; la media nos diría que la plantilla de la empresa en cuestión tiene como media 7,2 años de antigüedad (media aritmética de los valores 3, 5, 8, 9, 11), mientras que la mediana nos indica que «la mitad» de la plantilla de esta empresa tiene más de 8 años de antigüedad y la otra mitad menos de 8 años de antigüedad en la misma.

2. ***Que el número de observaciones, N, sea Par:*** En este caso hay dos términos centrales, y; la mediana será la media aritmética de esos dos valores.

Ejemplo 3.13. Calcular la mediana de una distribución estadística con los siguientes valores, indicativos del número de reclamaciones mensuales efectuadas por los clientes de una determinada empresa:

2, 3, 5, 6, 8, 9, 12, 16, 20, 24, 25, 26

En este caso N = 12, un número par; la distribución está ya ordenada, por lo que debemos fijarnos en los dos términos centrales, es decir, los que ocupan la posición 6ª y 7ª (9 y 12) y la mediana vendrá dada por:

$$M_e = \frac{9+12}{2} = 10{,}5$$

Si la variable no admite decimales, como sería el caso de este ejemplo, se acepta que existen conjuntamente dos medianas con los dos valores centrales 9 y 12, ya que valores iguales o menores a 9 hay 6 y valores iguales o superiores a 12 también hay 6; en el ejemplo podría decirse, pues, indiferentemente, que la mitad de los meses (es decir, en 6 meses) hay más de 9 reclamaciones y que en la mitad de los meses hay menos de 12.

Con este criterio, queda una vez más indicado que la estadística descriptiva no es una ciencia exacta; es más bien un conjunto de reglas y convenios que ayudan a simplificar e interpretar la información disponible; cuando gestionamos una empresa tenemos disponible la información del ejemplo en el registro de reclamaciones; en cualquier momento el administrador de la empresa puede pedirnos un informe sobre la marcha del mismo y no podemos confeccionarlo con vaguedades como «unos meses tenemos sólo dos quejas y otros 8, 10 6 26»; sí manejamos adecuadamente los ratios estadísticos, podremos indicar, por ejemplo, que a lo largo del año hemos tenido como media 13 reclamaciones mensuales, que la mitad de los meses tenemos menos de 12 y que la otra mitad tenemos más de 9, que cómo máximo hemos tenido 26 y cómo mínimo 2, etc..

Nótese que en el Ejemplo 3.13, **los valores observados están previamente ordenados**; cuando no se presenten ordenados es necesario proceder previamente a la ordenación de las observaciones, ya que de lo contrario el cálculo de la mediana es totalmente erróneo y puede carecer de significado.

Así, por ejemplo, sí no hubiésemos ordenado los datos del ejemplo 3.12 hubiésemos concluido que la mediana era 3, lo que sería erróneo, ya como es evidente «la mitad» de la plantilla no tiene más de 3 años de antigüedad en la empresa.

B) Cálculo de la mediana en las distribuciones de tipo II

Para determinar la mediana en distribuciones no unitarias o de tipo II, es preciso también ordenar los valores y trabajar con la frecuencia absoluta acumulada N_i, obteniendo en concreto el valor $N/2$.

Al igual que en las distribuciones de tipo I, por convenio generalmente aceptado entre los estadísticos, se distinguen dos casos:

1. Que exista un N_i igual a N/2

   En este caso la mediana es la media aritmética de X_i y del siguiente X_{i+1}; si la variable no admite decimales, la mediana serían los dos valores, conjuntamente.

2. Cuando no existe un N_i que iguale a $N/2$, la mediana corresponde al primer X_i cuyo valor supere al de $N/2$.

Ejemplo 3.14. Obtener la mediana de la siguiente distribución de frecuencias, en la que se indica el grado de satisfacción (valoración de 0 a 100) de 100 clientes que han utilizado el servicio de una oficina bancaria:

x_i	n_i	N_i
10	2	2
11	3	5
12	5	10
13	6	16
24	4	20
25	9	29
36	5	34
50	54	88
58	4	92
99	8	100
Suma	100	

La columna N_i es la columna de frecuencias absolutas acumuladas; obtenemos el valor $N/2 = 50$; el primer valor que lo supera en la columna N_i es el 88; que corresponde a $x_i = 50$; la mediana es en consecuencia el valor 50.

Podríamos decir, en consecuencia, que mas de la mitad de los clientes han dado puntuaciones iguales o superiores a 50, lo que es complementario de otro hecho relevante y a tener en cuenta en la interpretación de los datos, como sería que la media aritmética de las puntuaciones otorgadas sólo es de 44,16.

Ejemplo 3.15. Obtener la mediana de la siguiente distribución de frecuencias:

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

El valor de $N/2 = 100$; en este caso existe un valor de N_i cuya frecuencia acumulada es 100; en $x_i = 12$, N_i iguala a $N/2$; acudimos al primer supuesto anterior y consideramos como mediana la media aritmética del valor x_i y del siguiente x_{i+1} ($x_i = 12$ y $x_{i+1} = 16$), o sea:

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

En este caso, diríamos que la mitad de la distribución está por debajo del valor 14 y la otra mitad por encima de este valor.

C) Cálculo de la mediana en las distribuciones de tipo III o agrupadas por intervalos

Si la variable está agrupada en intervalos la mediana se calcula en parecida forma que el apartado anterior, es decir:

1. Si existe N_i que es igual a $N/2$, la mediana, por convenio, es el límite superior del intervalo mediano o intervalo en el que $N_i = N/2$.

2. Si no existe un $N_i = N/2$, la mediana está en el siguiente intervalo, es decir en el primer intervalo cuya N_i supera a $N/2$; diremos que dicho intervalo es el intervalo mediano.

En algunas ocasiones puede ser conveniente adjudicar a la mediana un valor concreto dentro del intervalo; en este caso, debemos estimar este valor con diversos procedimientos que explicamos en el siguiente ejemplo:

Ejemplo 3.16. Supongamos la siguiente distribución agrupada en intervalos sobre las cifras de ventas mensuales en euros de 2.000 establecimientos comerciales:

$L_{i-1} - L_i$	n_i	N_i
(0 - 8.000 €)	50	50
(8.001 - 10.000 €)	100	150
(10.001 - 11.000 €)	75	225
(11.001 - 12.000 €)	550	775
(12.001 - 13.000 €)	950	1.725
(13.001 - 14.000 €)	75	1.800
(14.001 - 15.000 €)	25	1.825
(15.001 - 16.000 €)	25	1.850
(16.001 - 24.000 €)	150	2.000
	2.000	

Operando con la columna de frecuencias absolutas acumuladas N_i se observa que el primer valor de N_i que supera a $N/2 = 1.000$, se corresponde con el intervalo (12.001-13.000 euros), intervalo que acumula 1.725 comercios y cuyas ventas mensuales se sitúan en torno a los 12.500 euros.

Dado que los datos se presentan en intervalos puede indicarse simplemente que el intervalo (12.001 – 13.000 €) es el intervalo mediano de la distribución y que en consecuencia hay tantos comercios cuyas ventas son inferiores a 12.001 euros mensuales, como aquellos que venden más de 13.000 euros.

Podría considerarse, no obstante, que conviene una mayor precisión de la mediana y que debe situarse ésta en un punto exacto dentro del intervalo mediano; en este caso podrían adoptarse tres alternativas:

1. Obtener la marca de clase y situar en la misma la mediana; en este caso diríamos que la mediana está en los 12.500 euros.
2. Aproximar la mediana mediante una regla de 3 que busque la proporcionalidad; para efectuar esta aproximación operamos de la siguiente forma:

   Suponemos que los 950 establecimientos comerciales que están en el intervalo (12.001-13.000 euros) se distribuyen uniformemente a lo largo del intervalo; sabemos que 775 establecimientos venden menos de 12.000 euros (N_i del intervalo anterior), luego faltan 225 establecimientos en llegar a las 1.000 que nos interesan ($N/2 = 1000$); haciendo una sencilla regla de tres en la que se tiene en cuenta la amplitud del intervalo, diremos:

   Si 950 agencias se distribuyen en 1.000 (amplitud del intervalo) 225 agencias ocuparán «x».

   De dónde:

$$x = \frac{225 \cdot 1.000}{950} = 237$$

Luego el valor exacto de la mediana será de 12.238 euros ($L_{i-1}+x$ = 12.001 + 237).

Es evidente que este método resulta más exacto que el de las marcas de clase y es el que debe utilizarse.

Existen otros convenios más o menos aplicados por los estadísticos; uno bastante habitual en los libros de estadística es el que resulta de la aplicación de la siguiente formulación:

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i \qquad [3.6.1]$$

c_i es la amplitud del intervalo; en el Ejemplo 3.16, dónde c_i = 1.000, tendríamos:

$$M_e = 12.001 + \frac{1.000 - 775}{950} \cdot 1.000 = 12.238$$

Repetimos que se trata de convenios más o menos aceptados, pero en todo caso discutibles; de forma que cualquier a de las 3 opciones manejadas en este supuesto (el intervalo mediano 12.001-13000, sin especificar ningún punto concreto, 12.500 ó 12.238) pueden considerarse como válidas.

Una vez más indicamos que la estadística es un instrumento para el gestor; la importancia de la exactitud de este último dato viene dada por los objetivos para los que sirva el mismo.

Ejemplo 3.17. Tenemos la siguiente distribución:

L_{i-1} - L_i	n_i	N_i
(0 - 8]	5	5
(8 - 10]	10	15
(10 - 11]	7	22
(11 - 12]	128	150
(12 - 13]	20	170
(13 - 14]	117	287
(14 - 15]	5	292
(15 - 16]	3	295
(16 - 24]	5	300
	300	

En este caso el valor de N_i que iguala a $N/2 = 150$ se corresponde con el intervalo (11-12); por el convenio más aceptado entre los estadísticos, consideraremos la mediana como el límite superior de dicho intervalo, es decir, el 12.

Nótese la exactitud de este convenio, ya que 150 observaciones (la mitad del total de las 300 disponibles) son inferiores a 12 y la otra mitad superiores a esta cantidad.

Las principales **ventajas** de la mediana son las siguientes:

— Es la medida *más representativa en el caso de las variables cualitativas o atributos* (recordemos que cuando hablamos de nacionalidades, por ejemplo, no podemos calcular la media aritmética, ni otras medidas de posición).
— Su cálculo es sencillo.
— Tiene una fácil interpretación.
— *No es sensible a los valores extremos de la distribución.*

El principal **inconveniente** es que en su determinación no se tienen en cuenta todos los valores de la variable; este inconveniente puede constituir incluso una ventaja, ya que es posible su cálculo cuando no se conocen los valores extremos pero sí su frecuencia.

En el ejemplo 3.18 se refleja esta ventaja.

Ejemplo 3.18. Determinar la mediana de la siguiente distribución de las ventas anuales a una muestra de 100 clientes:

Ventas anuales en euros	n_i	N_i
Menos de 1.000 €	5	5
De 1.001 a 2.000 €	15	20
De 2.001 a 5.000 €	30	50
De 5.001 a 10.000 €	10	60
Más de 10.000 €	40	100
	100	

En este caso ($N/2 = 50$), podemos situar, sin problemas, la mediana en el valor de 5.000 € (límite superior del intervalo de 2.001 a 5.000 €), valor que nos indica que hay 50 clientes a los que vendemos más de 5.000 € y otros 50 a los que vendemos menos de 5.000 €.

Tendríamos algún problema, sin embargo, sí tratásemos de calcular otros estadísticos como la media aritmética, la media geométrica o la media armónica, a no ser que realicemos unas hipótesis sobre la marca de clase de los dos intervalos abiertos (el primero y el último); además, como ya hemos indicado, el resultado de cualquiera de estos estadísticos variaría sí cambiásemos la hipótesis sobre la marca de clase asignada.

3.7. LA MODA

La Moda es el valor de la variable que se repite más veces; suele designarse por M_o y ***se define como el valor de la variable que presenta mayor frecuencia absoluta.***

En el caso de que existan varios valores en esta situación se dice que la distribución es *bimodal, trimodal o multimodal.*

Se diferencia entre moda o modas absolutas y moda o modas relativas.

Se dice que un valor de una variable constituye una *moda relativa cuando su frecuencia absoluta no es superada por la de sus valores contiguos.*

Obtención de la moda en los tres tipos de distribuciones

A) *En las distribuciones de frecuencias de tipo I no tiene sentido hablar de moda, ya que las frecuencias absolutas son todas unitarias.*

B) *Para obtener la moda en las distribuciones de tipo II basta con observar la columna de las n_i; veamos el siguiente ejemplo:*

Ejemplo 3.19. Determinar la moda de la siguiente distribución de frecuencias sobre una muestra de 100 visitantes a una feria de muestras:

Nacionalidad	N.º de visitantes (n_i)
Española	40
Francesa	15
Norteamericana	30
Otras nacionalidades	15
Total	100

La moda absoluta, es decir, la categoría que más visitantes aporta, corresponde a la nacionalidad española (40).

En este caso existiría también una moda relativa para el valor 30 (nacionalidad norteamericana), ya que cumple la condición de superar la frecuencia de los valores contiguos (15).

Sin embargo, la moda relativa sólo tiene interés cuando tratamos con variables «ordenables» y previamente ordenadas (variables cuantitativas o variables cualitativas ordenables); realmente, en el ejemplo anterior, bastaría con variar el orden de las nacionalidades (orden que realmente no aporta nada y que probablemente ha salido por casualidad) para que ya no existiese la moda relativa señalada; sería, por ejemplo, el siguiente supuesto de ordenación:

Nacionalidad	N.º de visitantes (n_i)
Española	40
Norteamericana	30
Francesa	15
Otras nacionalidades	15
Total	100

Veamos un ejemplo con variables ordenables, que puede extenderse, sin problemas, a variables cuantitativas.

Ejemplo 3.20. Obtener la moda de la siguiente distribución de 100 opiniones de clientes sobre la calidad de un determinado servicio:

Opinión manifestada	N.º de clientes (n_i)
Muy buena	5
Buena	40
Normal	15
Mala	30
Muy mala	10
Total	100

La moda absoluta está representada por la categoría «buena», pudiendo afirmarse que hay una moda relativa en la categoría «mala», cuya frecuencia supera a las categorías de «normal» y «muy mala».

C) *En las distribuciones de tipo III, es decir, cuando los datos están agrupados en clases o intervalos pueden darse dos supuestos:*

a) Que los intervalos sean de igual amplitud

En este caso la moda absoluta se situará en el intervalo que presente mayor frecuencia absoluta y las modas relativas en el intervalo o intervalos que superen la frecuencia absoluta de los intervalos contiguos.

Para determinar el valor exacto de la moda y al igual que hemos indicado con la mediana podríamos optar por considerar la marca de clase del intervalo o proceder a prorratear un valor dentro del intervalo.

Para este prorrateo el convenio más aceptado entre los estadísticos es el que resulta de aplicar la siguiente expresión:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i \qquad [3.7.1]$$

Ejemplo 3.21. Determinar la moda de la siguiente distribución de frecuencias:

$L_{i-1} - L_i$	(n_i)
[0 - 5)	2.000
[5 - 10)	3.500
[10 - 15)	500
	6.000

El intervalo modal es el intervalo [5 – 10), con una frecuencia absoluta de 3.500 superior a la de los otros 2 intervalos; para buscar el punto exacto aplicamos la expresión [3.7.1] en la siguiente forma:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i = 5 + \frac{500}{2.000 + 500} \cdot 5 = 6$$

Indicándonos que la moda está en el valor 6; bajo determinados supuestos este dato parece más preciso que sí se señalase la marca de clase del intervalo modal (7,5)[4].

Siendo este el criterio más aceptado, no sería un error estadístico especialmente grave, situar la moda en la marca de clase del intervalo modal.

b) Que los intervalos tengan distinta amplitud

En este caso, para determinar el intervalo modal es necesario obtener un ***ratio de densidad de frecuencia*** (frecuencia absoluta dividida por amplitud del intervalo); *el intervalo con mayor valor en este ratio constituirá el intervalo modal.*

Este ratio es imprescindible para evitar situaciones de desequilibrio que pudieran desvirtuar el propio concepto, interés e interpretación de la moda.

Veamos un ejemplo de intervalos de distinta amplitud.

[4] Esta fórmula se basa en considerar una cierta uniformidad en la distribución de valores dentro de cada intervalo; si la frecuencia del intervalo inferior, 2.000, es bastante superior a la frecuencia del último intervalo, 500, podría pensarse que los valores de las observaciones van disminuyendo, de forma que dentro del intervalo modal se acumularán más en los valores próximos al extremo inferior (5) que al extremo superior (10); la expresión utilizada permite reflejar o aproximar este hecho.

Ejemplo 3.22. *Determinar la moda de la siguiente distribución de frecuencias*

L_{i-1} - L_i	(n_i)
(0 - 25)	125
[25 - 30)	50
[30 - 35)	35
[40 - 45)	25
[45 - 55)	80
	315

Incurriríamos en una imprecisión estadística si considerásemos que el intervalo modal es el (0–25), ya que aunque es verdad que en este intervalo se han acumulado un mayor número de observaciones (125), también lo es que tiene una amplitud (25) superior al resto de los intervalos (5 ó 10); si construimos la columna de densidad de frecuencias n_i/c_i, dividiendo las frecuencias por la amplitud de los intervalos, tenemos:

L_{i-1} - L_i	n_i	c_i	$h_i = n_i / c_i$
(0 - 25)	125	25	5
[25 - 30)	50	5	10
[30 - 35)	35	5	7
[40 - 45)	25	5	5
[45 - 55)	80	10	8
	315		

De dónde se deduce que el intervalo modal, es decir el que concentra una mayor densidad de observaciones, es el intervalo [25 – 30).

Para hallar el punto modal exacto lo más habitual es operar con la siguiente expresión

$$M_o = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot c_i \qquad [3.7.2]$$

Equivalente a la expresión [3.7.1.], pero, en la que en vez de considerar las frecuencias n_i, se tiene en cuenta la densidad de frecuencias de cada intervalo (el cociente $h_i = n_i / c_i$, donde c_i representa la amplitud del intervalo).

Aplicando la expresión [3.7.2] a los datos del ejemplo, se tiene:

$$M_o = L_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot c_i = 25 + \frac{7}{5+7} \cdot 5 \approx 27{,}9$$

Las principales **ventajas** de la moda son las siguientes:

— Es un estadístico que *puede obtenerse en todas las distribuciones* (tanto en variables cuantitativas como en las cualitativas), ya que siempre es posible determinar el valor, la categoría o la modalidad que más se repite.
— Su cálculo es sencillo.
— Tiene una *fácil interpretación estadística*, ya que nos da el valor o modalidad que más se repite.

Al igual que en la mediana como principal **inconveniente** hay que señalar que en su determinación no intervienen todos los valores de la distribución, centrándonos sólo en la mayor frecuencia absoluta de un determinado valor de la variable o de la modalidad de los atributos.

3.8. MEDIDAS DE POSICIÓN NO CENTRALES: LOS CUANTILES

Los cuantiles son los valores de la variable que divíden una distribución de frecuencias en partes iguales.

Los más habituales son:

— Los **cuartiles,** que son tres valores que dividen a la serie de datos en cuatro partes iguales. La mediana coincide con el segundo cuartil y divide la distribución en dos partes iguales.
— Los **quintiles**, que son cuatro valores que dividen la distribución en 5 partes iguales.
— Los **deciles,** que son nueve valores que dividen la distribución en diez partes.
— Los **percentiles,** que son 99 valores que dividen la distribución en cien partes iguales.

Considerando N el número de datos de la distribución, o frecuencia absoluta acumulada, con carácter general los cuantiles se obtienen con la expresión $\frac{rN}{q}$, en la que r indica el cuantil correspondiente ($r = 1$, primer cuantil, $r = 2$, segundo cuantil, etc.) y q el número de intervalos con iguales frecuencias en los que se pretende dividir la distribución (si $q = 4$ hablamos de cuartiles, si $q = 10$ de percentiles, etc.).

Si $q = 4$ y $r = 1, 2, 3$ obtenemos los cuartiles.
Si $q = 10$ y $r = 1, 2, \ldots, 9$ obtenemos los deciles.
Si $q = 100$ y $r = 1, 2, \ldots, 99$ obtenemos los percentiles.

En el caso de los tres cuartiles la expresión a utilizar sería:

$\frac{1N}{4}$ para el primer cuartil Q_1

$\frac{2N}{4}$ para el primer cuartil Q_2

$\frac{3N}{4}$ para el primer cuartil Q_3

Para distribuciones agrupadas en intervalos utilizamos la siguiente expresión:

$$Q_{\frac{r}{q}} = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i} \cdot c_i \qquad [3.8.1]$$

El cálculo de los cuantiles se efectúa de manera similar a lo explicado para la mediana.

Ejemplo 3.23. Obtener los cuantiles de la distribución de frecuencias:

x_i	n_i
12	20
45	30
55	20
60	10
82	10
90	10
Total	100

Obtenemos la columna de frecuencias absolutas acumuladas N_i:

x_i	n_i	N_i
12	20	20
45	30	50
55	20	70
60	10	80
82	10	90
90	10	100
Total	100	

Y procedemos de la siguiente forma:

— **Obtención de cuartiles:**

Q_1; procedemos a obtener $\frac{1 \cdot N}{4} = \frac{100}{4} = 25$

El primer valor que supera a 25 en la columna de N_i es la frecuencia acumulada 50, que corresponde al valor de la variable $x_i = 45$; luego $Q_1 = 45$.

Q_2; 200/4 = 50; este valor coincide con la frecuencia absoluta acumulada de $x_i = 45$; al igual que se explicó en la mediana, cuando existe esta coincidencia, por convenio generalmente aceptado entre los estadísticos, se toma como cuartil la media aritmética de x_i y del siguiente x_{i+1}; en nuestro caso (45 + 55)/2 = 50

Q_3; 300/4 = 75; el primer valor de N_i que lo supera es 80, que corresponde al valor de la variable $x_i = 60$.

— **Obtención de Deciles:**

D_1; $\frac{1 \cdot N}{10} = \frac{100}{10} = 10$; el primer valor de N_i que supera a 10 corresponde al valor de la variable $x_i = 12$.

D_7; 700/10 = 70; el valor 70 coincide con la frecuencia absoluta acumulada de $x_i = 55$; por convenio, se toma como séptimo decil la media aritmética de x_i y del siguiente x_{i+1}; en nuestro caso (55 + 60)/2 = 57,5.

— **Obtención de Percentiles:**

P_{30}; 3000/100 = 30; el primer valor de N_i que lo supera corresponde al valor de la variable $x_i = 45$.

El alumno puede practicar calculando deciles y percentiles en ésta o en otras distribuciones, si bien debe tener en cuenta *la aplicación e interpretación lógica de los resultados*; así, por ejemplo, sí suponemos que la variable x_i del ejemplo 3.23 corresponde a algo indivisible, como sería el número de animales de una granja o número de aviones que llegan diariamente a un aeropuerto, puede no tener mucho sentido obtener el séptimo decil como media de 2 valores enteros e indicar en consecuencia que el 70% de las granjas tienen 57,5 animales o que el 70% de los días llegan 57,5 aviones al aeropuerto; la mediana, los cuartiles y en general todos los estadísticos tienen un valor instrumental (nos ayudan a sintetizar, analizar o interpretar una información más amplia) y no un valor matemático que puede no tener sentido lógico; en este caso concreto parecería más lógico y, en consecuencia, más deseable, indicar que el 70% de los días llegan menos de 57 aviones.

Otra observación a tener en cuenta es que algunos programas informáticos no utilizan los mismos criterios o algoritmos indicados con anterioridad; en concreto, la hoja de cálculo Excel de Microsoft considera a todas las distribuciones como si fueran continuas y sitúa los cuartiles no en el valor de la variable cuya frecuencia absoluta acumulada supera al establecido por el cuartil, sino en un punto intermedio que obtiene mediante un algoritmo particular; dado que muchos alumnos pueden utilizar este programa para efectuar o comprobar los cálculos, explicamos a continuación los resultados que da Excel con un sencillo ejemplo.

Ejemplo 3.24. Disponemos de la siguiente información sobre el gasto de 4 clientes en una cafetería:

	Gasto en euros en una cafetería (X_i)
Cliente 1	1
Cliente 2	8
Cliente 3	9
Cliente 4	85

Las frecuencias absolutas y las frecuencias acumuladas serían:

Gasto en euros en una cafetería (X_i)	n_i	N_i
1	1	1
8	1	2
9	1	3
85	1	4

Cálculo de Cuartiles:

Resultados según el criterio explicado	**Resultados según la Hoja Excel**
$Q_1 : \frac{4}{4} = 1 \Rightarrow x = \frac{1+8}{2} = 4{,}5$ (Media aritmética del valor que iguala $N_i = 1$ y el siguiente)	6,25
$Q_2 : \frac{8}{4} = 2 \Rightarrow x = \frac{8+9}{2} = 8{,}5$ (Media aritmética del valor que iguala $N_i = 2$ y el siguiente)	8,5
$Q_1 : \frac{12}{4} = 3 \Rightarrow x = \frac{9+85}{2} = 47$ (Media aritmética del valor que iguala $N_i = 3$ y el siguiente)	28

Lo que hace Excel es aplicar un criterio similar al explicado en el Ejemplo 3.21 para obtener un punto o valor concreto dentro de los intervalos; es decir, prorratea la «masa» de la distribución entre los valores de la variable, de forma que no se queda con el valor exacto de esta, sino con un valor intermedio entre los dos valores afectados; por ejemplo,

Para el cálculo del primer cuartil:

— Considera que entre los dos valores afectados hay una «masa» de 7 euros (8-1).
— Divide esta masa entre los clientes de la distribución 7/4 = 1,75 euros.
— Resta este valor del límite superior del cuartil (8-1,75 =6,25).

Para el cálculo del tercer cuartil:

— Considera que entre los dos valores afectados hay una «masa» de 76 euros (85-9).
— Divide esta masa entre los clientes de la distribución 76/4 = 19 euros.
— Suma este valor al límite inferior del cuartil (19 + 9 = 28 euros).

Dado el carácter instrumental de la Estadística Descriptiva cualquiera de estos dos criterios podría ser válido, aunque según los datos disponibles y según los fines perseguidos por el analista, podría ser más aconsejable uno u otro criterio; sí el ejemplo 3.24 estuviese referido a una muestra más amplia y estadísticamente representativa podría tener sentido el resultado aportado de Excel, ya que una vez inferidos los resultados al conjunto de la clientela probablemente le aporte al analista una información más precisa la afirmación de Excel que indica que «el 25% de la clientela gasta menos de 6,25 euros» que la aportada por el criterio tradicional; sí por el contrario volviésemos al ejemplo anterior y hablásemos de aviones parece más conveniente indicar que «un 25% de los días entran 4 aviones en un aeropuerto, prescindiendo en este caso de las cifras decimales».

3.9. MEDIDAS DE POSICIÓN ROBUSTAS

Como ya se ha comentado, la media aritmética se ve fuertemente afectada por los valores extremos. Las medidas robustas de tendencia central tratar de paliar los problemas de estimación asociados a distribuciones anómalas, siendo estadísticos que funcionan bien para varios tipos distintos de distribuciones teóricas, aunque pueden no ser el mejor estimador para ningún tipo concreto de distribución siendo, por tanto, el mejor compromiso. Estudiamos aquí las más significativas.

3.9.1. La media k-recortada

La media recortada o acotada al k por ciento es la media de los datos que quedan después de eliminar el k por ciento de los datos más grandes y el k por ciento de los

datos más pequeños. A la media recortada al 25% se la denomina **centrimedia**. Obviamente, la media recortada al 0% será igual a la media aritmética.

La media k-recortada elimina el efecto de los valores extremos en el caso en que la proporción de los mismos en cada extremo sea inferior a k; de este modo, puede considerarse un remedio muy adecuado para la «falta de robustez» de la media.

3.9.2. La media k-winsorizada

Se opera de forma análoga a las medias k-recortadas, sólo que en lugar de prescindir de los k por ciento datos más grandes y más pequeños, se sustituyen por el valor mayor y menor de los datos restantes.

Así, *si partimos de los siguientes datos:*

$$3,\ 4,\ 4,\ 5,\ 5,\ 6,\ 7,\ 8,\ 9,\ 11$$

La media recortada a nivel 2 implicaría eliminar las dos puntuaciones mayores y las 2 menores, es decir, calcularíamos la media de los siguientes datos.

$$4,\ 5,\ 5,\ 6,\ 7,\ 8$$

En la media winsorizada a nivel 2, los datos 3 y 4 (los dos menores) y el 9 y 11 (los dos mayores) se sustituyen por 4 y 8 respectivamente. Es decir, calcularemos la media de los siguientes datos.

$$4,\ 4,\ 4,\ 5,\ 5,\ 6,\ 7,\ 8,\ 8,\ 8$$

3.9.3. La trimedia

Es un índice de tendencia central que consiste en calcular una media aritmética ponderada de tres medidas, la Mediana (con peso doble) y el primer y tercer cuartel, es decir:

$$\text{Trimedia} = \frac{Q_1 + 2Q_2 + Q_3}{4}$$

Ejemplo 3.25. Calcular con los datos siguientes la media aritmética, la trimedia y las medias recortadas y winsorizadas al 10%. Comentar los resultados obtenidos.

$$2{,}1 \quad 3{,}2 \quad 3{,}5 \quad 4{,}1 \quad 4{,}4 \quad 4{,}9 \quad 5{,}2 \quad 5{,}8 \quad 6{,}9 \quad 73$$

$$\text{Media aritmética: } \bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 11{,}31$$

Media recortada al 10%:

Dado que el 10% de los datos es igual a: $0,1 \cdot 10 = 1$, debemos calcular la media aritmética de las ocho observaciones centrales, es decir:

$$3,2 \quad 3,5 \quad 4,1 \quad 4,4 \quad 4,9 \quad 5,2 \quad 5,8 \quad 6,9$$

$$\text{Media 1-recortada: } \frac{\sum_{i=1}^{8} x_i}{8} = 4,75$$

Media winsorizada al 10%:

La media 1-winsorizada se calcula sustituyendo 2,1 por 3,2 y 73 por 6,9, es decir, calcularemos la media aritmética de los siguientes datos:

$$3,2 \quad 3,2 \quad 3,5 \quad 4,1 \quad 4,4 \quad 4,9 \quad 5,2 \quad 5,8 \quad 6,9 \quad 6,9$$

$$\text{Media 1-winsorizada: } \frac{\sum_{i=1}^{10} x_i}{10} = 4,81$$

$$\text{Trimedia: } \frac{Q_1 + 2Q_2 + Q_3}{4} = \frac{3,5 + 2 \cdot 4,65 + 5,8}{4} = 4,65$$

Como vemos, la existencia de un valor extremo, desplaza la media a un valor de 11,31, que parece no estar acorde con la verdadera tendencia central de la serie. La influencia de este valor extremo es eliminada por las medidas robustas, que sitúan la tendencia e valores comprendidos entre 4,5 y 5.

3.10. MOMENTOS DE UNA DISTRIBUCIÓN UNIDIMENSIONAL DE FRECUENCIAS

Los momentos son medidas que caracterizan a una distribución de frecuencias y que tienen como principal utilidad su condición de operadores para el cálculo simplificado de las medidas de posición, dispersión o forma de una distribución; también tienen una importante utilidad para efectuar las regresiones estadísticas que abordaremos más adelante.

Existen dos clases de momentos:

— **Los momentos respecto al origen**, que se representan con a_h y se obtienen mediante la siguiente formulación:

$$a_h = \sum_{i=1}^{r} x_i^h \frac{n_i}{N}$$

Algunos ejemplos de estos momentos son:

— Si $h = 0$, se tiene $a_0 = 1$.
— Si $h = 1$, se tiene $a_1 = \bar{x}$, es decir la media aritmética de x_i.
— Por lo tanto a_h es la media aritmética de los valores observados elevados a la potencia h.

— **Los momentos centrales o respecto a la media**, que se representan con m_h y se obtienen mediante la siguiente formulación:

$$m_h = \sum_{i=1}^{r} \left(x_i - \bar{x}\right)^h \frac{n_i}{N}$$

Siendo $\bar{x}$ la media aritmética de la variable estadística.

El momento de orden 1 respecto a la media no tiene ningún interés, ya que siempre es cero:

$$m_1 = \sum_{i=1}^{r} \left(x_i - \bar{x}\right)^h \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{r} x_i n_i - \frac{1}{N} \bar{x} \sum_{i=1}^{r} n_i = \bar{x} - \bar{x} = 0$$

Al momento de orden 2 respecto a la media m_2; definido como

$$m_2 = \sum_{i=1}^{r} \left(x_i - \bar{x}\right)^2 \frac{n_i}{N}$$

se le denomina **varianza** y, cómo veremos en el capítulo siguiente, constituye la medida de dispersión más utilizada en Estadística.

Puede demostrarse, utilizando el desarrollo del Binomio de Newton, que los momentos respecto a la media están relacionados con los momentos respecto al origen $m_h = f(a_{h-j})$.

Por ejemplo,

$$m_2 = a_2 - a_1^2 = a_2 - \bar{x}^2$$

Es decir, la denominada **varianza,** o momento de segundo orden respecto a la media, es igual al momento de segundo orden respecto al origen menos la media aritmética elevada al cuadrado.

3.11. LAS FUNCIONES ESTADÍSTICAS EN HOJA DE CÁLCULO EXCEL Y EN SPSS

3.11.1. Las funciones estadísticas en Excel

Las funciones que incorpora la hoja de cálculo son herramientas especiales que realizan cálculos complejos, lo que nos permite llevar a cabo acciones y ejecutar operaciones que nos devuelven valores de forma automática. Excel 95 dispone de una amplia gama de funciones. Los grupos de funciones son los siguientes:

— Financieras. Ejecutan funciones sobre valores. Desarrollan cálculos sobre amortizaciones, préstamos, intereses, etc.

— *Fecha y hora.* Devuelven información cronológica.

— *Matemáticas y trigonométricas.* Es una lista de funciones de cálculo numérico, operaciones con matrices, trigonometría, etc.

— *Estadísticas.* Funciones de cálculo probabilístico, tendencias, desviaciones, distribuciones, estimaciones, etc.

— *Búsqueda y referencia.* Devuelven información sobre celdas, rangos de celdas y posiciones y trabajan con éstas.

— *Base de Datos.* Extraen información, manipulan y operan sobre listas.

— *Texto.* Convierten a datos tipo texto cualquier tipo de datos y operan con los caracteres de las cadenas de texto.

— *Lógicas.* Devuelven valores lógicos.

— *Información.* Ofrecen información de Microsof Excel.

Al igual que para otras tareas, Excel dispone de ayuda para todas las funciones.

Las funciones se pueden insertar en las celdas de las hojas de cálculo con el Asistente de funciones. Esta forma de introducir las funciones se realiza en dos pasos. En el primer paso se elige la función entre los múltiples grupos permitidos (estadísticas, financieras, etc.); en el segundo paso se rellenan los argumentos de la función. En las siguientes dos ilustraciones se ven los dos pasos citados tomando como ejemplo la función estadística contar.

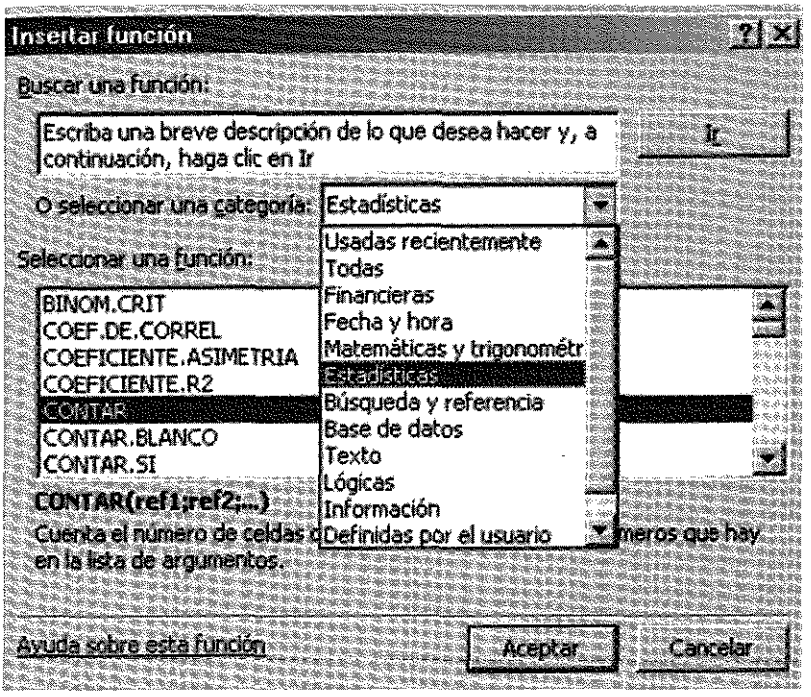


Imagen 1. Insertar función.

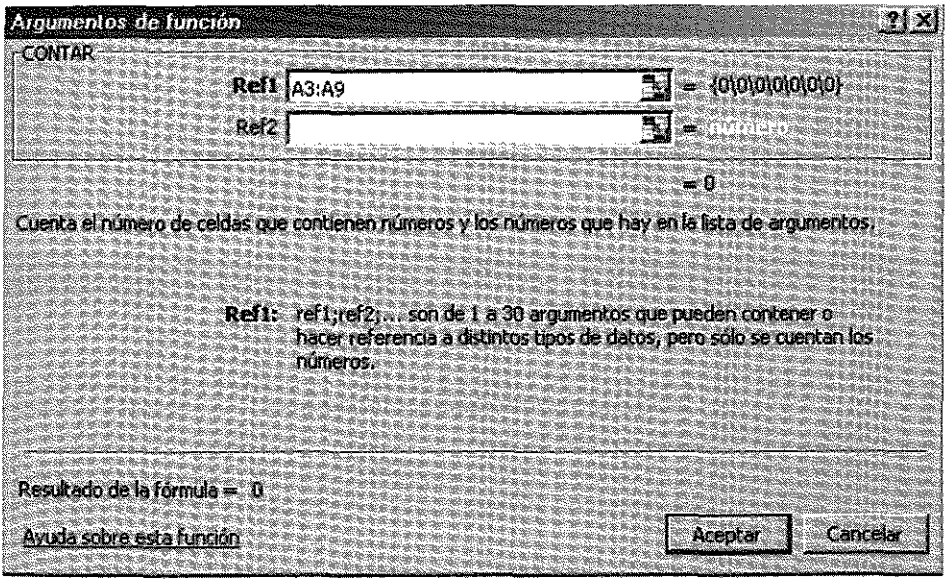


Imagen 2. Indicar los argumentos.

En la siguiente Tabla se muestran las funciones estadísticas para obtener las medidas de posición estudiadas en este capítulo y su formulación en Excel:

Funciones estadísticas más empleadas en Excel

ESTADÍSTICA	FÓRMULACIÓN
Media	=PROMEDIO(rango)
Media geométrica	=MEDIA.GEOM(rango)
Media armónica	= MEDIA.ARMO(rango)
Mediana	=MEDIANA(rango)
Moda	=MODA(rango)
Suma	=SUMA(rango)
Cuenta	=CONTAR(rango)
K-ésimo mayor	=K.ESIMO.MAYOR(rango;k)
K-ésimo menor	=K.ESIMO.MENOR(rango;k)

= SI(prueba_lógica;valor_si_verdadero;valor_si_falso)
= FRECUENCIA(matriz_datos;matriz_clases)
= CONTAR.SI(rango;criterio)
= SUMAR.SI(rango;criterio;rango_suma)

Entre estas funciones es importante destacar la función condicional SI que es utilizada para guardar un determinado valor en caso de que un contenido cumpla una determinada condición.

Por ejemplo en la fórmula: = SI (*A1*<220;50;10)

Excel devuelve el valor 50 si el valor de la celda *A1* es menor que 220; de lo contrario, devuelve 10.

La función condicional SI puede anidar otras funciones. Por ejemplo, la fórmula:

= SI (SUMA (*B1:B10*)>0; SUMA(*B1:B10*);0)

Excel devuelve la suma de *B1* hasta *B10* si ésta es mayor que 0; de lo contrario, devuelve el valor 0.

También se pueden utilizar argumentos de texto en las funciones SI:

=SI(F4>5;«Aprobado»;«Suspenso»)

Excel examina si el valor que contiene la celda F4 es mayor que 5 y si es así, la función devuelve el texto *Aprobado*; si la celda F4 es menor que o igual a 5, la función devolverá el texto *Suspenso*.

Incluso se pueden utilizar argumentos de texto en las funciones SI para no devolver nada, en lugar de 0:

=SI(SUMA(A1:A10)>0;SUMA(A1:A10);»»)

El argumento prueba lógica de una función SI puede verificar el contenido de un texto. Por ejemplo, la fórmula:

=SI(A1=«Valladolid»;160;250)

Excel devuelve el valor 160 si la celda A1 contiene la cadena de texto Valladolid y 250 si ésta contiene cualquier otro texto o valor. Recordar que la coincidencia entre los dos textos debe ser exacta.

La función SI, puede anidarse con los operadores lógicos Y, O y NO. Estos pueden ser utilizados junto a los operadores lógicos simples =, >, <, >=, <=, y <>.

Recordar que las funciones lógicas Y, O pueden tener hasta 30 argumentos lógicos, en tanto que la función NO sólo tiene un argumento.

Ejemplos del anidamiento de la función SI con los operadores lógicos descritos serían los siguientes:

=SI(Y(G4<5;F4>80%);«Bueno»;«Malo»)
=SI(O(G4<5;F4>80%);«Bueno»;«Malo»)
=SI(NO(A1=2);«Hacer»;«No hacer»

La función SI también puede emplearse para crear una jerarquía de pruebas. Por ejemplo, la fórmula siguiente:

=SI(A1=100;«Siempre»;SI(Y(A1>80;A1<100);«Normalmente»;SI(Y(A1>=60; A1<80);«A veces»;«¿A quién le importa?»)))

Excel utiliza tres funciones SI separadas. En este caso, si el valor de la celda A1 es 100, devuelve la cadena de texto *Siempre*; si el valor de la celda A1 está entre 80 y 100 (es decir, desde 81 hasta 99), devolverá la cadena de texto *normalmente*; pero, si el valor de la celda A1 está entre 60 y 80 (desde 60 hasta 79), devolverá la cadena de texto *A veces*; por último, si ninguna de estas condiciones es cierta, devolverá la cadena de texto *¿A quién le importa?*.

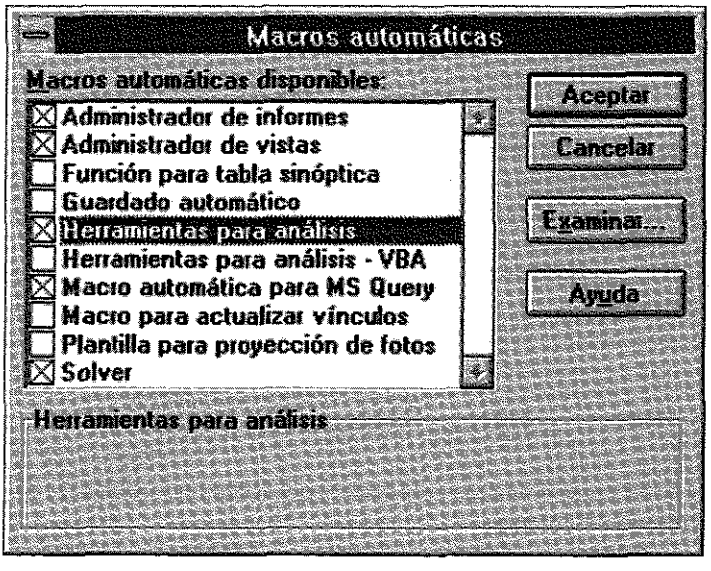
No obstante, hay que tener presente que EXCEL permite anidar hasta siete funciones SI, siempre que la sentencia no sobrepase el límite de 255 caracteres escritos en una sola celda.

LA MACRO HERRAMIENTAS DE EXCEL PARA ANÁLISIS ESTADÍSTICO

La hoja de cálculo Excel dispone también de una opción MACRO que facilita la rápida obtención de múltiples estadísticos descriptivos. Esta opción debe activarse adecuadamente, ya que es posible que según el tipo de instalación que se haya efectuado, no esté operativa.

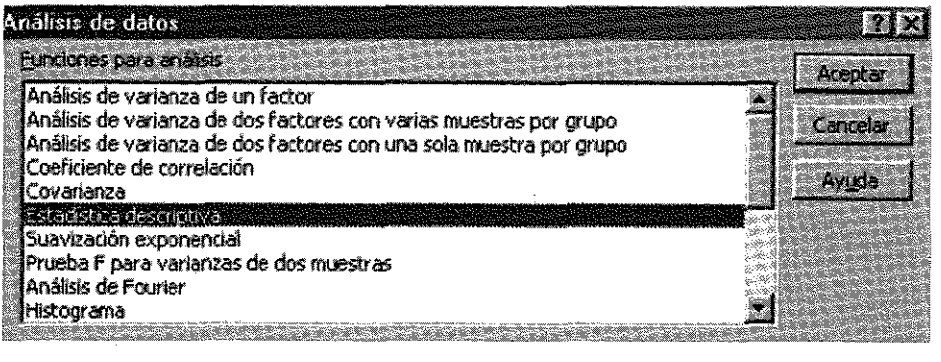
La forma de ver si la opción está activada es mirar en el menú «*Herramientas*» y comprobar que se encuentra el texto «*Análisis de Datos*» En caso de no encontrar esta

opción activada tendremos que cargar la macro «*Herramientas para análisis*» . Esta opción la encontramos en el apartado «*Macros automáticas*» (versiones más antiguas de Excel» en el apartado «Complementos», que están ambos en el menú de «*Herramientas*», con ello accedemos a la siguiente pantalla.



Los principales desarrollos estadísticos que contiene la hoja de cálculo Excel se encuentran en la macro de *Herramientas para análisis o Herramientas para análisis* -VBA. Una vez instalada se accede a ella desde el menú «*Herramientas*», apartado «*Análisis de datos*».

Las posibilidades de esta macro son muy amplias superando en bastantes casos el contenido de este libro; En este apartado nos quedaremos con la opción de Estadística Descriptiva, extraída del siguiente menú.

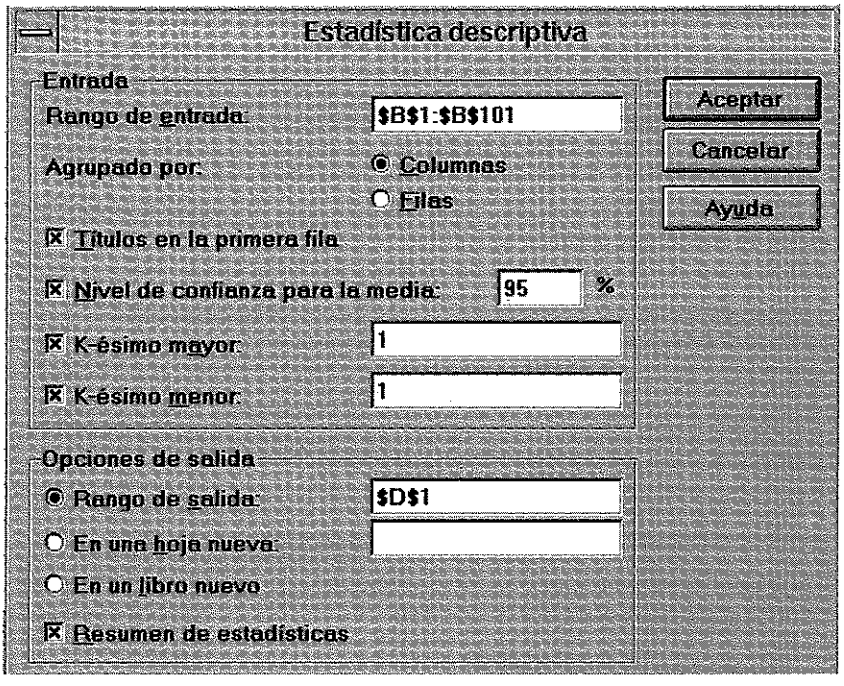


Dicha opción Genera un informe de estadísticas de una sola variable para datos del rango de entrada, y proporciona información acerca de la tendencia central y dispersión de los datos. En concreto, en genera información tanto sobre la Media, la Mediana y la Moda, como sobre los estadísticos de dispersión, concentración y forma que tratamos en el capítulo siguiente (el Error Típico, la Desviación estándar, la Varianza de la muestra, el Coeficiente de Curtosis, el Coeficiente de asimetría, el Rango, el Valor mínimo, el Valor máximo, etc..

Entre los aspectos estudiados hasta ahora también podemos trabajar con las opciones de *Histogramas, Jerarquía y Perceptil,* etc., dejándonos abiertas además otras múltiples posibilidades para trabajar con algunos estadísticos que veremos en capítulos posteriores (Covarianzas, Coeficientes de Correlación, Regresiones, etc.) y con otros instrumentos que pueden interesar a quién profundice en el estudio de esta disciplina.

Veamos dos sencillos ejemplos del funcionamiento de estas opciones.

En el primer ejemplo obtenemos los estadísticos que incluye la opción *Estadística Descriptiva* para la variable GRUPO2. En dicha opción encontramos el siguiente menú:



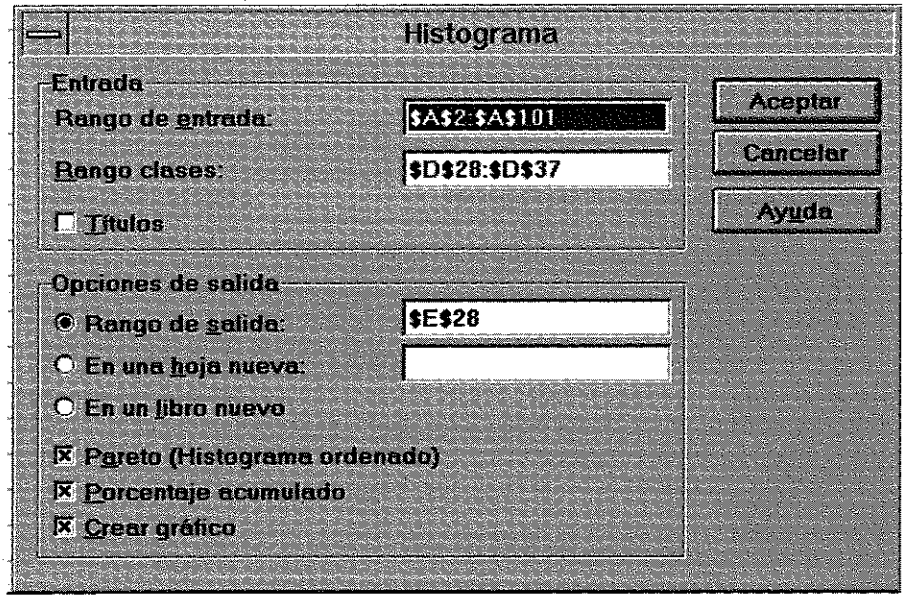
Se selecciona el rango de entrada de los datos (el correspondiente a la variable GRUPO2), señalando que éstos están agrupados por columnas, figuran los títulos en la primera fila y señalamos la celda en donde deseamos que deje los resultados obtenidos.

El resumen de estadísticas que EXCEL calcula para la variable GRUPO2 es el siguiente:

	A	B	C	D	E	F
1	GRUPO1	GRUPO2		*GRUPO2*		
2	10	9				
3	9	10		Media	5,71	
4	1	3		Error típico	0,28472741	
5	6	5		Mediana	6	
6	2	4		Moda	6	
7	1	10		Desviación estándar	2,84727408	
8	8	1		Varianza de la muestra	8,1069697	
9	10	6		Curtosis	-1,13516198	
10	10	6		Coeficiente de asimetría	-0,08587125	
11	4	7		Rango	9	
12	1	8		Mínimo	1	
13	10	4		Máximo	10	
14	9	1		Suma	571	
15	2	7		Cuenta	100	
16	3	5		Mayor (1)	10	
17	5	7		Menor(1)	1	
18	5	2		Nivel de confianza(95,0009	0,55805464	

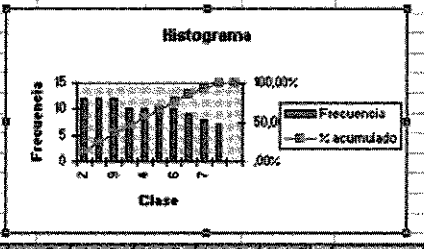
Hoja1 / **Hoja5** / Hoja6 / Hoja7 / Hoja8 / Hoja9 / Hoja1

En el segundo ejemplo realizamos un Histograma de frecuencias para la variable GRUPO1. La ilustración que sigue es la que corresponde a la opción «*Histograma*» de la macro «*Análisis de datos*». En el menú que despliega la opción es necesario especificar el rango de la hoja de cálculo donde están las clases para las cuales se va a elaborar el histograma de frecuencias. En el ejemplo son 10 rangos que van desde el 1 hasta el 10. El menú nos permite activar la opción gráfica.



La salida que nos ofrece la Hoja de cálculo EXCEL es la siguiente:

	D	E	F	G	H	I	J
26	RANGOS PARA HISTOGRAMA						
27							
28		*Clase*	*Frecuencia*	*% acumulado*	*Clase*	*Frecuencia*	*% acumulado*
29	1	1	10	10,00%	2	12	12,00%
30	2	2	12	22,00%	3	12	24,00%
31	3	3	12	34,00%	9	12	36,00%
32	4	4	10	44,00%	1	10	46,00%
33	5	5	10	54,00%	4	10	56,00%
34	6	6	10	64,00%	5	10	66,00%
35	7	7	8	72,00%	6	10	76,00%
36	8	8	9	81,00%	8	9	85,00%
37	9	9	12	93,00%	7	8	93,00%
38	10	10	7	100,00%	10	7	100,00%
39		y mayor...	0	100,00%	y mayor...	0	100,00%

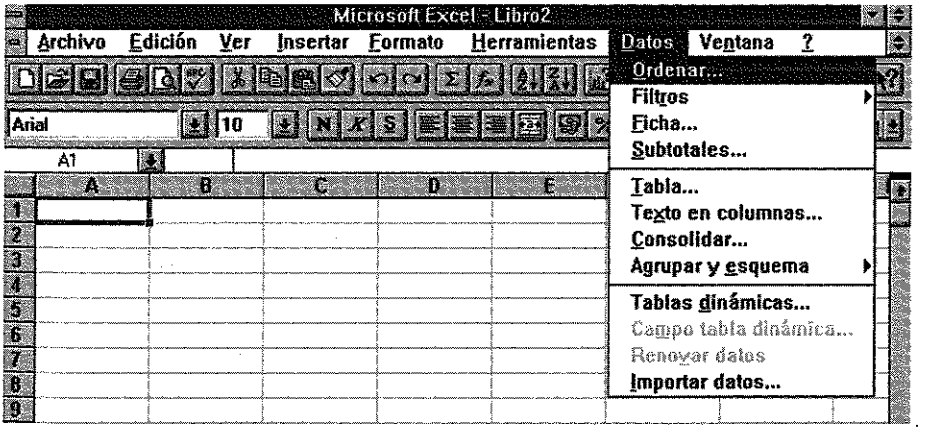


Hoja1 / Hoja4 / Hoja5 / Hoja6 / Hoja7 / Hoja8 / Hoja9

EL MENÚ DATOS

En esta sección avanzamos las posibilidades de trabajo con listas o Bases de Datos que presenta Microsoft Excel en el Menú Datos.

Esta opción presenta diversas alternativas tal como se observa en la siguiente pantalla:



Veamos las principales:

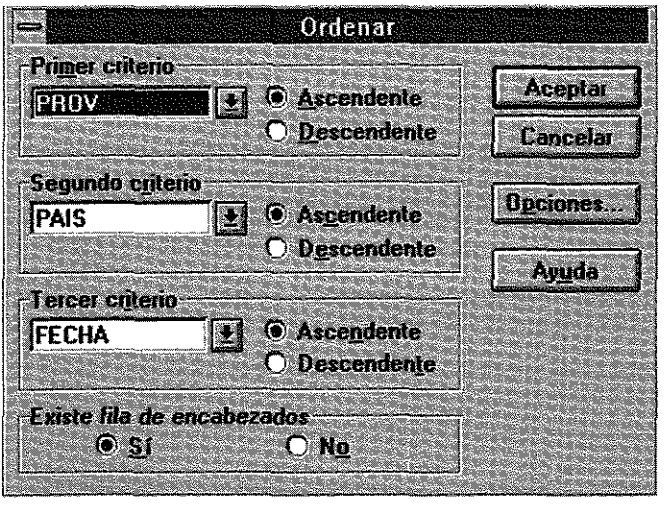
Ordenar datos

Esta tarea es una de las más frecuentes en la gestión de listas o bases de datos. Los datos en una lista están organizados de tal manera que las filas definen registros y las columnas los campos que constituyen la información de los registros. En Excel se pueden seleccionar hasta tres criterios para ordenar la base de datos y dentro de cada uno de ellos elegir el orden de presentación de los datos: ascendente o descendente. Si se quiere ordenar la base de datos siguiendo más de tres criterios, la acción de ordenar se ha de practicar dos o más veces. Para devolver la lista a su estado inicial cuando ésta se ha ordenado siguiendo varios criterios es necesario volver a enumerar previamente las filas, y proceder ordenando la lista por la columna en donde se ha situado la enumeración. Se recomienda antes de realizar estas operaciones enumerar previamente las filas. Con la opción de Ordenar, también se pueden construir criterios personalizados para ordenar datos.

Cuando se utiliza la opción Ordenar hay que tener presente que:

— En las operaciones de ordenación se ha de tener cuidado con las celdas que contienen fórmulas.
— Si se ordena por filas, después de la ordenación, las referencias a celda de la misma fila serán correctas pero no lo serán las referencias a otras filas.
— Si el criterio de ordenación es por columnas, las referencias a celdas de la misma columna serán correctas después de ordenar, pero serán incorrectas las fórmulas que hacen referencia a otras columnas.
— Una forma de evitar este problema es incluir en las fórmulas que se encuentran fuera de la lista, sólo referencias absolutas. Si ordenamos por filas (columnas) debemos evitar las fórmulas que hagan referencia a otras filas (columnas).

En el ejemplo siguiente, que utiliza una base de datos con procedencia de turistas por país y destino por trimestres a distintas provincias españolas; se han utilizado tres criterios para ordenar la lista: provincia, país y fecha.



3.11.2. Las funciones estadísticas en SPSS

Como continuación a las indicaciones dadas en el epígrafe 2.7.2, abordaremos en este apartado otros aspectos de interés en el manejo de SPSS y su uso para extraer los estadísticos indicados en este capítulo.

- **Lectura e importación de datos**: los archivos de datos pueden tener formatos muy diversos, SPSS está diseñado para trabajar con muchos de ellos: entre otros hojas de cálculo creadas con Excel y Lotus, tablas de bases de datos de diferentes orígenes de bases de datos, incluido Oracle, SQL Server, Access, dBASE, entre otros, archivos de texto delimitados por tabuladores y otros tipos de archivos de sólo texto, archivos de datos con formato SPSS creados en otros sistemas operativos, etc.
- ***Apertura de un archivo de datos;*** además de los archivos guardados en formato SPSS, puede abrir archivos de Excel, SAS, archivos delimitados por tabuladores y otros archivos sin necesidad de convertirlos a un formato intermedio ni de introducir información sobre la definición de los datos.
- ***Para abrir archivos de datos,*** elija en los menús: Archivo -> Abrir -> Datos... y en el cuadro de diálogo «Abrir archivo», seleccione el archivo que desea abrir.
- ***La ayuda en SPSS:*** la ayuda se proporciona de diversas formas; en la mayoría de las ventanas de SPSS el menú Ayuda proporciona acceso al sistema de ayuda principal además de a los tutoriales y al material de referencia técnica; por **Temas,** se proporciona acceso a las pestañas Contenido, Índice y Buscar, que pueden usarse para buscar temas específicos de la Ayuda; el «**Tutorial**» aporta instrucciones ilustradas paso a paso sobre cómo utilizar muchas de las funciones básicas de SPSS, la opción «**Estudios de casos** aporta ejemplos prácticos sobre cómo crear diferentes tipos de análisis estadísticos y cómo interpretar los resultados y finalmente, el «**Asesor estadístico**» incluye un método de asistencia para orientarle en el proceso de búsqueda del procedimiento que desea utilizar; el Asesor estadístico proporciona acceso a la mayoría de los procedimientos estadísticos y de generación de informes en el sistema Base y en los procedimientos de creación de gráficos, etc.

Analizaremos en este apartado tres funciones de interés: la transformación de datos, la gestión y transformación de ficheros de datos y la generación de estadísticos descriptivos.

La transformación de datos

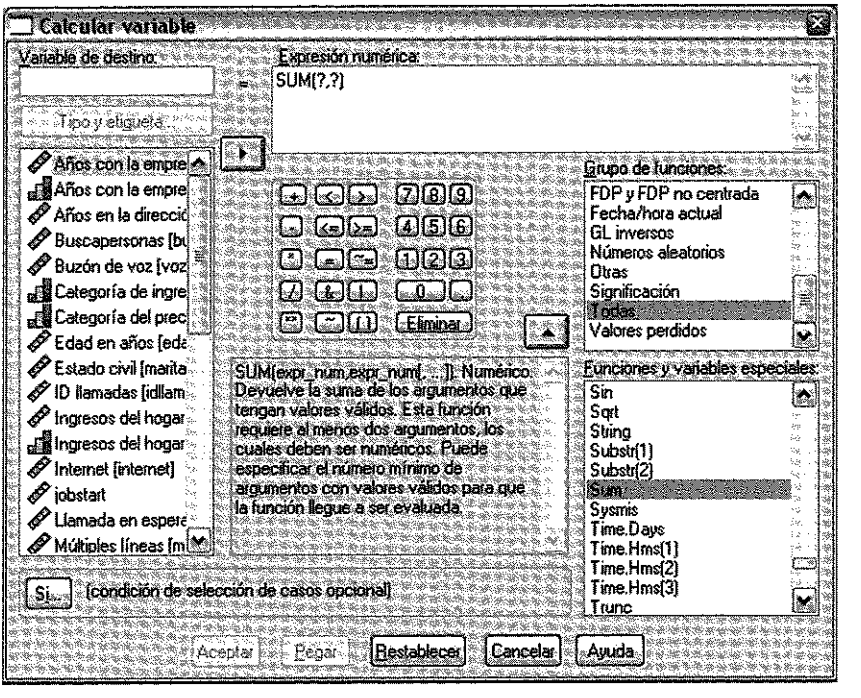
El análisis preliminar de la información a estudiar puede revelar esquemas de codificación poco prácticos o errores de codificación, o bien pueden requerirse transformaciones de los datos para trabajar posteriormente una mejor relación entre las variables.

SPSS puede realizar transformaciones de los datos de todo tipo, desde tareas sencillas, como la agrupación de categorías para su análisis posterior, hasta otras

más avanzadas, como la creación de nuevas variables basadas en ecuaciones complejas e instrucciones condicionales.

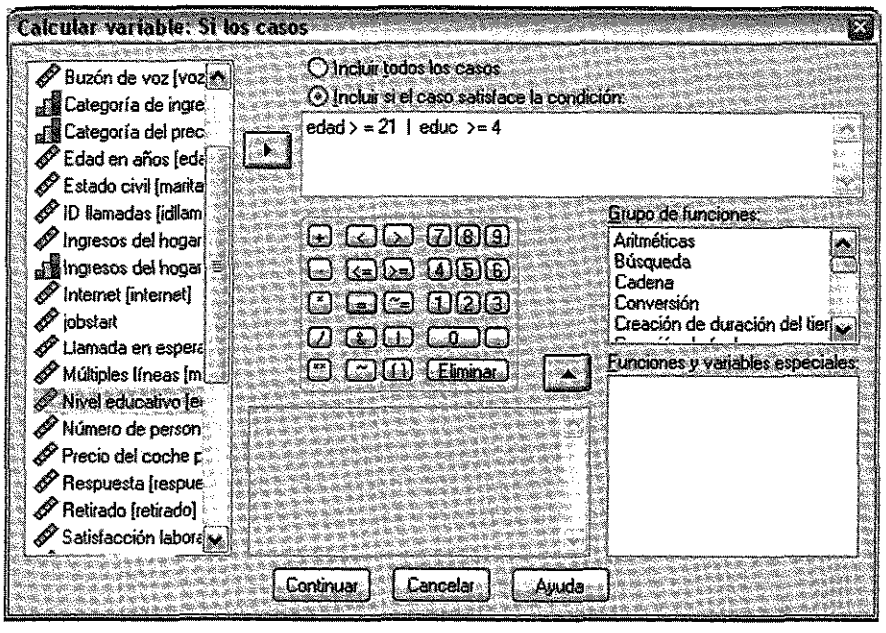
Para realizar el cálculo de nuevas variables, utilice el cuadro de diálogo «Calcular» para calcular los valores de una variable basándose en transformaciones numéricas de otras variables; con el mismo:

— Puede calcular valores para las variables numéricas o de cadena (alfanuméricas).
— Puede crear nuevas variables o bien reemplazar los valores de las variables existentes. Para las nuevas variables, también se puede especificar el tipo y la etiqueta de variable.
— Puede calcular valores de forma selectiva para subconjuntos de datos basándose en condiciones lógicas.
— Puede utilizar más de 70 funciones preincorporadas, incluyendo funciones aritméticas, funciones estadísticas, funciones de distribución y funciones de cadena.



Calcular variable: «Si los casos»

El cuadro de diálogo «Si los casos» permite aplicar transformaciones de los datos para subconjuntos de casos seleccionados utilizando expresiones condicionales. Una expresión condicional devuelve un valor *verdadero*, *falso* o *perdido* para cada caso.



La gestión y transformación de ficheros

Entre la amplia gama de posibilidades de transformación de archivos disponibles se encuentran las siguientes:

- **Ordenar datos.** Puede ordenar los casos en función del valor de una o más variables.
- **Transponer casos y variables.** El formato de archivo de datos de SPSS lee las filas como casos y las columnas como variables. Para los archivos de datos en los que el orden está invertido, se pueden intercambiar las filas y las columnas para leer los datos en el formato correcto.
- **Fundir archivos.** Puede fundir dos o más archivos de datos. Es posible combinar archivos con las mismas variables pero con casos distintos, o con los mismos casos pero variables diferentes.
- **Seleccionar subconjuntos de casos.** Puede restringir el análisis a un subconjunto de casos o efectuar análisis simultáneos de subconjuntos diferentes.
- **Agregar datos.** Puede cambiar la unidad de análisis agregando casos basados en el valor de una o más variables de agrupación.
- **Ponderar datos.** Puede ponderar los casos para un análisis basado en el valor de una variable de ponderación.
- **Reestructurar datos.** Puede reestructurar los datos para crear un único caso (registro) a partir de varios casos o crear varios casos a partir de un único caso.

La generación de estadísticos descriptivos

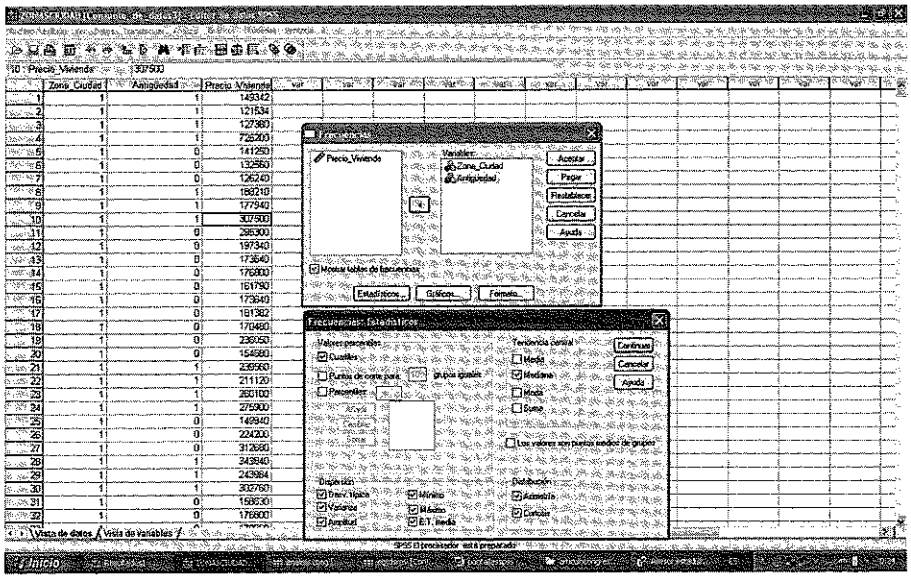
Trabajando con sintaxis, SPSS utiliza las siguientes funciones de Estadística Descriptiva en cuanto a medidas de posición

SUM (A)	HALLA LA SUMA DE LAS OBSERVACIONES DE LA VARIABLE A
SUM (A, B, C ...)	HALLA EL VECTOR DE LAS SUMAS DE LAS OBSERVACIONES DE LAS VARIABLES A, B, C...
MEAN(A)	HALLA LA MEDIA DE LA VARIABLE A
MEAN(A, B, C, ...)	HALLA EL VECTOR DE LAS MEDIAS DE LAS VARIABLES A, B, C, ...
LAG(numvar;n)	DESPLAZA EL COMIENZO DE LA VARIABLE NUMÉRICA NUMVAR N POSICIONES HACIA ADELANTE Y SUSTITUYE LAS N PRIMERAS POSICIONES POR VALORES DESAPARECIDOS. SE TRATA DE LA TÍPICA VARIABLE RETARDO DE ORDEN N.

Sin necesidad de emplear la sintaxis, en general puede trabajarse directamente con los comandos para analizar los datos, bien obteniendo sus Frecuencias o haciendo un procedimiento descriptivo:

Obtención de Frecuencias

Una vez elegido el procedimiento frecuencias donde se escogen las variables de interés y se marcan los estadísticos que queremos calcular en la siguiente pantalla tipo:



Los resultados que se obtienen son del tipo:

Estadísticos

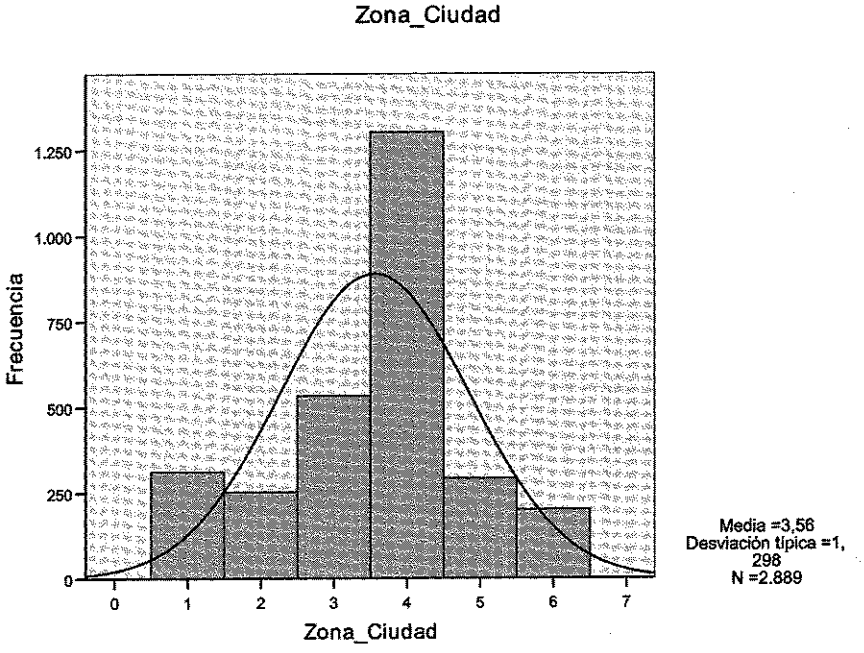
		Zona Ciudad	Antigüedad
N	Válidos	2889	2889
	Perdidos	0	0
Error tip. de la media		,024	,009
Mediana		4,00	1,00
Desv. tip		1,298	,499
Varianza		1,684	249
Asimetría		−,371	−,152
Errot tip. de asimetría		,046	,046
Curtosis		−,158	−1,978
Error tip. curtosis		,091	,091
Rango		5	1
Mínimo		1	0
Máximo		6	1
Percentiles	25	3,00	,00
	50	4,00	1,00
	75	4,00	1,00

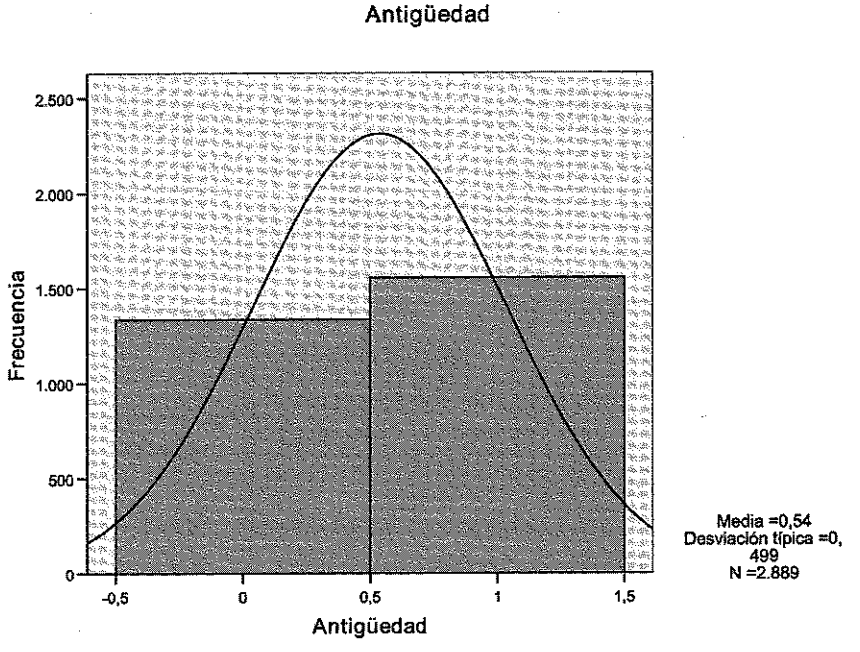
Zona_Ciudad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Zona A	311	10,8	10,8	10,8
	Zona B	252	8,7	8,7	19,5
	Zona C	533	18,4	18,4	37,9
	Zona D	1302	45,1	45,1	83,0
	Zona E	291	10,1	10,1	93,1
	Zona F	200	6,9	6,9	100,0
	Total	2889	100,0	100,0	

Antigüedad

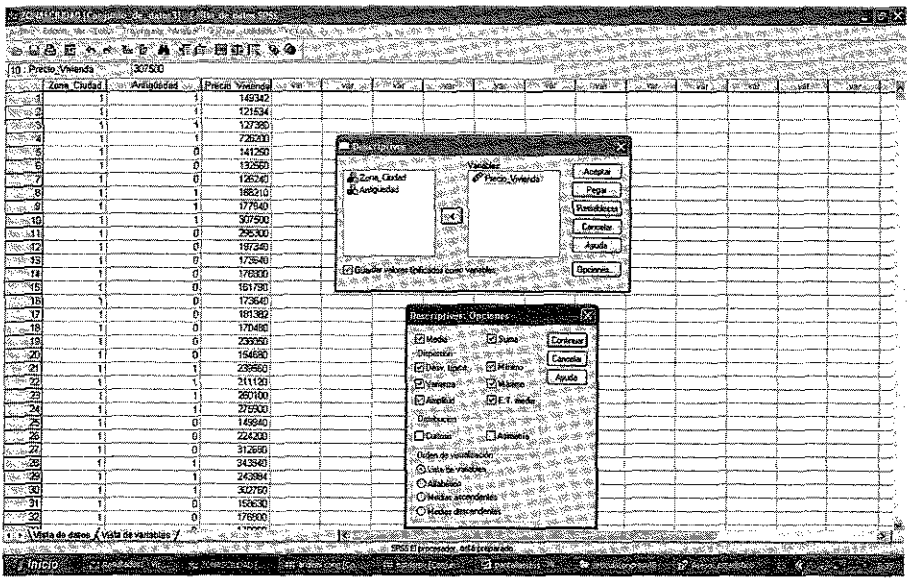
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Vivienda antigua	1335	46,2	46,2	46,2
	Vivienda nueva	1554	53,8	53,8	100,0
	Total	2889	100,0	100,0	


Zona_Ciudad
Frecuencia
1.250
1.000
750
500
250
0
0
1
2
3
4
5
6
7
Zona_Ciudad
Media =3,56
Desviación típica =1,
298
N =2.889


Antigüedad
Frecuencia
2.500
2.000
1.500
1.000
500
0
-0,5
0
0,5
1
1,5
Antigüedad
Media =0,54
Desviación típica =0,
499
N =2.889

Procedimientos Descriptivos

De forma similar al procedimiento anterior procedemos a calcular las medidas estadísticas con una variable numérica; la pantalla tipo será:



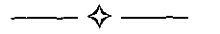
El grupo de resultados aportados es del tipo:

Estadísticos descriptivos (ejemplo):

	Precio_Vivienda	N válido (según lista)
N	Estadístico	2889
Rango	Estadístico	721620
Mínimo	Estadístico	12480
Máximo	Estadístico	734100
Suma	Estadístico	379524352
Media	Estadístico	131368,76
	Error típico	1317,0
Desviación típica	Estadístico	70790,82
Varianza	Estadístico	5011339978,57
Asimetría	Estadístico	1,994
	Error típico	0,046
Curtosis	Estadístico	8,098
	Error típico	0,091

3.12. EJERCICIOS

Sobre medidas de posición en distribuciones unidimensionales



Ejercicio 3.1. *Calcular la media, la mediana y la moda de la masa salarial de una empresa con 500 trabajadores que tiene la siguiente distribución de salarios por intervalos:*

Salario Mensual en euros $(L_{i-1} - L_i)$	N.º de trabajadores (n_i)
600-800 €	85
800-1.000 €	45
1.000-1.200 €	10
1.200-1.400 €	15
1.400-1.600 €	10
1.600-1.800 €	85
1.800-2.000 €	155
2.000-2.200 €	10
2.200-2.400 €	45
2.400-2.600 €	5
2.600-2.800 €	10
2.800-3.000 €	20
3.000-3.200 €	5
Total	450

Respuesta

Una primera observación de esta distribución nos indica que se presenta de una forma técnicamente errónea; los intervalos deben ser excluyentes entre sí, lo que no se respeta en esta presentación; un trabajador que cobra 800 euros de salario podría estar entre los 85 del primer intervalo o entre los 45 del segundo intervalo; dado que coinciden los extremos de los intervalos, deberíamos indicar, en consecuencia, si son cerrados por la izquierda y abiertos por la derecha o viceversa; no obstante, hay que reseñar que muchas presentaciones de datos adoptan un formato similar al indicado, lo que probablemente es un signo de que se ha elegido una presentación de este tipo más por su comodidad o versatilidad que por su pretensión de exactitud.

En todo caso, adoptemos, por más correcto, una presentación más exacta cerrando los intervalos por la derecha; aproximando las marcas de clase y obteniendo las columnas auxiliares de frecuencias absolutas (N_i) y de producto de valores por frecuencias ($X_i\, n_i$), tendremos:

Salario mensual en euros L_{i-1}-L_i	Marca de clase (m_i)	N.º de trabajadores (n_i)	Total Salarios ($x_i\, n_i$)	N.º acumulado de trabajadores (N_i)
(600-800 €]	700 €	85	59.500	85
(800-1.000 €]	900 €	45	40.500	130
(1.000-1.200 €]	1.100 €	10	11.000	140
(1.200-1.400 €]	1.300 €	15	19.500	155
(1.400-1.600 €]	1.500 €	10	15.000	165
(1.600-1.800 €]	1.700 €	85	144.500	250
(1.800-2.000 €]	1.900 €	155	294.500	405
(2.000-2.200 €]	2.100 €	10	21.000	415
(2.200-2.400 €]	2.300 €	45	103.500	460
(2.400-2.600 €]	2.500 €	5	12.500	465
(2.600-2.800 €]	2.700 €	10	27.000	475
(2.800-3.000 €]	2.900 €	20	58.000	495
(3.000-3.200 €]	3.100 €	5	15.500	500
Suma / Total		500	822.000	

La media aritmética viene dada por:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + \ldots + x_r n_r}{n_1 + n_2 + n_3 + \ldots + n_r} = \frac{\sum_{i=1}^{i=r} x_i n_i}{N} = \frac{822.000}{500} = 1.644\ €$$

Para el cálculo de la mediana comprobamos que se trata de una distribución adecuadamente ordenada y obtenemos $N/2 = 250$; calculamos la columna de frecuencias acumuladas y verificamos:

— Si existe N_i que es igual a $N/2$, en cuyo caso, la mediana, por convenio, es el límite superior del intervalo mediano.

— Si existe un N_i que supera a $N/2$, en cuyo caso, la mediana está en el intervalo mediano (L_{i-1}, L_i) que corresponde a dicho N_i.

Dado que de $N_6 = 250$ el intervalo mediano será (1.600-1.800 €] y estaremos en el primer supuesto. La mediana será por tanto en este caso igual a 1.800 €.

Para el cálculo de la moda tenemos en cuenta que todos los intervalos son de la misma amplitud; en este caso la moda absoluta se situará en el intervalo que presente mayor frecuencia absoluta y las modas relativas en el intervalo o intervalos que superen la frecuencia absoluta de los intervalos contiguos.

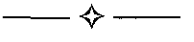
El intervalo que más se repite y, consecuentemente, el intervalo modal absoluto es (1.800-2.000 €].

Examinando la información, se obtienen también diversas modas relativas:

— En el intervalo (1.200-1.400 €], ya que en este intervalo el valor $n_i = 15$ supera al de sus dos intervalos contiguos $n_{i-1} = 10$ y $n_{i+1} = 10$
— En el intervalo (2.200-2.400 €], dónde $n_i = 45$ es superior a $n_{i-1} = 10$ y a $n_{i+1} = 5$
— En el intervalo (2.800-3.000 €], dónde $n_i = 20$; $n_{i-1} = 10$ y $n_{i+1} = 5$

Para obtener el punto modal exacto, aplicamos la expresión [3.7.1.] al intervalo modal absoluto:

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i = 1.800 + \frac{10}{10 + 85} \cdot 200 = 1.821$$



Ejercicio 3.2. *Se ha realizado una Encuesta a 100 clientes, obteniendo la siguiente distribución de edades:*

Edad de los entrevistados	n_i
Menores de 10 años	5
De 10 a 20 años	10
De 21 a 30 años	15
De 31 a 40 años	20
De 41 a 50 años	15
De 51 a 60 años	20
De 61 a 70 años	10
Más de 70 años	5
Total	100

Obtener los principales estadísticos descriptivos de la distribución.

Respuesta

Tenemos una distribución de frecuencias de una variable continua (la edad) que aparece agrupada por intervalos abiertos, cuyos límites o extremos vienen expresados en años.

A fin de facilitar los trabajos construimos la siguiente tabla:

Edad de los entrevistados	m_i	n_i	f_i	N_i	F_i	$m_i n_i$
Menos de 10 años	5	*5*	5%	5	5%	25
De 10 a 20 años	15	*10*	10%	15	15%	150
De 21 a 30 años	25	*15*	15%	30	30%	375
De 31 a 40 años	35	*20*	20%	50	50%	700
De 41 a 50 años	45	*15*	15%	65	65%	675
De 51 a 60 años	55	*20*	20%	85	85%	1.100
De 61 a 70 años	65	*10*	10%	95	95%	650
Más de 70 años	75	*5*	5%	100	100%	375
Total		*100*	100%			4.050

En las que las columnas representan respectivamente:

— La edad de los entrevistados en intervalos
— La marca de clase «aproximada» de cada intervalo, m_i.
— La frecuencia absoluta, n_i.
— La frecuencia relativa, f_i.
— La frecuencia absoluta acumulada, N_i.
— La frecuencia relativa acumulada, F_i.
— Columna $m_i n_i$ obtenida para facilitar los cálculos de la media aritmética.

Como puede comprobarse, la primera decisión que se ha tomado es *imputar* una marca de clase a los dos intervalos abiertos, el menor (5 años) y el mayor (75 años) y *aproximar* una marca de clase para el resto de los intervalos; en el primer intervalo abierto hemos considerado la media entre 0 y 10 años y en el segundo una edad media que parece razonable para los clientes mayores de 70 años; para realizar correctamente esta imputación puede efectuarse un pequeño muestreo entre los entrevistados en ambos grupos de edad.

Media aritmética

Bajo el supuesto de que las marcas de clase del primer y último intervalo sean, respectivamente, de 5 y de 75 años, la media aritmética de la distribución es de 40,5 años (4.050 / 100 = 40,5 años).

Medias geométrica y armónica

Con este tipo de datos no tiene un claro significado las medias geométrica y armónica, por lo que no sería necesario obtenerlas; el alumno puede calcularlas como práctica, comprobando la desigualdad que relaciona las tres medias.

Mediana

$N/2$ da un valor de 50; mirando en la columna de frecuencias absolutas acumuladas N_i, se deduce que hay un N_i que coincide exactamente con 50; se trata del correspondiente al intervalo de 31 a 40 años (intervalo mediano).

En este caso, el valor exacto de la mediana coincide con el límite superior del intervalo, es decir, 40 años, de forma que puede afirmarse que la mitad de la población (50 personas) tiene menos de 40 y la otra mitad más de 40 años.

Moda absoluta

La moda absoluta corresponde a 2 intervalos con frecuencia 20 (distribución bimodal); se trata de los intervalos de edad de 31 a 40 años y de 51 a 60 años. Para obtener el valor exacto aplicaríamos la expresión [3.7.1].

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

Obteniendo como valor exacto:

$M_o = 31 + (15 / (15 + 15) \cdot 10) = 36$ años.
$M_o = 51 + (10 / (15 + 10) \cdot 10) = 55$ años.

Habría que tener en cuenta, sin embargo, que en este caso pueden no ser totalmente aplicables los criterios válidos para la aplicación de la expresión [3.7.1], en el sentido de que podría considerarse que no todos los intervalos (primero y último) tienen la misma amplitud.

Modas relativas

No existen modas relativas.

Cuantiles

El primer cuartil se encuentra en el intervalo donde $100/4 = 25$, el primer N_i que supera esta masa acumulada es el intervalo de 21 a 30 años.

La edad exacta en la que se situaría el cuartil, se tendría, aplicando la expresión [3.8.1].

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

modificada en la siguiente forma:

$$Q_1 = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i} \cdot c_i = 21 + \frac{\frac{1 \cdot 100}{4} - 15}{15} \cdot 10 \cong 27,7$$

Es decir, aproximadamente, 28 años.

El segundo cuartil se corresponde con la mediana (40 años)

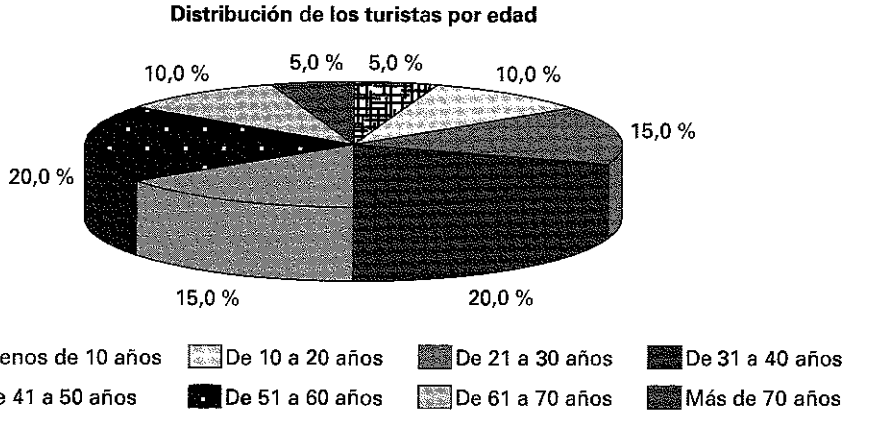
El tercer cuartil se encuentra en el intervalo donde las frecuencias absolutas acumuladas superan a 75 (100 · 3/4 = 75), es decir, en el intervalo de 51 a 60 años; sí se quiere calcular el punto exacto, aplicando la expresión [3.8.1], se obtiene:

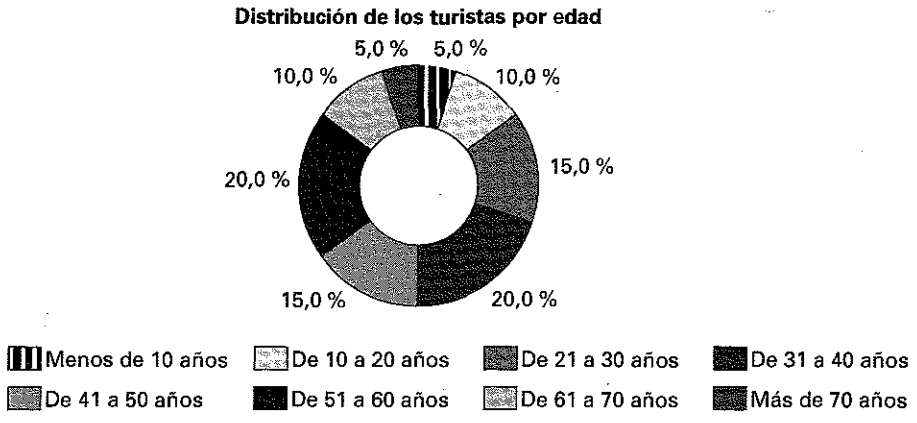
$$Q_1 = L_{i-1} + \frac{\frac{rN}{q} - N_{i-1}}{n_i} + c_i = 51 + \frac{\frac{3 \cdot 100}{4} - 65}{20} \cdot 10 = 56$$

Es decir, aproximadamente, 56 años.

Representación Gráfica

Para representar gráficamente este tipo de datos puede utilizarse un gráfico de sectores o de anillos; empleando cualquier programa de gráficos obtendríamos:





Ejercicio 3.3. *Suponga que cinco empleados de una oficina que son llamados a declarar sus gastos mensuales en telefonía móvil, reportan las siguientes cifras:*

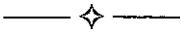
35 € 35 € 35,50 € 35,50 € 180 €

El gasto medio de los 5 empleados es de 64,20 euros ¿Es representativo este valor?

Respuesta

A simple vista puede apreciarse que el promedio no es muy representativo para los datos de la distribución, ya que hay un valor atípico (180 €), que genera una importante alteración de la media (recordemos, en este punto la sensibilidad de la media a los valores extremos).

Para mejorar esta valoración deben calcularse las medidas de dispersión (particularmente la desviación típica y el coeficiente de variación de Pearson) y las medidas de concentración, que veremos, en ambos casos, en el capítulo siguiente.



Ejercicio 3.4. *En la siguiente tabla se presentan los datos de una compañía que utiliza tres niveles de mano de obra (no cualificada, semi-cualificada y cualificada) para producir dos productos.*

La compañía desea saber el coste promedio de trabajo por hora para cada uno de los productos.

Nivel de trabajo	Salario por hora	Horas de trabajo por unidad producida	
		Producto 1	Producto 2
No cualificado	5	1	4
Semi-cualificado	7	2	3
Cualificado	9	5	3

Respuesta

Debemos realizar un promedio ponderado.

— Para el producto 1 sobre 8 horas de trabajo (1 + 2 + 5):

$$\bar{x}_{p_1} = \frac{1 \cdot 5 + 2 \cdot 7 + 5 \cdot 9}{8} = 8$$

— Para el producto 2 sobre 10 horas de trabajo (4 + 3 + 3):

$$\bar{x}_{p_2} = \frac{4 \cdot 5 + 3 \cdot 7 + 3 \cdot 9}{10} = 6{,}8$$

Por lo tanto el producto 1 tiene un coste promedio de trabajo por hora de 8 € y el producto 2 de 6,8 €.

Ejercicio 3.5. *Un tren recorre 100, 300 y 350 Km. a las velocidades medias de 50, 60 y 70 Km. por hora. Calcule la velocidad media para el recorrido total.*

Respuesta

Se debe realizar la media armónica, ya que esta medida es más representativa que el resto de las medias para obtener promedios de velocidades, rendimientos y productividades.

$$H = \frac{100 + 300 + 350}{\frac{100}{50} + \frac{300}{60} + \frac{350}{70}} = \frac{750}{12} = 62{,}5$$

Ejercicio 3.6. *Los siguientes valores corresponden al tiempo que esperan para ser atendidos 15 clientes de una oficina bancaria (datos en minutos):*

20, 25, 22, 20, 25, 20, 21, 22, 22, 24, 23, 20, 23, 20, 23

Obtener:

a) *El tiempo medio de espera.*

b) *El tiempo máximo que esperó el 50% de los clientes.*

c) *El tiempo más frecuente de espera.*

Respuesta

Ordenamos los valores de menor a mayor

20, 20, 20, 20, 20, 21, 22, 22, 22, 23, 23, 23, 24, 25, 25

a) La media de los valores es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20 \cdot 5 + 21 + 22 \cdot 3 + 23 \cdot 3 + 24 + 25 \cdot 2}{15} = 22 \text{ minutos.}$$

b) El valor pedido coincide con la mediana.
Al tratarse de una distribución ordenada y de tipo I, con un número de observaciones impar (15 observaciones), la mediana coincide con el valor que se encuentra justo en medio, es decir, en octavo lugar, o sea 22 minutos; este valor deja 7 observaciones a su derecha y otras 7 a su izquierda.

c) La moda es 20 minutos, el valor que más se repite (5 veces).

Ejercicio 3.7. *Los datos que se dan a continuación representan las edades de los clientes de un establecimiento de venta de vehículos en el último mes.*

35 25 66 43 40 38 30 56 67 56
39 33 65 53 25 37 33 52 44 48

a) *Construya una distribución de frecuencias con clases 20 – 29, 30 – 39, etc.*

b) *Calcule la media de la muestra a partir de la distribución de frecuencias agrupadas.*

c) *Calcule la media de la muestra a partir de los datos sin agrupar.*

d) *Compare las respuestas a los apartados b) y c).*

Respuesta

a) Construyamos una tabla de frecuencias.

Clase L_{i-1} - L_i	Frecuencia n_i
20 - 29	2
30 - 39	7
40 - 49	4
50 - 59	4
60 - 69	5

b) Busquemos la media a partir de los datos agrupados.

Clase L_{i-1} - L_i	Marca de clase m_i	Frecuencia n_i	$m_i \cdot n_i$
20 - 29	24,5	2	49
30 - 39	34,5	7	241,5
40 - 49	44,5	4	178
50 - 59	54,5	4	218
60 - 69	64,5	3	193,5
Total		20	880

$$\bar{x} = \frac{\sum_{i=1}^{r} m_i\, n_i}{\sum_{i=1}^{r} n_i} = \frac{880}{20} = 44$$

c) Busquemos la media a partir de los datos sin agrupar.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{885}{20} = 44{,}25$$

d) Como era de esperar, son parecidas pero no exactamente iguales, ya que al agrupar los datos siempre se pierde precisión en el cálculo de los estadísticos.

Ejercicio 3.8. *Una empresa de fabricación de maquinaria tuvo durante el último año 190 clientes; una vez clasificados dichos clientes por intervalos de ventas se dispone de la siguiente información:*

Ventas en miles de euros	N.º de clientes
100 a 120	40
120 a 140	70
140 a 160	60
160 a 180	20

Obtener:

a) *El promedio de ventas por cliente.*

b) *El gerente de la empresa desea conocer el coste que tendría aplicar un descuento del 1% a sus mejores clientes, entendiendo como tales los que facturaron más de 160.000 ⇔¿Cuál sería el coste de dicha operación comercial?*

b) *¿Qué porcentaje de clientes compraron entre 120 y 160 mil euros?*

Respuesta:

a) $\bar{x} = \dfrac{110 \cdot 40 + 130 \cdot 70 + 150 \cdot 60 + 170 \cdot 20}{190} = 136{,}31$ miles de euros.

Para ello se calcularon previamente las marcas de clase: 110, 130, 150 y 170.

b) El número de clientes cuyo volumen de ventas superó los 160 mil euros es 20; sí para el intervalo 160.000-180.000 se considera una marca de clase de 170.000 €, se tendrá unas ventas totales para este grupo de clientes de 3.400.000 €; el coste de aplicar un 1% de descuento a este grupo sería, en consecuencia, de 34.000 €.

c) El número de clientes que compraron entre 120 y 160 mil euros es: 70 + 60 = 130, esto corresponde al 68,4% del total.

Ejercicio 3.9. *Los siguientes datos corresponden a la superficie en hectáreas de una muestra de explotaciones agrarias de una determinada comarca.*

49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68,
60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53,
67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31, 84

a) *Calcule la media y la mediana.*

b) *Encuentre los cuartiles inferior y superior.*

c) *Encuentre los percentiles quinto y noveno.*

d) *Elimine la observación más pequeña y vuelva a calcular los apartados* a), b) *y* c). *Explique los resultados.*

Respuesta

Ordenamos los datos de menor a mayor

31 40 45 49 52 53 57 58 58 60 61 61 63 66 67 67 67 67
68 69 70 70 70 70 72 73 75 75 76 76 78 79 80 81 83 84

Y construimos una tabla de frecuencias, con las columnas auxiliares i (orden), N_i (frecuencia absoluta) y $x_i n_i$:

i	x_i	n_i	N_i	$x_i n_i$	i	x_i	n_i	N_i	$x_i n_i$
1	31	1	1	31	14	68	1	19	68
2	40	1	2	40	15	69	1	20	69
3	45	1	3	45	16	70	4	24	280
4	49	1	4	49	17	72	1	25	72
5	52	1	5	52	18	73	1	26	73
6	53	1	6	53	19	75	2	28	150
7	57	1	7	57	20	76	2	30	152
8	58	2	9	116	21	78	1	31	78
9	60	1	10	60	22	79	1	32	79
10	61	2	12	122	23	80	1	33	80
11	63	1	13	63	24	81	1	34	81
12	66	1	14	66	25	83	1	35	83
13	67	4	18	268	26	84	1	36	84
Total							36		2.371

a) La media en este caso toma el valor $\bar{x} = \frac{2.371}{36} = 65{,}86$ Has.

Como se trata de una distribución de tipo II y $N/2 = 18$ coincide exactamente con $N_{(13)}$ (N situado en el orden 13), la mediana será el promedio de los dos valores centrales

$$M_e = \frac{67 + 68}{2} = 67{,}5 \text{ Has.}$$

b) El primer cuartil viene dado por el primer valor de la variable cuya frecuencia acumulada supera a $N/4$; en nuestro caso $\frac{36}{4}$; en la tabla anterior puede comprobarse que el valor cuya frecuencia acumulada es $N_i = 9$ corresponde a una superficie de 58 Has,

$$\text{Luego } Q_1 = \frac{58 + 60}{2} = 59 .$$

Esta situación se interpreta como que el 25% (un cuarto) de las explotaciones analizadas tienen menos de 59 Has de superficie.

El tercer cuartil vendrá dado por el primer valor de la variable cuya frecuencia acumulada supera a $3\,N/4$; en nuestro caso $\frac{3 \cdot 36}{4}$; ; en la tabla anterior puede comprobarse que el primer valor cuya frecuencia acumulada supera a 27 ($N_{19} = 28$) corresponde a 75 Has.

Esta situación se interpreta como que el 75% (tres cuartas partes) de los días analizados tienen menos de 75 Has.

c) Para hallar el quinto percentil, $k = 0{,}05$. Luego, $nk = 36 \cdot 0{,}05 = 1{,}8$. Como el valor que se encuentra en el lugar 2 (primero que supera a 1,8) es 40, se concluye que $P_5 = 40$.

Ello quiere decir que un 5% de las explotaciones analizadas tienen menos de 40 Has.

Para hallar el noveno percentil, $k = 0{,}09$. Luego, $nk = 36 \cdot 0{,}09 = 3{,}24$. Como el valor que se encuentra en el lugar 4 es 49, se concluye que $P_9 = 49$; es decir, el 9% de las explotaciones tienen un máximo de 49 Has de superficie.

d) Lo mejor para obtener los ratios sin el valor más pequeño (31 Has), es volver a construir la nueva tabla, en este caso con 35 explotaciones:

i	x_i	n_i	N_i	$x_i n_i$	i	x_i	n_i	N_i	$x_i n_i$
1	40	1	1	40	14	68	1	19	68
2	45	1	2	45	15	69	1	19	69
3	49	1	3	49	16	70	4	20	280
4	52	1	4	52	17	72	1	24	72
5	53	1	5	53	18	73	1	25	73
6	57	1	6	57	19	75	2	26	150
7	58	2	7	116	20	76	2	28	152
8	60	1	9	60	21	78	1	30	78
9	61	2	10	122	22	79	1	31	79
10	63	1	12	63	23	80	1	32	80
11	66	1	13	66	24	81	1	33	81
12	67	4	14	268	25	83	1	34	83
13	68	1	18	68	26	84	1	35	84
						Total	35		2.340

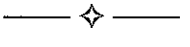
d) La media para los datos sin 31 $\bar{x} = \dfrac{2.340}{35}$ = es 66,86 Has. Ahora la cantidad de datos es impar (35), y $N/2 = 17{,}5$, por lo tanto la mediana es el primer valor cuya frecuencia acumulada supere este valor, es decir, 68 Has.

El primer cuartil es $Q_1 = \dfrac{35}{4} = 8{,}75 \Rightarrow$ *el valor 60*

y el tercero es $Q_3 = \dfrac{3 \cdot 35}{4} = 26{,}25 \Rightarrow$ *el valor 76 Has.*

Para hallar el quinto percentil, $k = 0{,}05$. Luego, $nk = 35 \cdot 0{,}05 = 1{,}75$. El N_i acumulado que lo supera es 2, equivalente a 45 Has, se concluye que $P_5 = 40$.

Para hallar el noveno percentil, $k = 0{,}09$. Luego, $nk = 35 \cdot 0{,}09 = 3{,}15$; el primer valor que lo supera N > 3,15 es 4, por lo que se concluye que $P_9 = 52$ Has.



Ejercicio 3.10. *La facturación mensual de 50 compañías de transporte durante un determinado mes se recogen en la siguiente tabla en miles de euros.*

$L_{i-1} - L_i$	(40, 100]	(100, 200]	(200, 500]	(500, 1000]
n_i	10	20	15	5

Calcule la mediana y la moda.

Respuesta

a) Mediana:

Tenemos una distribución agrupada en la que los intervalos tienen distinta amplitud; construimos una tabla de frecuencias, en la que calculamos una columna con la amplitud de los intervalos y otra columna con la densidad de los intervalos:

$L_{i-1} - L_i$	n_i	c_i	$h_i = n_i / c_i$	N_i
(40, 100]	10	60	0,166667	10
(100, 200]	20	100	0,2	30
(200, 500]	15	300	0,05	45
(500, 1000]	5	500	0,010	50
	$N = 50$			

Obtenemos el intervalo mediano; para ello calculamos

$$\frac{N}{2} = \frac{50}{2} = 25$$

El valor de $N_i = 30$ supera a 25; en consecuencia, el intervalo mediano será (100, 200].

Para calcular el punto exacto en el que se encuentra la mediana utilizamos la expresión

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

en la que c_i es la amplitud del intervalo ($c_i = L_i - L_{i-1}$); en nuestro caso la amplitud del intervalo mediano será de 100 y la expresión tomará el siguiente valor.

$$M_e = 100 + \frac{25 - 10}{20} \cdot 100 = 175$$

Se dice, en consecuencia que la mediana se encuentra en el intervalo (100, 200] y que, por convenio reconocido entre los estadísticos, el punto mediano es 175.

b) Moda:

Para el cálculo de la moda también debe tenerse en cuenta que los intervalos son de diferente amplitud. En la columna de densidad del intervalo h_i,

puede observarse que el intervalo de mayor frecuencia relativa es el segundo; en consecuencia el intervalo modal es el intervalo (100, 200].

Para el cálculo del punto modal se utiliza la expresión:

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

Operando se tiene:

$$M_o = 100 + \frac{0{,}05}{0{,}05 + 0{,}1667} \cdot 100 \cong 123$$

Se dice, en consecuencia, por convenio reconocido entre los estadísticos, que la moda de la distribución está en el punto 123, es decir, que la moda es aproximadamente de 123.000 euros.

Ejercicio 3.11. *El indicador del nivel de renta mensual de cien familias de un pequeño municipio, muestra la siguiente distribución en miles de euros:*

$L_{i-1} - L_i$	[0, 1]	(1, 2]	(2, 4]	(4, 6]	(6, 10]	(10, 15]
n_i	20	30	20	15	10	5

Calcule la mediana y la moda.

Respuesta

$L_{i-1} - L_i$	n_i	c_i	$h_i = n_i / c_i$	N_i
[0, 1]	20	1	20	20
(1, 2]	30	1	30	50
(2, 4]	20	2	10	70
(4, 6]	15	2	7,5	85
(6, 10]	10	4	2,5	95
(10, 15]	5	5	1	100
	$N = 100$			

a) Para obtener la mediana calculamos $\frac{N}{2} = \frac{100}{2}$; en las frecuencias acumuladas el valor que iguala a $N_i = 50$ con lo que el intervalo mediano será (1, 2] y la mediana será el límite superior, luego $M_e = 2$.

b) El intervalo modal será también el intervalo (1, 2] y la moda se obtiene resolviendo la siguiente expresión.

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

$$M_o = 1 + \frac{10}{20 + 10} \cdot 1 \cong 1{,}33$$

Ejercicio 3.12. *En una encuesta aleatoria efectuada a 120 clientes de una empresa se obtuvieron las siguientes calificaciones (valoración de 1 a 10 sobre la calidad del servicio recibido) y frecuencias:*

Calidad del servicio x_i	Frecuencia n_i	Frecuencia acumulada N_i
1	20	20
3	30	50
4	20	70
5	40	110
7	7	117
9	3	120

Calcule la moda y la mediana.

Respuesta

a) La moda es el valor de la variable con mayor frecuencia absoluta, por lo tanto $M_o = 5$; esta puntuación es el valor otorgado por el mayor número de clientes.

b) La mediana representa un valor que divide a la serie de datos ordenada en dos partes iguales;

$$\frac{N}{2} = \frac{120}{2} = 60$$

La posición 60 de la distribución ordenada (de hecho desde la posición 60 a la 70), corresponde al valor 4; por tanto la mediana corresponde a este valor.

Ejercicio 3.13. *Los salarios mensuales de 200 trabajadores de una determinada empresa se recogen en la siguiente distribución de frecuencias:*

$L_{i-1} - L_i$	[750-1250]	(1250-1750]	(1750-2250]	(2250-2750]
n_i	25	100	50	25

Calcule la moda y la mediana.

Respuesta

$L_{i-1} - L_i$	Marca de clase m_i	n_i	N_i
[750-1.250]	1.000	25	25
(1.250-1.750]	1.500	100	125
(1.750-2.250]	2.000	50	175
(2.250-2.750]	2.500	25	200

a) La mayor frecuencia absoluta es $n_i = 100$ por lo que el intervalo modal será: (1.250, 1.750]. Entonces el valor de la moda es:

$$M_o = 1.250 + \frac{50}{25+50} \cdot 500 \cong 1.583$$

Realmente, con la información disponible sólo podemos decir que el salario más habitual en la empresa se encuentra en al intervalo 1.250-1.750 euros; o obstante, con una estimación aproximada, podemos decir que este salario es en concreto 1.583 euros.

Esta cifra tiene la siguiente interpretación:
La marca de clase del intervalo modal es 1500 euros

$$M_i = \frac{1.250 + 1.750}{2} = 1.500;$$

algún convenio estadístico podría decir que esta es la cantidad que constituye la moda; sin embargo, el convenio antes aplicado, y el más común entre los estadísticos, tiene en cuenta las frecuencias de los intervalos anterior y posterior; en nuestro caso, sí existen 50 trabajadores que ganan entre 1.750 y 2.250 euros y sólo 25 trabajadores que ganan entre 750 y 1.250 euros, lo que se concluye es que probablemente haya más trabajadores de la mitad del intervalo modal en adelante (es decir de 1.500 a 1.750 €) que los que hay de la mitad hacia el límite inferior (es decir de 1.250 a 1.500 €); es por ello más probable que sea 1.583 que 1.500.

b) $\frac{N}{2} = 100$. El valor de $N_i = \mathbf{125}$ supera a 100, el intervalo mediano será (1.250, 1.750] y la mediana se calcula a través de

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

siendo c_i la amplitud del intervalo ($c_i = L_i - L_{i-1}$).

$$M_e = 1.250 + \frac{\frac{200}{2} - 25}{100} \cdot 500 \quad \Rightarrow \quad M_e = 1.625$$

La interpretación del dato es similar a la indicada para la moda.

Ejercicio 3.14. *Una asociación creada con la finalidad de ofrecer a sus socios productos de alta calidad a precios menores de los que se observan en un determinado mercado, registra en la actualidad 200 socios.*

Debido a la imposibilidad de tener un trato directo con cada uno de sus miembros, decide realizar una encuesta para conocer las características socio-demográficas y los hábitos de compra de sus socios. Para ello tomó una muestra de 40 familias y, entre otras, se consideraron las siguientes variables:

a) *Nivel de educación del cabeza de familia, con las siguientes categorías (1 Primaria (completa o incompleta), 2 Secundaria incompleta, 3 Secundaria completa, 4 Universitaria incompleta, 5 Universitaria completa).*
b) *Cantidad de personas por familia.*
c) *Ingresos mensuales de la familia).*

Los resultados fueron:

TABLA: Educación, cantidad de personas e ingresos mensuales por familia

N.º orden	Nivel de educación	Cantidad de personas por familia	Ingresos mensuales familiares	N.º orden	Nivel de educación	Cantidad de personas por familia	Ingresos mensuales familiares
1	1	1	250	21	1	3	800
2	1	1	280	22	3	4	1.350
3	1	1	200	23	5	5	1.300
4	1	1	500	24	3	8	2.100
5	2	1	800	25	1	1	500
6	2	1	700	26	4	2	1.800
7	2	1	350	27	3	3	2.400
8	1	2	280	28	4	3	2.200
9	2	2	600	29	2	4	900
10	5	5	2.450	30	2	4	2.000
11	2	3	450	31	1	5	450
12	2	3	350	32	4	6	2.450
13	3	4	1.700	33	5	1	1.400
14	2	4	900	34	5	3	1.200
15	2	4	600	35	5	4	2.400
16	1	2	280	36	3	4	1.600
17	3	2	450	37	5	3	2.000
18	2	2	350	38	2	3	900
19	2	3	800	39	2	4	900
20	3	4	1.100	40	4	4	1.550

En relación con la situación planteada:

— *Calcule la moda para las tres variables consideradas.*

— *Construya la tabla de frecuencias adecuada para cada caso.*

Respuesta

— Variable: *nivel de educación:*

Construimos la tabla unidimensional de frecuencias:

Nivel de educación	Frecuencia
1	9
2	14
3	7
4	4
5	6
Total	40

Moda: La moda viene dada por el Nivel 2, es decir: Secundario incompleto.

— Variable: *Cantidad de personas por familia:*

Construimos su tabla unidimensional de frecuencias:

Cantidad de personas por familia	Frecuencia
1	9
2	6
3	9
4	11
5	3
6	1
8	1
Total	40

La moda viene dada por el valor 4, es decir las familias más frecuentes son las de 4 miembros.

— Variable: *ingresos mensuales familiares*

Para estudiar esta variable conviene establecer intervalos; una determinación arbitraria, pero apropiada, de los intervalos de clase da como resultado la siguiente tabla de frecuencias:

Intervalos de clase	n_i
0 - 500	11
500 - 1.000	12
1.000 - 1.500	5
1.500 - 2.000	4
2.000 - 2.500	8
Total	40

El intervalo modal es el de 500-1000 euros; sí queremos determinar el punto modal exacto de este intervalo utilizaremos la expresión

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

Cuyo valor sería:

$$M_o = 500 + \frac{5}{11+5} \cdot 500 = 656,2$$

Es decir, sería inferior a la marca de clase del intervalo modal (750), ya que hay más densidad en el intervalo anterior (11 familias) que en el posterior (5 familias).

Ejercicio 3.15. *Una empresa dispone de los siguientes datos sobre su producción semanal (unidades del bien fabricado):*

TABLA: Frecuencias absolutas. Producción por semana

Producción	Cantidad de semanas
3- 5	6
5 - 7	9
7 - 9	15
9 - 11	18
11 - 13	2
Total	50

Calcule la mediana de producción semanal.

Respuesta

Calculando las frecuencias acumuladas:

Producción	n_i	N_i
3 - 5	6	6
5 - 7	9	15
7 - 9	15	30
9 - 11	18	48
11 - 13	2	50
Total	50	

Tenemos una distribución con intervalos de la misma amplitud ($c_i = 2$).
La mediana estará situada en el intervalo 7-9, ya que:

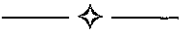
$$\frac{N}{2} = \frac{50}{2} = 25$$

En el intervalo 7-9 la frecuencia acumulada supera al valor medio de 25.
Para aproximar el punto mediano exacto, podemos utilizar la expresión:

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

Que nos da:

$$M_e = 7 + 2 \cdot \frac{25 - 15}{15} = 7 + 2 \cdot \frac{10}{15} \cong 8{,}3 \text{ unidades}$$



Ejercicio 3.16. *Los siguientes datos corresponden a los meses de vida operativa de la flota de una compañía de transporte aéreo. Los aviones fueron divididos en grupos de acuerdo con su tamaño (en cuanto al número de pasajeros que pueden ocuparlos) en las siguientes categorías: pequeño, mediano, grande y muy grande, con el siguiente resultado:*

TABLA: Frecuencias absolutas.
Meses de vida operativa de aviones

Pequeño	Mediano	Grande	Muy grande	Pequeño	Mediano	Grande	Muy grande
72	30	8	177	999	25	24	52
4	84	92	162	112	21	18	164
228	4	35	216	242	13	31	19
126	54	117	553	991	87	51	5
118	13	132	278	111	2	90	15
18	23	12	12	1	20	52	43
81	97	162	260	587	7	73	340
110	153	3	200	389	24	8	133
314	59	95	15	38	99	36	111
100	117	48	182	25	8		231
42	16	7	143	357	99		
8	151	140	10	467	61		
144	22	186	103	1	25		
25	56	84	250	10	95		
11	21	19	100	30	80		
44	18	45	378	15	52		
283	139	80	49		29		

Efectúe un análisis exploratorio de los datos a través del cual pueda obtener algunas primeras conclusiones.

Respuesta

El análisis de estos datos descarta la media aritmética como medida de posición relevante, ya que la variable «meses de vida» presenta valores con importantes extremos en cada una de las categorías, lo que desvirtúa la utilización de esta medida para cualquier clase de análisis.

Aunque podrían efectuarse otros análisis más complejos, la medida de posición más conveniente es la mediana; que nos aporta la siguiente información:

Medida	Tamaño			
	Pequeño	Mediano	Grande	Muy grande
M_e	100	29,5	49,5	143

Que indica que la mitad de los aviones pequeños tenían más de 100 meses de vida útil y la otra mitad menos de 100 meses; en el caso de los aviones medianos, el 50% tiene menos de 29,5 meses de vida operativa, etc.

También es interesante destacar la frecuencia de cada una de las categorías en la flota:

Pequeño	Mediano	Grande	Muy Grande
27,5%	28,3%	21,7%	22,5%

Ejercicio 3.17. *Los siguientes datos provienen de una investigación efectuada en el Departamento de Diseño de Producto de una empresa de software.*

El ensayo consistió en la aplicación de un test para medir el tiempo transcurrido entre los fallos producidos por un cierto software. La variable «tiempo entre fallos» fue medida en segundos y agrupada en los siguientes intervalos de clase:

Tabla: Cantidad de fallos según tiempo

Intervalos de clase	Cantidad de fallos	Intervalos de clase	Cantidad de fallos
0 - 50	2	250 - 300	4
50 - 100	3	300 - 350	10
100 - 150	10	350 - 400	5
150 - 200	15	400 - 450	3
200 - 250	16	450 - 500	2

Con los datos de la investigación:

a) *Represente gráficamente la distribución de frecuencias.*

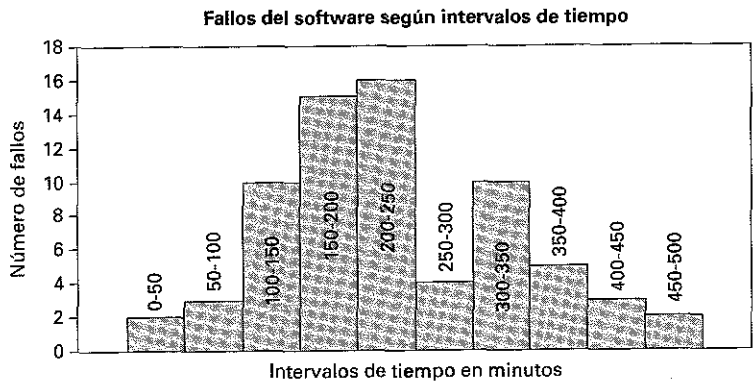
b) *Calcule la media aritmética ponderada.*

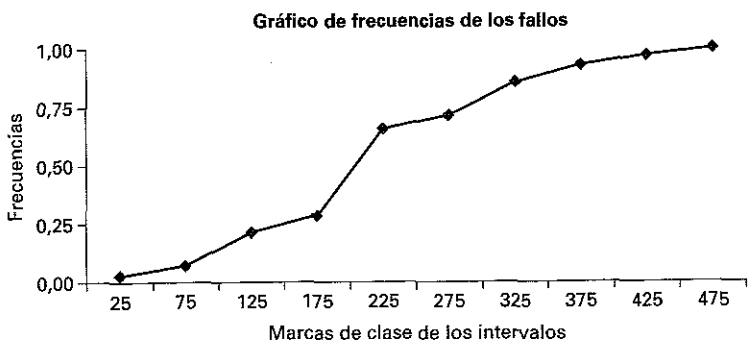
c) *Responda gráfica y analíticamente, a partir de qué tiempo fallan más del 75% del software sometido a dicho test?*

Respuesta

a) Variable: Tiempo entre fallos.

Intervalos de clase	n_i	Marca de clase (m_i)	$m_i \cdot n_i$	N_i	f_i	F_i
0 - 50	2	25	50	2	0,0286	0,0286
50 - 100	3	75	225	5	0,0429	0,0715
100 - 150	10	125	1.250	15	0,1429	0,2144
150 - 200	15	175	2.625	30	0,2143	0,4287
200 - 250	16	225	3.600	46	0,2286	0,6573
250 - 300	4	275	1.100	50	0,0571	0,7144
300 - 350	10	325	3.250	60	0,1429	0,8573
350 - 400	5	375	1.875	65	0,0714	0,9287
400 - 450	3	425	1.275	68	0,0429	0,9716
450 - 500	2	75	950	70	0,0286	1,0000
Total	70		16.200		1,0000	





b) $\bar{x} = \dfrac{16.200}{70} = 231{,}43$ seg. CPU

c) $\dfrac{3N}{4} = 3 \qquad \dfrac{70}{4} = 52{,}5$

$$Q_3 = 300 + 50\frac{(52{,}5 - 50)}{(60 - 50)} = 300 + 12{,}5 = 312{,}5 \text{ seg. CPU}$$

Ejercicio 3.18. *Se registraron las medidas en metros de 100 productos y se agruparon los datos en la siguiente tabla de distribución de frecuencias:*

TABLA: Frecuencias absolutas. Distribución de medidas

Intervalos de clase (metros)	Cantidad de productos
1,20 - 1,30	5
1,30 - 1,40	18
1,40 - 1,50	42
1,50 - 1,60	27
1,60 - 1,70	8

a) *Calcule la moda, mediana y el promedio; compare los resultados.*

b) *Calcule los cuartiles* Q_1 *y* Q_3.

c) *Grafique la distribución de frecuencias relativas acumuladas y marque en la misma* Q_1, Q_2 *y* Q_3..

Respuesta

Intervalos de clase	n_i	Marca de clase (m_i)	$m_i \cdot n_i$	N_i	f_i	F_i
1,20 -1,30	5	1,25	6,25	5	0,05	0,05
1,30 -1,40	18	1,35	24,30	23	0,18	0,23
1,40 -1,50	42	1,45	60,90	65	0,42	0,65
1,50 -1,60	27	1,55	41,85	92	0,27	0,92
1,60 -1,70	8	1,65	13,20	100	0,08	1,00
Total	100		146,50		1,00	

a) $\bar{x} = \dfrac{146,5}{100} = 1,465$ m.

Los intervalos son de la misma amplitud. La moda está en el intervalo 1,40-1,50; para la aproximación del valor exacto operamos con la expresión

$$M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c \quad \Rightarrow \quad M_o = 1,40 + 0,10\frac{27}{27+18} = 1,46 \text{ m.}$$

La mediana está también en el mismo intervalo ($N/2 = 50$; 65 es el primer valor de la frecuencia acumulada que supera a 50 y corresponde a este intervalo); para la aproximación del valor exacto operamos con la expresión

$$M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i \quad \Rightarrow \quad M_e = 1,40 + 0,10\frac{(50-23)}{42} = 1,46 \text{ m.}$$

b)

$$Q_1 = 1,40 + 0,10\frac{(25-23)}{(65-23)} = 1,40 + 0,005 = 1,405 \text{ m.}$$

$$Q_3 = 1,50 + 0,10\frac{(75-65)}{(92-65)} = 1,50 + 0,037 = 1,573 \text{ m.}$$

c)

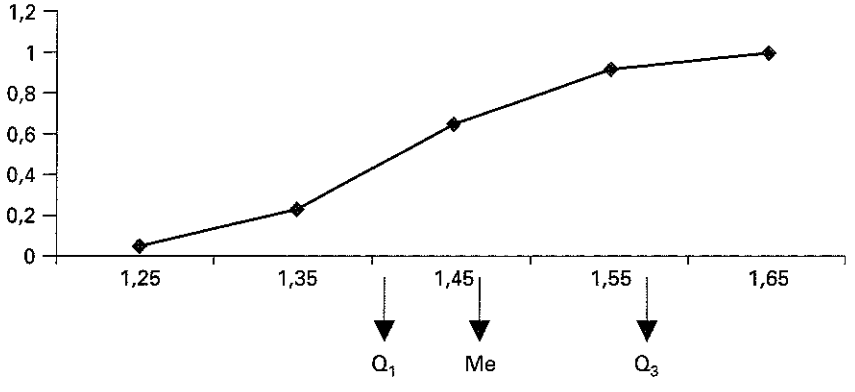


TABLA RESUMEN DEL CAPÍTULO 3:
Principales medidas de posición de una distribución unidimensional

Media Aritmética	Es la suma de todos los valores de la distribución dividida por el número total de observaciones.	$\overline{x} = \frac{1}{N}\sum_{i=1}^{r} x_i n_i$
Media Geométrica	Es la raíz N-ésima del producto de los N valores observados.	$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \ldots x_r^{n_r}}$
Media Armónica	Es la media aritmética de los inversos de los valores de la variable.	$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_r}{x_r}} = \frac{N}{\sum_{i=1}^{r} \frac{n_i}{x_i}}$
Mediana	En una distribución de frecuencias con los valores ordenados de menor a mayor, se denomina Mediana al valor de la variable que deja a su izquierda y a su derecha el mismo número de frecuencias, es decir aquel valor cuya frecuencia acumulada es $N/2$.	Cuando el n.º de valores de la variable es impar: $M_e = N/2$. Cuando el n.º de valores de la variable es par: M_e = media aritmética de los 2 términos centrales de la distribución. En las distribuciones agrupadas en intervalos: $M_e = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot c_i$
Moda	Es el valor de una variable que se repite más veces, es decir, aquel que tiene mayor frecuencia absoluta.	Para distribuciones agrupadas en intervalos: $M_o = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot c_i$
Cuartiles	Son los tres valores que dividen la distribución en cuatro partes iguales; cada parte incluye, pues, el 25% de los valores de la distribución.	C_1 = es el valor que ocupa el lugar $N/4$. C_2 = es el valor que ocupa el lugar $2N/4$. C_3 = es el valor que ocupa el lugar $3N/4$. Para distribuciones agrupadas en intervalos ($k = 4$; $r = 1, 2, 3$): $Q_{\frac{r}{q}} = L_{i-1} + \frac{\frac{rN}{K} - N_{i-1}}{n_i} \cdot c_i$

TABLA RESUMEN DEL CAPÍTULO 3:
Principales medidas de posición de una distribución unidimensional
(Continuación)

Deciles	Son los nueve valores que dividen la distribución en diez partes iguales; cada parte incluye, pues, el 10% de los valores de la distribución.	D_1 = es el valor que ocupa el lugar $N/10$. D_2 = es el valor que ocupa el lugar $2N/10$. [...] D_9 = es el valor que ocupa el lugar $9N/10$. Para distribuciones agrupadas en intervalos se emplearía la misma fórmula de los cuartiles en la que $k = 10$ y $r = 1, 2, \ldots, 9$.
Perceptiles	Son los 99 puntos o valores que dividen la distribución en cien partes iguales.	P_1 = es el valor que ocupa el lugar $N/100$. P_2 = es el valor que ocupa el lugar $2N/100$. [...] P_{99} = valor que ocupa el lugar $99N/100$. Para distribuciones agrupadas en intervalos se emplearía la misma fórmula de los cuartiles en la que $k = 100$ y $r = 1, 2, \ldots, 99$.

Capítulo 4

LAS MEDIDAS DE DISPERSIÓN, DE CONCENTRACIÓN Y DE FORMA EN UNA DISTRIBUCIÓN DE FRECUENCIAS UNIDIMENSIONAL

4.1. DEFINICIÓN Y CLASIFICACIÓN

En múltiples casos las medidas de posición son *insuficientes* para resumir las características fundamentales de una distribución; deben estudiarse, por ello, otro grupo de estadísticos, las medidas de dispersión, que son un complemento de las medidas de posición y que permiten medir lo más o menos esparcida que se encuentra la variable estadística entorno a la medidas de posición. *A mayor dispersión menor representatividad tienen las medidas de posición para describir la distribución de frecuencias y viceversa.*

Veamos un ejemplo:

Ejemplo 4.1. Analizar la edad de dos grupos de individuos representados en las distribuciones unitarias X e Y.

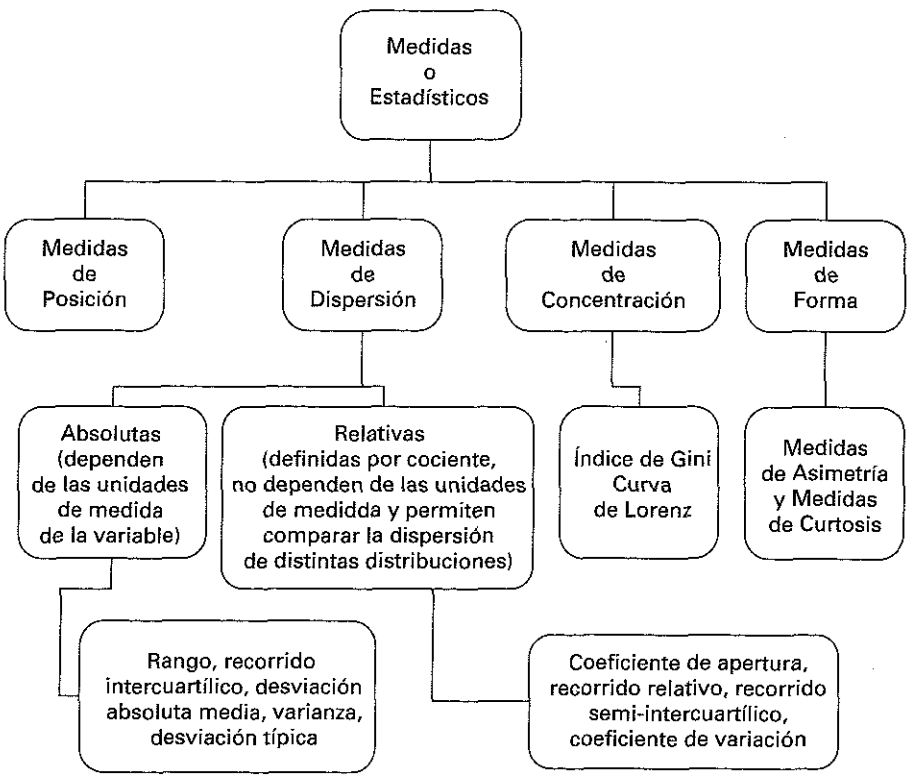
	x_i	y_i
Individuo 1	14	2
Individuo 2	16	4
Individuo 3	18	5
Individuo 4	20	39
Individuo 5	22	40
Suma de edades	90	90
Media aritmética	18	18

Aplicando las medidas de posición descritas hasta ahora, se concluiría el primer y el segundo grupo de individuos tienen la misma media de edad, pero es evidente que los dos grupos son muy diferentes entre sí; uno corresponde a un grupo de jóvenes, mientras que el otro podría corresponder a una familia con 3 hijos. Otras medidas de posición, como la mediana, la moda o los cuartiles, tampoco aportan información suficiente para conocer adecuadamente la distribución.

Se hace necesario, pues, ampliar la información aportada por las medidas de posición, investigando la «*distancia*» entre los valores de la distribución y los valores centrales; esta información la proporcionan las denominadas medidas de dispersión o de variabilidad.

En este capítulo veremos que también aportan información significativa sobre este extremo las denominadas medidas de concentración y las medidas de forma.

Veamos un esquema que nos clasifica este grupo de medidas, antes de pasar a realizar una breve descripción del concepto y significado de cada una de ellas.



4.2. LAS MEDIDAS DE DISPERSIÓN

Las principales medidas de dispersión son las siguientes:

4.2.1. Rango, recorrido o amplitud total de la distribución

En una distribución, con los valores previamente ordenados de menor a mayor, ***se define como la diferencia entre el mayor y el menor valor de la distribución.***

Se denota como R y se obtiene mediante la expresión:

$$R_x = x_r - x_1 = \max\{x_i\} - \min\{x_i\} \qquad \text{para } 1 \le i \le r;$$

O lo que es lo mismo:

$$\boxed{R_x = x_{(n)} - x_{(1)}}$$

En el Ejemplo 4.1: $R_x = 22 - 14 = 8$

$R_y = 40 - 2 = 38$

De lo que se deduce que la distribución de x_i tiene menor rango (8) que la de y_i (38).

4.2.2. Coeficiente de apertura

Se define como la relación entre el mayor y el menor valor de la distribución:

$$\boxed{C_{ap} = \frac{x_{(n)}}{x_{(1)}}}$$

En el Ejemplo 4.1:

$$C_{apx} = 22/14 = 1{,}57$$

$$C_{apy} = 40/2 = 20$$

Diríamos con ello que la distribución de x_i tiene menor apertura que la de y_i.

Las medidas anteriores tienen un interés limitado por estar construidas con los valores extremos; para amortiguar la influencia de estos valores se emplean las siguientes 4 medidas suavizadoras.

4.2.3. Recorrido intercuartílico

El **Recorrido o Rango intercuartílico** se define como la diferencia entre el tercer y el primer cuartil de la distribución.

$$\boxed{R_i = Q_3 - Q_1}$$

En el Ejemplo 4.1, con frecuencias unitarias tenemos, para la variable x:

	x_i	n_i	N_i	y_j	n_i	N_i
Individuo 1	14	1	1	2	1	1
Individuo 2	16	1	2	4	1	2
Individuo 3	18	1	3	5	1	3
Individuo 4	20	1	4	39	1	4
Individuo 5	22	1	5	40	1	5

Para x_i ,aplicando; $RR = \dfrac{rN}{q}$

Q_1 está en el valor cuya frecuencia supera a 5/4 = 1,25, o sea en $x_2 = 16$;
Q_3 está en el valor cuya frecuencia supera a 15/4 = 3,75, es decir en $x_q = 20$.

En consecuencia:

$$R_{i_x} = 20 - 16 = 4$$

Análogamente, para la variable y:

$$R_{i_j} = 39 - 4 = 35$$

4.2.4. Rango entre percentiles

El **Rango entre percentiles** se define como la diferencia entre el percentil 90 y el 10:

$$\boxed{R_p = P_{90} - P_{10}}$$

En el ejemplo 4.1:

Para x_i:

P_{90} está en el valor cuya frecuencia supera a 450/100 = 4,5, o sea en $x_5 = 22$;
P_{10} está en el valor cuya frecuencia supera a 50/100 = 0,5, es decir en $x_1 = 14$.

En consecuencia:

$$R_{p_x} = 22 - 14 = 8$$

Análogamente, para la variable y:

$$R_{p_y} = 40 - 2 = 38$$

4.2.5. Recorrido relativo

Se define como el cociente entre el recorrido y la media aritmética y expresa el número de veces que el recorrido contiene a la media aritmética; su formulación viene dada por la expresión:

$$RR_x = \frac{R_x}{\bar{x}}$$

$$RR_x = \frac{R_x}{\bar{x}} = \frac{8}{18} = 0,44 \qquad RR_y = \frac{R_y}{\bar{y}} = \frac{38}{18} = 2,11$$

4.2.6. Recorrido semi-intercuartílico

El **Recorrido semi-intercuartílico** queda definido como el cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil:

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

En el Ejemplo 4.1, tendríamos:

$$R_{si_x} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{20 - 16}{20 + 16} = 0,11 \qquad R_{si_y} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{39 - 4}{39 + 4} = 0,81$$

4.2.7. Desviación media

Se llama *desviación* a la diferencia entre el valor de la variable y la medida de posición central que se considere (media aritmética, geométrica, armónica, mediana o moda).

Cada valor de la variable tiene una desviación respecto a la media (d_i); la suma de todas estas desviaciones multiplicadas por sus respectivas frecuencias $[\sum_{i=1}^{r}(x_i - \bar{x})n_i = 0]$ es siempre cero (Véase el epígrafe 3.2.4).

Para obtener un estadístico de desviación media resolviendo esta dificultad operamos de tres formas, generándose tres medidas alternativas:

a) La **Desviación media** o *Desviación absoluta media:* se define como la media de los valores absolutos de las desviaciones respecto a la media aritmética; se denota como D_x, y viene dada por la expresión:

$$D_{\bar{x}} = \sum_{i=1}^{r} |x_i - \bar{x}| \frac{n_i}{N}$$

b) La **Desviación mediana,** viene dada por la siguiente expresión:

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

En la que la media se ha sustituido por la mediana.

La tercera de estas medidas o estadísticos es la varianza, que, por su importancia, tratamos en el epígrafe siguiente.

Ejemplo 4.2. Calcular la desviación media y la desviación mediana en una distribución de frecuencias de tipo I en la que la variable X_i toma los valores 1, 6, 8 y 9.

Los pasos a efectuar son los siguientes:

1. Se procede a calcular la media aritmética

$$\bar{x} = \frac{\sum_{i=1}^{i=4} x_i}{N} = \frac{24}{4} = 6$$

2. Se obtiene la columna $x_i - \bar{x}$ (sgunda columna de la Tabla adjunta), restando de la media (6) cada uno de los valores de x_i.
3. Se obtiene el valor absoluto de esta columna $|x_i - \bar{x}|$ (tercera columna de la Tabla adjunta).
4. Se obtiene la mediana: aplicando el criterio convencional, calculamos $N/2 = 2$, frecuencia que coincide con N_2; en consecuencia se toma como mediana la media de dicho valor y del siguiente: $(6 + 8)/2 = 7$.
5. Se obtiene $x_i - M_e$ (cuarta columna de la Tabla adjunta) y su valor absoluto $|x_i - M_e|$ (última columna de la Tabla adjunta).

	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$	$x_i - M_e$	$\lvert x_i - M_e \rvert$
	1	–5	5	–6	6
	6	0	0	–1	1
	8	2	2	1	1
	9	3	3	2	2
Suma	24	0	10	–4	10
Media aritmética	6				
Mediana	7				

6. Aplicando las oportunas formulaciones tendremos:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{N} = \frac{10}{4} = 2{,}5 \qquad D_{Me} = \frac{\sum_{i=1}^{n} |x_i - Me|}{N} = \frac{10}{4} = 2{,}5$$

En parecida forma se operaría con las distribuciones de tipo II y de tipo III.

4.2.8. La varianza

La varianza de una distribución se define como la media aritmética de los cuadrados de las desviaciones respecto a la media; se denota por s^2 o por σ^2 y su formulación es la siguiente:

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

Este estadístico se mide en el cuadrado de la unidad de la variable, por ejemplo, si la variable viene dada en segundos la varianza vendrá en segundos al cuadrado, lo que tiene una difícil interpretación.

4.2.9. La desviación típica

Para evitar este inconveniente suele operarse con la raíz cuadrada positiva de este estadístico; así, se conoce como ***desviación típica o desviación estándar*** (σ_x o s_x) *de una distribución a la raíz cuadrada* ***positiva*** *de la varianza.*

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

La desviación típica es la medida de dispersión más importante en estadística aplicada; una desviación típica pequeña significa que todos los valores de la distribución se sitúan próximos a la media, mientras que una desviación típica elevada implica la existencia de valores, por exceso o por defecto, muy alejados de la media.

Las principales *propiedades* de la desviación típica son:

— Es siempre mayor o igual que cero, ya que aunque la raíz cuadrada de la varianza tiene dos valores (positivo o negativo), por convenio se toma siempre el positivo.

- No está afectada por cambios de origen.
- Si que está afectada por cambios de escala, quedando multiplicada por el factor de escala cuando efectuamos un cambio de escala de la variable.

Junto a la varianza y a la desviación típica se utilizan 2 medidas parecidas, las denominadas **Cuasivarianza** y **Cuasidesviación típica,** cuya diferencia con las anteriores es que en sus formulaciones se divide por $N-1$ en vez de por N.

Para su notación hay criterios diferentes: algunos autores prefieren utilizar la letra griega sigma (σ) para la varianza y para la desviación típica y la letra s para la cuasivarianza y para la cuasidesviación típica; otros las indican con s_{n-1} o σ_{n-1}.

Vienen dada por las siguientes expresiones:

$$s_x^2 = \sum_{i=1}^{r} \frac{(x_i - \bar{x})^2 n_i}{N-1} \qquad s_x = \sqrt{\sum_{i=1}^{r} \frac{(x_i - \bar{x})^2 n_i}{N-1}}$$

En el nivel que se trata en este texto no nos detendremos en exceso en interpretar estos dos nuevos estadísticos; baste decir que su interés está relacionado con ***su importancia como estimadores***; cuando tomamos datos de una muestra estadísticamente representativa y queremos inferir resultados a la población total (inferencia estadística), la media de la muestra puede utilizarse como estimador de la media de la población total (puede demostrarse que la media de la muestra es un estimador insesgado de la media de la población); sin embargo, la varianza y la desviación típica de la muestra no son estimadores adecuados para inferir la varianza y la desviación típica de la población total; por el contrario, en cambio, puede demostrarse que los estimadores más adecuados para inferir la varianza y la desviación típica poblacional son, precisamente, la cuasivarianza y la cuasidesviación típica muestrales.

De esta forma, cuando estemos tratando con datos muestrales haremos mejor en obtener la cuasivarianza y la cuasidesviación típica y cuando estemos trabajando con datos poblaciones debemos obtener la varianza y la desviación típica.

Ejemplo 4.3. Cálculo de la varianza y de la desviación típica en la siguiente distribución de frecuencias de tipo I:

x_i
4
6
10
16

Se obtiene la media aritmética

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{N} = \frac{36}{4} = 9$$

y se procede al cálculo de la segunda $x_i - \bar{x}$ y de la tercera columna $(x_i - \bar{x})^2$, cuya suma, dividida por el número de elementos (4) proporciona la varianza.

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	4	–5	25
	6	–3	9
	10	1	1
	16	7	49
Suma	36	0	84
Varianza			21

Y tendremos:

$$\sigma_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N} = \frac{84}{4} = 21 \qquad s_x = \sigma_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N}} = \sqrt{21} \approx 4{,}58$$

Si consideramos que los valores corresponden a una muestra y que los cálculos se utilizan para obtener estimadores de los estadísticos poblacionales debemos operar con la cuasivarianza y la cuasidesviación típica, cuyos cálculos serían:

$$s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N-1} = \frac{84}{3} = 28 \qquad s_x = \sigma_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{N-1}} = \sqrt{28} \approx 5{,}29$$

Ejemplo 4.4. Cálculo de la varianza y de la desviación típica en la siguiente distribución de frecuencias de tipo II:

x_i	n_i
4	7
7	5
8	8
9	5
12	3
16	4
21	3
25	7
30	8

Obtenemos la siguiente tabla auxiliar:

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
4	7	28	–11	–77	121	847
7	5	35	–8	–40	64	320
8	8	64	–7	–56	49	392
9	5	45	–6	–30	36	180
12	3	36	–3	–9	9	27
16	4	64	1	4	1	4
21	3	63	6	18	36	108
25	7	175	10	70	100	700
30	8	240	15	120	225	1.800
Suma	50	750		0		4.378
Media aritmética		15				
Varianza						87,56
Desviación típica						9,36

En la que:

— En la primera y segunda columna tenemos los valores de la variable (x_i) y frecuencias (n_i).

— En la tercera columna, se construye, fila a fila, el producto necesario para obtener la media aritmética.

$$\bar{x} = \sum_{i=1}^{r} \frac{x_i n_i}{N} = \frac{750}{50} = 15.$$

— Conocida esta media aritmética, en la cuarta columna se obtiene, fila a fila, el valor de $(x_i - \bar{x})$; para ello se ha procedido a restar a cada valor de la variable la media aritmética de la distribución.
— En la quinta columna hemos realizado la operación $(x_i - \bar{x})n_i$, simplemente para comprobar que

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

— En la sexta columna, siempre trabajando fila a fila, se eleva al cuadrado el valor de la columna 4, obteniéndose con ello el valor $(x_i - \bar{x})^2$.
— Finalmente en la séptima columna se obtiene, también fila a fila, el valor de la expresión $(x_i - \bar{x})^2 n_i$; para ello multiplicamos el valor de la columna quinta por el valor de la columna segunda.
— La suma de la séptima columna dividida por 50 ($n = 50$), nos dará el valor de la varianza (87,56).

$$\sigma_x^2 = \sum_{i=1}^{r} \frac{(x_i - \bar{x})^2 n_i}{N} = \frac{4.378}{50} = 87,56.$$

La raíz cuadrada positiva de de la varianza nos dará la desviación típica:

$$s_x = \sqrt{87,56} \approx 9,36.$$

En las distribuciones de tipo III se opera en la misma forma tomando x_i como la marca de clase de los intervalos.

Cálculo de la varianza mediante los momentos

En el epígrafe 3.9 del capítulo anterior, se estableció el concepto estadístico de «momento» (definiendo los momentos respecto al origen, representados por la letra «*a*» y los momentos respecto a la media, representados por la letra «*m*»).

En aquel capítulo se avanzaba ya que al momento de orden 2 respecto a la media m_2, definido como $m_2 = \sum_{i=1}^{r} (x_i - \bar{x})^2 \frac{n_i}{N}$, se le denomina **varianza.**

Se indicaba asimismo una relación matemática que permite obtener los momentos de segundo orden en función de los momentos de primer orden y en que en el caso concreto de la varianza viene dada por la expresión:

$$m_2 = a_2 - a_1^2 = a_2 - \bar{x}^2 \qquad [4.2.9]$$

O lo que es lo mismo, *la varianza también se define como el momento de orden 2 respecto al origen menos la media aritmética elevada al cuadrado.*

Ejemplo 4.5. Obtener la varianza de la distribución del ejemplo 4.4. utilizando el método de los momentos:

Para aplicar este método necesitamos conocer

$$a_1 = \sum_{i=1}^{r} \frac{x_i n_i}{N} \qquad \text{y} \qquad a_2 = \sum_{i=1}^{r} \frac{x_i^2 n_i}{N}$$

Para ello construimos la siguiente tabla

x_i	n_i	$x_i \cdot n_i$	x_i^2	$x_i^2 \cdot n_i$
4	7	28	16	112
7	5	35	49	245
8	8	64	64	512
9	5	45	81	405
12	3	36	144	432
16	4	64	256	1.024
21	3	63	441	1.323
25	7	175	625	4.375
30	8	240	900	7.200
Suma	50	750	3017	15.628

$$a_1 = \sum_{i=1}^{r} \frac{x_i n_i}{N} = \frac{750}{50} = 15$$

$$a_2 = \sum_{i=1}^{r} \frac{x_i^2 n_i}{N} = \frac{15.628}{50} = 312{,}6$$

Varianza: $m_2 = a_2 - a_1^2 = 312{,}6 - 15^2 = 312{,}6 - 225 = 87{,}56$

En la que las columnas 1.ª, 2.ª y 3.ª coinciden con la tabla anterior y las columnas cuarta y quinta corresponden, respectivamente, al cuadrado de los valores de la variable y a este valor multiplicado por las frecuencias; aplicando la expresión [4.2.9.] obtenemos para la varianza el mismo resultado que el procedimiento anterior.

El alumno puede elegir para el cálculo de la varianza y de la desviación típica el procedimiento que le resulte más cómodo.

Como hemos indicado con anterioridad, tanto la varianza como la desviación típica vienen influidas por la unidad en la que se mide la variable, de forma que si cambiamos de unidad de medida, realizando un cambio de escala, como los indicados en el epígrafe 3.2.4 para la media aritmética, los valores de estos estadísticos se verían a su vez modificados.

Para eliminar la influencia de la unidad de medida y poder comparar la dispersión de dos distribuciones entre sí, se emplea, una medida «*escalar*», es decir, que no lleve asociado ninguna unidad de medida; la más empleada en Estadística es el Coeficiente de Variación de Pearson.

4.2.10. El Coeficiente de Variación de Pearson

El *Coeficiente de Variación de Pearson se define* como el *cociente entre la desviación típica y la media aritmética.* Representa, en consecuencia, el número de veces que la desviación típica contiene a la media

*S*e expresa indistintamente como

$$\gamma = \frac{\sigma}{\bar{x}}$$

O en porcentajes, como:

$$\gamma = \frac{\sigma}{\bar{x}} \cdot 100$$

Es una medida de dispersión relativa que permite comparar distribuciones diferentes, es decir, distribuciones que no vienen expresadas en las mismas medidas.

Este coeficiente es adimensional, ya que al venir expresada tanto la desviación típica como la media en la misma unidad y estar definido como un coeficiente, dicha unidad de medida queda anulada (simplificada).

El coeficiente no varía, por tanto, ante cambios de escala, pero sí ante cambios de origen. Para la interpretación de este ratio debe tenerse en cuenta, que:

— Si $\gamma = 0$ la representatividad de la media es máxima.
— Valores menores de la unidad indican que el promedio representa adecuadamente a la distribución de frecuencias, ya que la dispersión es inferior a la me-

dia aritmética; en concreto, a partir de $\gamma > 0{,}5$ podríamos considerar que la media tiene una baja representatividad.

— A partir de la unidad hay que rechazar el promedio (media aritmética) como parámetro representativo de los datos de la distribución.

Volviendo al Ejemplo 4.1, tendríamos:

	x_i	y_i
Individuo 1	14 años	2 años
Individuo 2	16 años	4 años
Individuo 3	18 años	5 años
Individuo 4	20 años	39 años
Individuo 5	22 años	40 años
Suma de edades	90 años	90 años
Media aritmética	18 años	18 años
Desviación típica	2,83 años	17,6 años
Coeficiente de variación de Pearson	0,16	0,98

Lo que aplicando el criterio anterior nos llevaría a concluir que la media aritmética no es una medida representativa para la variable y, mientras que sí tiene una representatividad adecuada para la variable x.

Nótese que mientras la variable y los estadísticos media aritmética y desviación típica van expresados en una unidad de medida (años) el Coeficiente de Variación de Pearson es una unidad adimensional.

4.3. MEDIDAS DE CONCENTRACIÓN

Las medidas o estadísticos de concentración miden el mayor o menor grado de equidad o igualdad en la distribución o reparto de los valores de una variable; se utilizan para analizar los aspectos redistributivos de variables como la renta, la riqueza, los salarios, etc.

Las posibilidades de concentración van desde la **concentración máxima** (cuando uno sólo individuo percibe el total y los demás nada —reparto no equitativo—, hasta la **concentración mínima,** cuando el total analizado está repartido por igual entre todos los valores de la variable —reparto equitativo—.

Las medidas de concentración más utilizadas son el índice de Gini y la curva de Lorenz.

4.3.1. Índice de Gini

Es la medida de concentración más utilizada; se calcula a través de la expresión:

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

Este índice toma valores comprendidos entre 0 y 1; toma el valor 0 cuando la variable está distribuida de forma muy homogénea y valor 1 cuando está muy concentrada (toda la renta está en las manos de un individuo).

Veamos su cálculo, significado e interpretación con el siguiente ejemplo:

Ejemplo 4.6. Estudiar la concentración de la masa salarial en dos empresas, cuya distribución de salarios es la siguiente:

Salario Mensual en euros L_{i-1} - L_i	N.º de trabajadores de la Empresa A	N.º de trabajadores de la empresa B
[500-900]	200	125
[900-1.300]	40	135
[1.300-1.700]	25	165
[1.700-2.100]	135	575

Obtenemos la siguiente tabla auxiliar para la empresa A:

L_{i-1} - L_i	n_i	N_i	$p_i = N_i/n$	x_i	$x_i \cdot n_i$	a_i = Acumulado de $x_i \cdot n_i$	q_i	$p_i - q_i$
[500-900]	200	200	0,500	700	140.000	140.000	0,292887	0,2071
[900-1.300]	40	240	0,600	1.100	44.000	184.000	0,384937	0,21506
[1.300-1.700]	25	265	0,663	1.500	37.500	221.500	0,463389	0,19911
[1.700-2.100]	135	400	1,000	1.900	256.500	478.000	1	0
Suma	400				478.000			0,62128
$\sum_{i=1}^{n-1} p_i$			1,763					

En la que:

— Se comprueba que los valores de la variable están ordenados de menor a mayor.
— Se calculan las frecuencias acumuladas N_i.
— Se obtiene la columna p_i dividiendo la columna N_i por el valor $N = 400$.
— Se calculan las marcas de clase (x_i) y el producto $x_i \cdot n_i$.
— Se obtiene el valor de la columna u_i de la siguiente forma:

- Para el primer intervalo (500-900] será $u_1 = x_1 \cdot n_1 \Rightarrow 700 \cdot 200 = 140.000$.
- Para el segundo intervalo (900-1.300] será: $u_2 = u_1 + x_2 \cdot n_2 \Rightarrow 140.000 + (1.100 \cdot 40) = 184.000$.
- Para el tercer intervalo (1.300-1.700] será: $u_3 = u_2 + x_3 \cdot n_3 \Rightarrow 184.000 + (1.500 \cdot 25) = 221.500$.
- Etc.

— Se obtiene la columna q_i dividiendo la columna u_i por el valor de la suma de la columna $x_i \cdot n_i$.
— Finalmente se obtiene la columna $p_i - q_i$, cuyo sumatorio constituye el numerador del cociente que da lugar al Índice de Gini.

$$IG_a = \frac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i} = \frac{0,621286}{0,5+0,6+0,663} = \frac{0,621286}{1,7625} = 0,35$$

Para la empresa B tendremos:

$L_{i-1} - L_i$	n_i	N_i	$p_i = N_i/n$	x_i	$x_i \cdot n_i$	u_i = Acumulado de $x_i \cdot n_i$	q_i	$p_i - q_i$
[500-900]	125	125	0,125	700	87.500	87.500	0,055520	0,069479
[900-1.300]	135	260	0,260	1.100	148.500	236.000	0,149746	0,110253
[1.300-1.700]	165	425	0,425	1.500	247.500	483.500	0,306789	0,118210
[1.700-2.100]	575	1.000	1,000	1.900	1.092.500	1.576.000	1	0
Suma	1.000				1.576.000			0,297944
$\sum_{i=1}^{n-1} p_i$			0,810					

De dónde:

$$IG_b = \frac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i} = \frac{0,297944}{0,81} = 0,3678$$

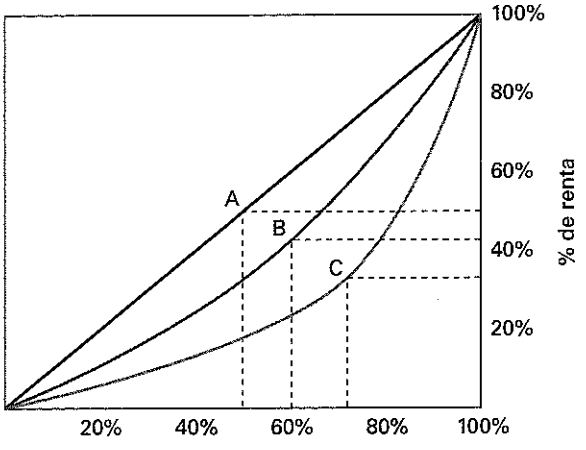
Como $IG_b > IG_a$ deducimos que la concentración de salarios es más alta en la empresa B que en la empresa A.

4.3.2. Curva de Lorenz

La curva de Lorenz es una forma gráfica de mostrar la dispersión o concentración de una distribución; tanto en abscisas como en ordenadas la gráfica parte del origen (0,0) y termina en el punto (100,100); sí la variable estuviese distribuida de forma equitativa la curva coincidiría con la línea de 45 grados que pasa por el origen, mientras que si es totalmente desigual (un valor concentra toda la masa de la distribución) la curva coincidiría con el eje horizontal hasta el punto (100,0) donde saltaría al punto (100,100); en general la curva se encuentra en una situación intermedia entre estos dos extremos.

Sí se trata de representar el grado de concentración de la renta en una población se representa en el eje de abscisas la población ordenada de forma que los percentiles de renta más baja quedan a la izquierda y los de renta más alta quedan a la derecha; en el eje de ordenadas se representan de abajo a arriba los percentiles acumulados de renta.

En el gráfico adjunto se representan 3 distribuciones de población; en la primera (línea diagonal A) la distribución de la renta es equitativa (el 50% de la población dis-



pone del 50% de la renta), en la segunda (línea B) la distribución es menos equitativa que en la primera (el 60% de la población sólo dispone de un 45% de la renta), pero más equitativa que en la tercera (dónde el 75% de la población apenas dispone de un 35% de la renta).

El Índice de Gini es aproximadamente el área comprendida entre la diagonal y la curva de Lorenz, dividida por el área del triangulo formado por los puntos (0, 0), (100, 0) y (100, 100).

Ejemplo 4.7. Obtener el Índice de Gini y la curva de Lorenz en una empresa que tiene la siguiente distribución de salarios de sus empleados (expresado en €):

x_i (Sueldos de los empleados, €)	n_i (N.º de personas)
850	20
1.000	12
1.250	5
2.000	3
3.500	2

Construimos la tabla que necesitamos para calcular el Índice:

$N_i^{\downarrow}$	$p_i = \frac{N_i^{\downarrow}}{N} \cdot 100$	$u_i = \sum_{j=1}^{i} x_j n_j$	$q_i = \frac{u_i}{u_r} \cdot 100$	$p_i - q_i$
20	47,62	17.000	35,23	12,39
32	76,19	29.000	60,10	16,09
37	88,10	35.250	73,06	15,04
40	95,24	41.250	85,49	9,75
42	100,00	48.250	100,00	0

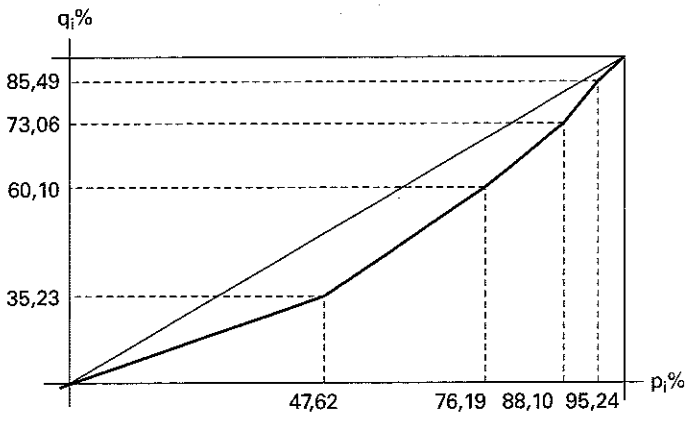
El Índice de Gini será:

$$I_g = \frac{(47,62 - 35,23) + (76,19 - 60,10) + (88,10 - 73,06) + (95,24 - 85,49)}{(47,62 + 76,19 + 88,10 + 95,24)} =$$

$$= \frac{53,26}{307,14} = 0,173$$

El valor del Índice es bastante próximo a 0, por lo que se puede afirmar que la renta está bastante distribuida entre todos los empleados.

Representamos gráficamente la curva de Lorenz:



4.4. LAS MEDIDAS DE FORMA

Las medidas de forma son de dos tipos: medidas de asimetría y medidas de curtosis.

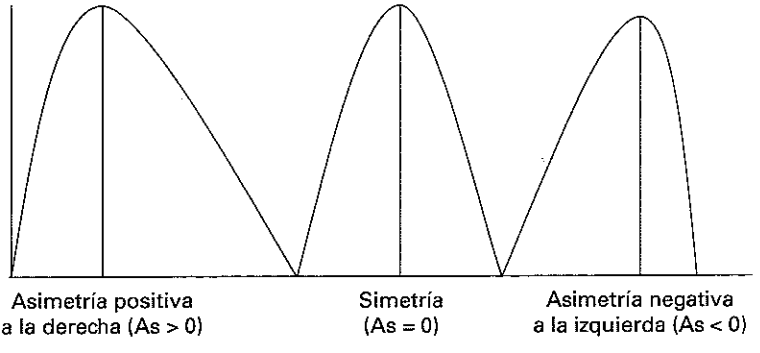
4.4.1. Medidas de asimetría

Son medidas que tratan de indicar el grado de simetría con el que se agrupan los valores de una distribución en torno a sus medidas centrales (generalmente la media aritmética o la mediana).

Decimos que una distribución es simétrica respecto a la media o respecto a la mediana sí al lado derecho de la misma queda la misma masa, es decir la misma cantidad de frecuencias que al lado izquierdo.

Visualmente decimos que una distribución es simétrica cuando el gráfico que la representa es simétrico respecto de la recta $x = \bar{x}$ o respecto a la recta $x = M_e$.

Gráficamente:



Cuando realizamos un estudio descriptivo es altamente improbable que la distribución de frecuencias sea totalmente simétrica, por lo que en la práctica diremos que la distribución de frecuencias es simétrica cuando lo es de un modo aproximado (Coeficiente de asimetría (As) próximo a 0).

El estadístico más empleado para obtener la asimetría es el **Coeficiente de asimetría de Fisher**; este estadístico está basado en el momento de tercer orden y se obtiene con la siguiente expresión:

$$g_1 = \frac{m_3}{\sigma^3} = \frac{\sum_{i=1}^{r}(x_i - \bar{x})^3 n_i}{N\sigma^3}$$

En la que σ es la desviación típica[5], de forma que:

$$g_1 = \frac{m_3}{g_3} = \frac{\frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})^3 n_i}{\left[\frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i\right]^{3/2}}$$

Si $g_1 = 0 \Rightarrow$ la distribución puede ser simétrica o no, pero si es simétrica se dará siempre que $g_1 = 0$.
Si $g_1 < 0 \Rightarrow$ la distribución es asimétrica a la izquierda.
Si $g_1 > 0 \Rightarrow$ la distribución es asimétrica a la derecha.

Algunos autores prefieren utilizar otras medidas de asimetría relacionadas con la mediana y/o con los cuartiles; algunas de las formulaciones propuestas vienen dadas por las siguientes expresiones:

Coeficiente de Asimetría de Pearson:	$A_p = \frac{\bar{x} - M_e}{s}$
Coeficiente de asimetría de Bowley:	$C_a = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}$
Coeficiente empleado por la hoja de cálculo Excel de Microsoft[3]:	$A = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$

[5] En vez de la desviación típica σ, algunos autores consideran que es mejor trabajar con la cuasidesviación típica, $s = \sqrt{\frac{\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i}{N-1}}$; formalmente debería trabajarse con la cuasidesviación típica si tenemos datos muestrales y con la desviación típica si los datos son poblaciones; en la práctica ambas fórmulas son equivalentes a efectos de la medición de la asimetría.

Valores nulos de estos coeficientes indican que puede tratarse de una distribución simétrica, mientras que valores positivos indican que la asimétrica es a la derecha de la media o de la mediana y valores negativos señalan una asimetría a la izquierda.

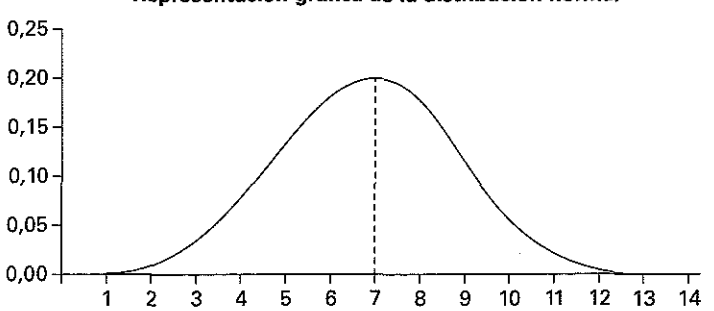
4.4.2. Medidas de apuntamiento o curtosis

Las medidas de apuntamiento o curtosis tratan de estudiar la distribución de frecuencias en la zona media, es decir, el mayor o menor número de valores de la variable alrededor de la media dará lugar a una distribución más o menos apuntada.

Para estudiar el apuntamiento hay que definir una distribución tipo que nos sirva de referencia. Esta distribución es conocida como la distribución Normal o la curva de Gauss y se corresponde con numerosos fenómenos de la naturaleza. Su forma es la de una campana en donde la gran mayoría de los valores se encuentran concentrados alrededor de la media, siendo escasos los valores que están, en ambos extremos, muy distanciados de ésta.

La representación gráfica de la distribución Normal es:

Representación gráfica de la distribución normal



Al tomar como referencia esta curva se dice que otra distribución es más apuntada que la distribución Normal (leptocúrtica) o menos apuntada (platicúrtica). A las distribuciones que se asemejan a la distribución Normal se les denomina mesocúrticas.

El **Coeficiente de Curtosis de Fisher** nos mide el grado de apuntamiento de la distribución sin necesidad de efectuar la representación gráfica de la misma. Este coeficiente está relacionado con el momento respecto a la media de cuarto orden m_4 y viene dado por la expresión:

$$g_2 = \frac{m_4}{\sigma_x^4} - 3$$

En la que recordemos que m_4 se obtiene mediante la expresión:

$$m_4 = \frac{\sum_{i=1}^{r}(x_i - \bar{x})^4 n_i}{N}$$

De forma que:

$$g_2 = \frac{\frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})^4 n_i}{\left[\frac{1}{N}\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i\right]^2} - 3$$

Si $g_2 = 0 \Rightarrow$ la distribución será mesocúrtica o Normal.
Si $g_2 < 0 \Rightarrow$ la distribución es platicúrtica o menos apuntada que la Normal.
Si $g_2 > 0 \Rightarrow$ la distribución es leptocúrtica o más apuntada que la Normal.

Existen otras formulaciones para obtener este coeficiente de curtosis; en concreto, el empleado por la hoja de cálculo Excel para medir la curtosis viene dado por la siguiente expresión:

$$C = \left\{\frac{N(N-1)}{(N-1)(N-2)(N-3)}\sum_{i=1}^{n}\left(\frac{(x_i - \bar{x})}{s}\right)^4\right\} - \frac{3(N-1)^2}{(N-2)(N-3)}$$

Que en la práctica es equivalente a la aquí formulada.

4.5. LAS MEDIDAS DE DISPERSIÓN, FORMA Y CONCENTRACIÓN EN HOJA DE CÁLCULO EXCEL Y EN SPSS

4.5.1. Las medidas de dispersión, forma y concentración en Excel

Siguiendo el esquema indicado en el capítulo 3, la siguiente Tabla muestra las funciones estadísticas para obtener las principales medidas de dispersión, forma y concentración estudiadas en este capítulo y su formulación en Excel:

Funciones estadísticas más empleadas en Excel

ESTADÍSTICA	FORMULACIÓN
Desviación estándar (muestra)	=DESVEST(rango)
Desviación estándar (población)	=DESVESTP(rango)
Varianza (muestra)	=VAR(rango)
Varianza (población)	=VARP(rango)
Curtosis	=CURTOSIS(rango)
Coeficiente de asimetría	=COEFICIENTE.ASIMETRÍA(rango)
Rango	=MAX(rango)-MIN(rango)
Mínimo	=MIN(rango)
Máximo	=MAX(rango)

Como hemos indicado en el epígrafe 3.10.1 para mayor comodidad puede utilizarse el mismo método de las Macro herramientas de Excel para obtener de forma conjunta este grupo de estadísticos.

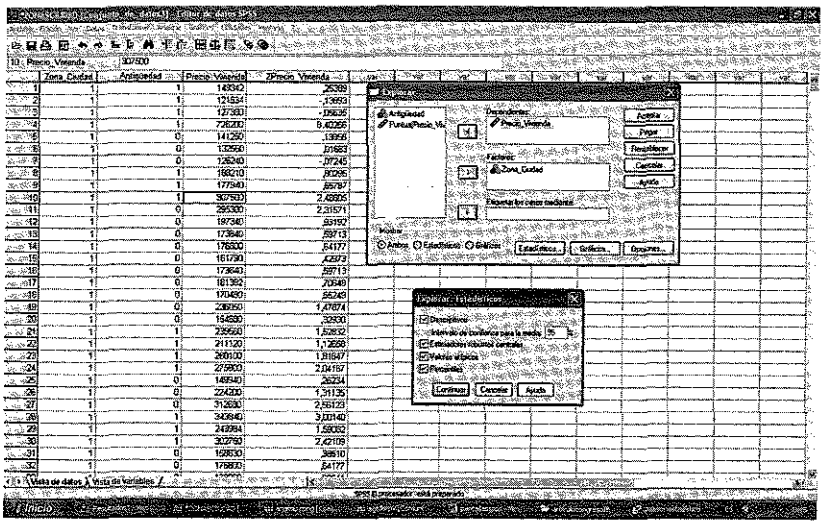
4.5.2. Las medidas de dispersión, forma y concentración en SPSS

Trabajando con sintaxis, SPSS utiliza las siguientes funciones de Estadística Descriptiva en cuanto a medidas de dispersión

SD(A)	Halla la desviación típica de la variable A
SD(A, B, C, ...)	Halla el vector de las desviaciones típicas de las variables A, B, C, ...
VARIANCE(A)	Halla la varianza de la variable A.
VARIANCE(A, B, C,...)	Halla el vector de las varianzas de las variables A, B, C, ...
CFVAR(A)	Halla el coeficiente de variación de la variable A.
CFVAR(A, B, C, ...)	Halla el vector de coeficientes de variación de las variables A, B, C, ...
MAX(A)	Halla el máximo de las observaciones de la variable A.
MAX(A, B, C, ...)	Halla el vector de los máximos de las observaciones de las variables A, B, C, ...
MIN(A)	Halla el mínimo de las observaciones de la variable A.
MIN(A, B, C, ...)	Halla el vector de los mínimos de las observaciones de las variables A, B, C, ...

Los estadísticos de dispersión, forma y concentración se obtienen de la forma ya indicada en el epígrafe 3.10.2; desarrollamos aquí, además, la opción «Explorar» que nos ofrece un conjunto de operaciones para representar gráficamente los datos, examinar visualmente sus distribuciones para varios grupos y otras opciones más complejas para las que más adelante explicaremos su interpretación.

Veamos un ejemplo:



Zona ciudad	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Precio vivienda						
Zona A	311	100,0%	0	,0%	311	100,0%
Zona B	252	100,0%	0	,0%	252	100,0%
Zona C	533	100,0%	0	,0%	533	100,0%
Zona D	1302	100,0%	0	,0%	1302	100,0%
Zona E	291	100,0%	0	,0%	291	100,0%
Zona F	200	100,0%	0	,0%	200	100,0%

Zona ciudad		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado (definición 1)	Precio Vivienda							
	Zona A	85180,40	99924,00	138282,00	200500,00	266860,00	325758,00	377460,00
	Zona B	70410,70	91498,20	120876,50	159200,00	228065,50	299013,00	382866,00
	Zona C	100723,00	113600,00	129321,00	149500,00	171620,00	196900,00	218230,00
	Zona D	37783,70	49802,00	69543,50	93850,00	117110,00	136702,00	151822,60
	Zona E	34244,80	49095,69	78121,90	113031,18	151815,63	179443,49	194054,72
	Zona F	56351,26	71247,78	103779,62	142828,59	174058,56	197050,40	226557,85
Bisagras de Tukey	Precio Vivienda							
	Zona A			138361,00	200500,00	265280,00		
	Zona B			120885,00	159200,00	227631,00		
	Zona C			129400,00	149500,00	171620,00		
	Zona D			69552,00	93850,00	117110,00		
	Zona E			76650,95	113031,18	151815,63		
	Zona F			104413,80	142828,59	173534,31		

DESCRIPTIVOS

	Zona ciudad		Estadístico	Error tip.
Precio_ vivienda	Zona A	Media	212464,01	5633,457
		Intervalo de confianza Límite inferior	201379,36	
		para la media al 95% Límite superior	223548,66	
		Media recortada al 5%	205960,84	
		Mediana	200500,00	
		Varianza	9,9E+009	
		Desv. tip.	99347,091	
		Mínimo	39340	
		Máximo	734100	
		Rango	694760	
		Amplitud intercuartil	128578	
		Asimetría	1,392	,138
		Curtosis	4,481	,276
	Zona B	Media	182624,67	5958,385
		Intervalo de confianza Límite inferior	170889,87	
		para la media al 95% Límite superior	194359,48	
		Media recortada al 5%	174960,51	
		Mediana	159200,00	
		Varianza	8,9E+009	
		Desv. tip.	94586,424	
		Mínimo	20380	
		Máximo	591900	
		Rango	571520	
		Amplitud intercuartil	107189	,153
		Asimetría	1,437	,306
		Curtosis	2,928	
	Zona C	Media	153270,26	1773,210
		Intervalo de confianza Límite inferior	149786,91	
		para la media al 95% Límite superior	156753,62	
		Media recortada al 5%	151434,58	
		Mediana	149500,00	
		Varianza	1,7E+009	
		Desv. tip.	40937,726	
		Mínimo	18800	
		Máximo	410042	
		Rango	391242	
		Amplitud intercuartil	42299	
		Asimetría	1,333	,106
		Curtosis	6,937	,211

DESCRIPTIVOS (continuación)

	Zona ciudad		Estadístico	Error tip.
Precio_ vivienda	Zona D	Media	94926,86	1102,052
		Intervalo de confianza Límite inferior	92764,87	
		para la media al 95% Límite superior	97088,86	
		Media recortada al 5%	93242,71	
		Mediana	93850,00	
		Varianza	1,6E+009	
		Desv. tip.	39765,617	
		Mínimo	12480	
		Máximo	403530	
		Rango	391050	
		Amplitud intercuartil	47567	
		Asimetría	1,651	,068
		Curtosis	9,054	,136
	Zona E	Media	115911,49	3109,154
		Intervalo de confianza Límite inferior	109792,12	
		para la media al 95% Límite superior	122030,86	
		Media recortada al 5%	113887,00	
		Mediana	113031,18	
		Varianza	2,8E+009	
		Desv. tip.	53038,186	
		Mínimo	20764	
		Máximo	421427	
		Rango	400663	
		Amplitud intercuartil	73694	
		Asimetría	,918	,143
		Curtosis	3,315	,285
	Zona F	Media	142042,78	3942,082
		Intervalo de confianza Límite inferior	134269,17	
		para la media al 95% Límite superior	149816,40	
		Media recortada al 5%	139218,11	
		Mediana	142828,59	
		Varianza	3,1E+009	
		Desv. tip.	55749,465	
		Mínimo	34245	
		Máximo	487631	
		Rango	453387	
		Amplitud intercuartil	70279	
		Asimetría	1,422	,172
		Curtosis	7,162	,342

VALORES EXTREMOS

	Zona Ciudad			Número del caso	Valor
Precio_Vivienda	Zona A	Mayores	1	303	734100
			2	4	726200
			3	291	595500
			4	210	536600
			5	207	484900
		Menores	1	239	39340
			2	267	50242
			3	238	50400
			4	234	50400
			5	233	50400
	Zona B	Mayores	1	442	591900
			2	512	552050
			3	489	523960
			4	510	505000
			5	443	489640
		Menores	1	563	20380
			2	371	23540
			3	530	34600
			4	520	34600
			5	379	34600
	Zona C	Mayores	1	631	410042
			2	629	388204
			3	632	345070
			4	651	331042
			5	630	311924
		Menores	1	817	18800
			2	815	21960
			3	963	26700
			4	842	40480
			5	858	45186
	Zona D	Mayores	1	1738	403530
			2	1392	379232
			3	1387	354900
			4	1391	354900
			5	1390	313820
		Menores	1	2188	12480
			2	1442	12480
			3	2067	12796
			4	2153	13270
			5	2097	13428a

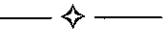
Valores extremos (continuación)

	Zona Ciudad			Número del caso	Valor
Precio_ Vivienda	Zona E	Mayores	1	2543	421427
			2	2446	301899
			3	2537	262057
			4	2501	256664
			5	2502	243643
		Menores	1	2491	20764
			2	2637	21363
			3	2679	22262
			4	2644	23760
			5	2588	24059
	Zona F	Mayores	1	2694	487631
			2	2746	348791
			3	2735	289916
			4	2798	261458
			5	2705	253828
		Menores	1	2781	34245
			2	2814	44730
			3	2693	51919
			4	2696	52219
			5	2702	53258

[a] En la tabla de valores extremos menores sólo se muestra una lista parcial de los casos con el valor 13428.

4.6. EJERCICIOS

Sobre medidas de dispersión, concentración y forma en distribuciones unidimensionales



Ejercicio 4.1. *Los siguientes datos representan los porcentajes del ingreso familiar asignados a la compra de alimentos en una muestra de 30 compradores.*

26 28 30 37 33 30 29 39 49 31 38 36 33 24 34
40 29 41 40 29 35 26 42 36 37 35 44 32 45 35

a) *Calcule el rango o amplitud.*

b) *Calcule el rango o recorrido* intercuartílico.

c) *Calcule el rango entre percentiles.*

d) *Calcule el recorrido relativo.*

e) *Calcule el recorrido semi-*intercuartílico.

Respuesta

Ordenamos los datos de menor a mayor.

24 26 26 28 29 29 29 30 30 31 32 33 33 34 35
35 35 36 36 37 37 38 39 40 40 41 42 44 45 49

Y procedemos a obtener la columna de frecuencias acumuladas:

i	x_i	f_i	F_i	i	x_i	f_i	F_i
1	24	1	1	11	36	2	19
2	26	2	3	12	37	2	21
3	28	1	4	13	38	1	22
4	29	3	7	14	39	1	23
5	30	2	9	15	40	2	25
6	31	1	10	16	41	1	26
7	32	1	11	17	42	1	27
8	33	2	13	18	44	1	28
9	34	1	14	19	45	1	29
10	35	3	17	20	49	1	30

a) Tomamos el valor máximo y el mínimo, para obtener, posteriormente la diferencia:

$$\left.\begin{array}{l} x_{máx} = 49 \\ x_{mín} = 24 \end{array}\right\} \Rightarrow R = x_{máx} - x_{mín} = 49 - 24 = 25$$

b) Se procede a calcular Q_1, obteniendo

$$\frac{1N}{4} = \frac{30}{4} = 7{,}5$$

el primer valor que supera a 7,5 en las frecuencias acumuladas es $F_5 = 9$; por lo tanto $Q_1 = 30$; en la misma forma se tiene, para calcular Q_3, que

$$\frac{3N}{4} = \frac{90}{4} = 22{,}5$$

el primer valor que supera el 22,5 en la columna de frecuencias acumuladas es el que tiene como frecuencia acumulada 23, es decir el 39.

Por lo tanto, $R_i = Q_3 - Q_1 = 9$.

c) Para calcular P_{10}, $k = 0{,}10 \Rightarrow nk = 30 \cdot 0{,}10 = 3$. Por lo tanto hay que hacer el promedio entre los valores situados en los lugares 3 y 4. Luego,

$$P_{10} = \frac{26 + 28}{2} = 27$$

Para calcular P_{90}, $k = 0{,}90 \Rightarrow nk = 30 \cdot 0{,}90 = 27$. Por lo tanto hay que hacer el promedio entre los valores entre los lugares 27 y 28. Luego,

$$P_{90} = \frac{42 + 44}{2} = 43$$

Concluimos que el rango entre percentiles es: $R_p = P_{90} - P_{10} = 43 - 27 = 16$.

d) El recorrido relativo es

$$\frac{R_i}{\bar{x}} = \frac{25}{34{,}76} = 0{,}72$$

e) El recorrido semi-intercuartílico es

$$R_{SI} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{9}{68} = 0{,}13$$

Ejercicio 4.2. *Con los datos del Ejercicio 4.1 calcule las desviaciones media y mediana.*

Respuesta

Como la media es 34,76 y la mediana 35 se obtienen:

x_i	n_i	$\lvert x_i-\bar{x}\rvert n_i$	$\lvert x_i-Me\rvert n_i$	x_i	n_i	$\lvert x_i-\bar{x}\rvert n_i$	$\lvert x_i-Me\rvert n_i$
24	1	10,77	11	36	2	2,47	2
26	2	17,53	18	37	2	4,47	4
28	1	6,77	7	38	1	3,23	3
29	3	17,30	18	39	1	4,23	4
30	2	9,53	10	40	2	10,47	10
31	1	3,77	4	41	1	6,23	6
32	1	2,77	3	42	1	7,23	7
33	2	3,53	4	44	1	9,23	9
34	1	0,77	1	45	1	10,23	10
35	3	0,70	0	49	1	14,23	14
				Suma		145,47	145

$$D_{\bar{x}}=\frac{1}{N}\sum_{i=1}^{r}\left|x_i-\bar{x}\right|n_i \quad\Rightarrow\quad D_{\bar{x}}=\frac{1}{30}\sum_{i=1}^{20}\left|x_i-\bar{x}\right|n_i=\frac{145{,}47}{30}=4{,}85$$

$$D_{Me}=\frac{1}{n}\sum_{i=1}^{r}\left|x_i-M_e\right|n_i \quad\Rightarrow\quad D_{Me}=\frac{1}{30}\sum_{i=1}^{20}\left|x_i-M_e\right|n_i=\frac{145}{30}=4{,}83$$

Ejercicio 4.3. *Calcule la varianza de los datos del ejercicio 4.1.*

Respuesta

Vamos a obtenemos la varianza mediante el método normal y mediante el método de los momentos.

a) Aplicamos la fórmula de la varianza $\sigma_x^2=\sum_{i=1}^{r}\frac{\left(x_i-\bar{x}\right)^2 n_i}{N}$ para lo que necesitamos obtener primero la columna $(x_i-\bar{x})n_i$ y luego la columna $(x_i-\bar{x})^2 n_i$.

x_i	n_i	$(x_i-\bar{x})\,n_i$	$(x_i-\bar{x})^2\,n_i$	x_i	n_i	$(x_i-\bar{x})n_i$	$(x_i-\bar{x})^2 n_i$
24	1	–10,77	115,92	36	2	2,47	3,04
26	2	–17,53	153,71	37	2	4,47	9,98
28	1	–6,77	45,79	38	1	3,23	10,45
29	3	–17,30	99,76	39	1	4,23	17,92
30	2	–9,53	45,44	40	2	10,47	54,78
31	1	–3,77	14,19	41	1	6,23	38,85
32	1	–2,77	7,65	42	1	7,23	52,32
33	2	–3,53	6,24	44	1	9,23	85,25
34	1	–0,77	0,59	45	1	10,23	104,72
35	3	0,70	0,16	49	1	14,23	202,59
				Suma		0	1.069,37

Donde la varianza, aplicando el método directo tomará el siguiente valor:

$$\sigma_x^2 = \sum_{i=1}^{20} \frac{(x_i-\bar{x})^2 n_i}{N} = \frac{1.069{,}37}{30} = 35{,}65$$

b) Aplicamos la fórmula de la varianza en función de los momentos respecto al origen: $\sigma^2 = m_2 = a_2 - a_1^2 = a_2 - \bar{x}^2$.

Conocemos que $a_1 = \bar{x} = 34{,}76$; por lo que $a_1^2 = 1.208{,}72$.

Necesitamos conocer

$$a_2 = \sum_{i=1}^{r} x_i^2 \frac{n_i}{N}$$

para ello obtenemos la columna $x_i^2 n_i$ y procedemos a su sumatorio:

x_i	n_i	$x_i^2 n_i$	x_i	n_i	$x_i^2 n_i$
24	1	576	36	2	2.592
26	2	1.352	37	2	2.738
28	1	784	38	1	1.444
29	3	2.523	39	1	1.521
30	2	1.800	40	2	3.200
31	1	961	41	1	1.681
32	1	1.024	42	1	1.764
33	2	2.178	44	1	1.936
34	1	1.156	45	1	2.025
35	3	3.675	49	1	2.401
			Suma		37.331

De donde se deduce que:

$$a_2 = \sum_{i=1}^{r} \frac{x_i^2 n_i}{N} = \frac{37.331}{30} = 1.244,37$$

Y, por consiguiente:

$$\sigma_x^2 = a_2 - \bar{x}^2 = 1.244,37 - 1.208,72 = 35,65$$

Ejercicio 4.4. *Teniendo en cuenta los datos de los Ejercicios 4.1 y 4.2, calcule la desviación estándar.*

Respuesta

Una vez conocida la varianza en el ejercicio anterior, se trata simplemente de obtener la raíz cuadrada positiva de este estadístico:

$$\sigma^2 = 35,65 \quad \Rightarrow \quad \sigma = \sqrt{\sigma^2} = \sqrt{35,65} = 5,97$$

Ejercicio 4.5. *Los siguientes valores corresponden a la facturación diaria, en miles de euros, durante 15 días consecutivos de un supermercado:*

20 25 22 20 25 20 21 22 22 24 23 20 23 20 25

Calcule la desviación estándar.

Respuesta

Obtenemos la media aritmética de los valores:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{\sum_{i=1} x_i}{N} = \frac{332}{15} = 22{,}13$$

Calculamos las columnas $x_i - \bar{x}$ y $(x_i - \bar{x})^2$.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	20	–2,13	4,55
	25	2,87	8,22
	22	–0,13	0,02
	20	–2,13	4,55
	25	2,87	8,22
	20	–2,13	4,55
	21	–1,13	1,28
	22	–0,13	0,02
	22	–0,13	0,02
	24	1,87	3,48
	23	0,87	0,75
	20	–2,13	4,55
	23	0,87	0,75
	20	–2,13	4,55
	25	2,87	8,22
Suma	332		53,73

Y aplicamos la fórmula de la desviación típica:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N}} = \sqrt{\frac{53{,}73}{15}} = 1{,}89$$

Dado que los datos pueden considerarse provenientes de una muestra, podíamos haber empleado la cuasivarianza y la cuasidesviación típica, dividiendo por n – 1 = 14;

en este caso:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{53,73}{14}} = 1,96$$

—— ✧ ——

Ejercicio 4.6. *Los quince edificios más altos de un ciudad tienen 47, 42, 43, 40, 38, 36, 33, 33, 26, 22, 27, 32, 27, 32 y 27 plantas.*

Se pide:

a) *Hallar el rango o amplitud.*

b) *Indicar la desviación estándar de la muestra.*

Respuesta

a) El rango se define como $R = x_{(r)} - x_{(1)} = \text{máx}\{x_i\} - \text{mín}\{_i\}$ para $1 \le i \le r$. Si ordenamos los valores en orden decreciente, tendremos que el máximo es 47 y el mínimo 22, por lo que:

$$R = x_{(r)} - x_{(1)} = 47 - 22 = 25 \text{ plantas.}$$

b) $$\sigma = \sqrt{\frac{\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{733,33}{15}} = 6,99$$

$s = 7,24$.

—— ✧ ——

Ejercicio 4.7. Un establecimiento de venta de equipo deportivo desea realizar una campaña publicitaria teniendo la edad de los clientes. Para ello, se recoge la edad, agrupada en intervalos, de los clientes durante el último mes, obteniéndose la siguiente tabla.

Edad en años	N.º de clientes
0 a 10	20
10 a 20	50
20 a 30	60
30 a 40	40
40 a 50	20
50 a 60	10
Total	200

Calcule la varianza de la variable edad.

Respuesta

Calculamos la columna $x_i \cdot n_i$ operando con las marcas de clase:

	x_i	n_i	$x_i \cdot n_i$
	5	20	100
	15	50	750
	25	60	1.500
	35	40	1.400
	45	20	900
	55	10	550
Suma		200	5.200

Y obtenemos la media aritmética

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{5.200}{200} = 26$$

Procedemos al cálculo de las columnas $(x_i - \bar{x})$, $(x_i - \bar{x})^2$ y $(x_i - \bar{x})^2 n_i$.

	x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
	5	20	100	–21	441	8.820
	15	50	750	–11	121	6.050
	25	60	1.500	–1	1	60
	35	40	1.400	9	81	3.240
	45	20	900	19	361	7.220
	55	10	550	29	841	8.410
Suma		200	5.200		1.846	33.800

Y aplicamos la formulación de la varianza

$$\sigma^2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_i}{N} = \frac{33.800}{200} = 169$$

La cuasivarianza será:

$$s^2 = \frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_i}{N-1} = \frac{33.800}{199} = 169,84$$

Como puede comprobarse, cuando las muestras son suficientemente grandes, la varianza y la cuasivarianza prácticamente coinciden.

Ejercicio 4.8. *Las estadísticas de la liga de baloncesto ponen de manifiesto, para los tiros de 3 puntos, que un jugador A obtiene una media del 40% por partido con una desviación estándar de 10. El jugador B obtiene un promedio del 60% de encestes, pero con una desviación estándar de 15. ¿Cuál de los dos jugadores tiene mayor variabilidad en los tiros de 3?*

Respuesta

Una alta desviación típica o estándar nos indica que el jugador obtiene su media con partidos en los que encesta un alto porcentaje y otros en los que su proporción de éxitos es muy reducida; por el contrario una baja desviación típica nos indica que el jugador obtiene en todos los partidos un porcentaje próximo a su media.

A primera vista parecería que el jugador B tiene una variabilidad mayor, ya que su desviación estándar (15) es mayor que la del jugador A (10), pero B obtiene un promedio mayor que A. Tomando en cuenta toda esta información, podemos calcular el coeficiente de variación para ambos jugadores.

Para A: $CV = \frac{10}{40} \cdot 100 = 25$

Para B: $CV = \frac{15}{60} \cdot 100 = 25$

Por lo tanto para ambos jugadores la variabilidad es del 25%. Luego, el jugador B, quien posee una variabilidad absoluta mayor que A, tiene una variación relativa similar debido a que su promedio de aciertos es mayor que el de A.

Ejercicio 4.9. *Una muestra de las edades de cinco mujeres, que asisten a una clase de ejercicios aeróbicos, permitió registrar las siguientes edades, en años, según el cumpleaños más próximo: 22, 18, 26, 20 y 24. Sus pesos en Kg. son: 52, 54, 56, 49 y 54. ¿Cuál de los dos conjuntos de datos es más variable?*

Respuesta

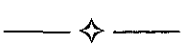
Para la edad se obtiene una media de 22 años y para los pesos de 53 Kg.

La desviación típica de la variable edad es 2,828 y la de los pesos 2,366; obteniendo el coeficiente de variación de Pearson:

Para la edad: $CV = \frac{2,828}{22} \cdot 100 = 12,86$

Para el peso: $CV = \frac{2,366}{53} \cdot 100 = 4,46$

El peso es menos variable que la edad, ya que 12,86 es mayor que 4,46 (coeficientes de variabilidad respectivos).



Ejercicio 4.10. *Calcule la varianza y la desviación estándar de la siguiente distribución de frecuencias.*

Clase	Frecuencia n_i
40 - 49	4
50 - 59	4
60 - 69	3
70 - 79	2
80 - 89	7

Respuesta

Vamos a obtener la varianza por el método de los momentos; calculamos las marcas de clase m_i y los sumatorios de las columnas $m_i \cdot n_i$ y $m_i^2 \cdot n_i$ que nos permiten el cálculo de los momentos de primer y segundo orden respeto al origen.

Clase	Marca de la clase m_i	Frecuencia n_i	$m_i \cdot n_i$	$m_i^2 \cdot n_i$
40 - 49	44,5	4	178	7.921
50 - 59	54,5	4	218	11.881
60 - 69	64,5	3	193,5	12.480,75
70 - 79	74,5	2	149	11.100,5
80 - 89	84,5	7	591,5	49.981,75
Suma		20	1.330	93.365

Recordemos que la varianza es el momento de segundo orden respecto a la media (m_2) y que la relación con los momentos de primer y segundo orden respeto al origen a_1 y a_2 viene dada por: $m_2 = a_2 - a_1^2$.

— a_1 es la media aritmética, que viene dada por la expresión:

$$a_1 = \sum_{i=1}^{r} \frac{x_i n_i}{N},$$

en nuestro caso, utilizando las marcas de clase, será

$$a_1 = \sum_{i=1}^{5} \frac{m_i n_i}{20} = \frac{1.330}{20} = 66,5.$$

— Por su parte a_2 viene dada por

$$a_2 = \sum_{i=1}^{5} \frac{m_i^2 n_i}{20} = \frac{93.365}{20} = 4.668,25.$$

De dónde $m_2 = a_2 - a_1^2 = 4.668,25 - (66,5)^2 = 246$.

La desviación típica vendrá dada por $\sigma = \sqrt{\sigma^2} = \sqrt{246} = 15,68$.

Si considerásemos que los datos proceden de una muestra y que en consecuencia nuestro mejor estimador no es la varianza sino la cuasivarianza o la cuasidesviación típica, tendríamos:

$$s^2 = \frac{1}{N-1}\left[\sum_{i=1}^{r} m_i^2 n_i - \frac{\left(\sum_{i=1}^{r} m_i n_i\right)^2}{N}\right] = \frac{1}{19}\left[93.365 - \frac{(1.330)^2}{20}\right] = 258{,}95$$

$$s = \sqrt{s^2} = 16{,}09$$

Ejercicio 4.11. *En la siguiente tabla se presentan los niveles de renta de los empleados de una compañía. Obtener el índice de Gini.*

Niveles de renta (€)	Cantidad de empleados
500 - 1.000	50
1.000 - 1.500	100
1.500 - 2.000	200
2.000 - 2.500	100
2.500 - 3.000	50

Respuesta

Marca de clase x_i	N.º de empleados n_i	N.º acum de empleados N_i	Total de salarios $n_i \cdot x_i$	$\mu_i = \sum x_i \cdot n_i$	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{\mu_i}{\mu_n} \cdot 100$
750	50	50	37.500	37.500	10	4,29
1.250	100	150	125.000	162.500	30	18,57
1.750	200	350	350.000	512.500	70	58,57
2.250	100	450	225.000	737.500	90	84,29
2.750	50	500	137.500	875.000	100	100
	$N = 500$		$_n = 875.000$			

$$I_G = \frac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i} = \frac{(10-4{,}29)+(30-18{,}57)+(70-58{,}57)+(90-84{,}29)}{200} = \frac{34{,}29}{200} = 0{,}17$$

Como el valor es cercano a cero, puede afirmarse que la variable se encuentra distribuida en forma homogénea.

Ejercicio 4.12. *En un muestreo se entrevistó a 10 asesores financieros sobre la rentabilidad de un determinado activo en el próximo año.*

Los datos en porcentaje son los siguientes:

15 16 17 24 18 29 30 14 37 24

a) *Calcule el coeficiente de asimetría de Pearson.*

b) *Calcule el coeficiente de asimetría de Fisher.*

c) *Calcule el coeficiente de curtosis.*

Respuesta

Primero obtenemos los valores de los siguientes estadísticos:

— $n = 10$.

— $\bar{x} = \sum \frac{x_i}{N} = \frac{224}{10} = 22{,}4$.

— $M_e = 21$; recordemos que para el cálculo de la Mediana debemos ordenar previamente la distribución y construir la tabla de frecuencias:

i	x_i	n_i	$x_i n_i$	N_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1	14	1	14	1	–8,4	70,56	–592,70	4.978,71
2	15	1	15	2	–7,4	54,76	–405,22	2.998,66
3	16	1	16	3	–6,4	40,96	–262,14	1.677,72
4	17	1	17	4	–5,4	29,16	–157,46	850,31
5	18	1	18	5	–4,4	19,36	–85,18	374,81
6	24	2	48	7	3,2	5,12	8,19	13,11
7	29	1	29	8	6,6	43,56	287,50	1.897,47
8	30	1	30	9	7,6	57,76	438,98	3.336,22
9	37	1	37	10	14,6	213,16	3.112,14	45.437,19
Suma		10	224			534,4	2.344,08	61.564,19

Al ser $N/2$ igual N_5, la mediana viene dada por la media de los valores correspondientes a x_5 y x_6, (18 y 24); por lo que:

$$M_e = \frac{18+24}{2} = 21.$$

— Al indicar en el enunciado que se trata de un muestreo, calculamos la cuasivarianza:

$$s^2 = \frac{\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i}{N} = \frac{534,4}{9} = 59,38$$

— $s = \sqrt{59,38} = 7,71$

— $s^3 = 7,71^3 = 457,54.$

— $\frac{1}{n}\sum(x_i - \bar{x})^3 n_i = \frac{2.344,08}{10} = 234,41$

— $\frac{1}{n}\sum(x_i - \bar{x})^4 n_i = \frac{61.564,19}{10} = 6.156,42$

Con esta información procedemos al cálculo de los estadísticos demandados:

a) El coeficiente de asimetría de Pearson viene dado por la expresión:

$$A_P = \frac{\bar{x} - M_e}{\sigma} = \frac{22,4-21}{7,71} = 0,18 > 0.$$

Este valor positivo indica que las observaciones tienen un sesgo hacia la derecha de la mediana.

b) El Coeficiente de asimetría de Fisher está basado en el momento de tercer orden y se obtiene con la expresión:

$$g_1 = \frac{\frac{1}{N}\sum(x_i - \bar{x})^3 n_i}{\left[\frac{1}{N}\sum(x_i - \bar{x})^2 n_i\right]^{3/2}} = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3 n_i}{s^3} = \frac{234,408}{457,54} = 0,51 > 0.$$

Este valor positivo indica que las observaciones tienen un sesgo hacia la derecha de la media.

c) Empleamos el coeficiente de curtosis de Fisher, que viene dado por la expresión:

$$g_2 = \frac{\frac{1}{N}\sum (x_i - \bar{x})^4 n_i}{\left[\frac{1}{N}\sum (x_i - \bar{x})^2 n_i\right]^2} - 3 = \frac{\frac{1}{N}\sum (x_i - \bar{x})^4 n_i}{(Var)^2} - 3 = \frac{6.156,42}{3.525,72} - 3 = 1,75 - 3 = -1,25$$

Este valor menor que 0 indica que tenemos una distribución con colas ligeras.

Ejercicio 4.13. *El cuadro que se presenta corresponde a la distribución de las compras realizadas por una empresa a sus diversos proveedores.*

Suministros en miles de euros	Cantidad de empresas
0 - 100	20
100 - 150	40
150 - 210	42
210 - 500	28

a) *Calcule el promedio de compras por empresa y analice su representatividad.*

b) *¿Cuál es la cantidad más frecuentemente suministrada?*

c) *¿Es simétrica la distribución? Justifique la respuesta.*

d) *Se prevé un descuento para aquellas empresas que hayan suministrado más de 145.000 euros. ¿Qué porcentaje de empresas podrá acogerse a esta bonificación?*

e) *¿Qué porcentajes de empresas han suministrado menos de 50.000 euros?*

Respuesta

Construimos la siguiente tabla auxiliar en la que hemos aproximado una marca de clase para cada intervalo:

Compras en miles de €	x_i (marca de clase)	n_i (número de empresas)	Frecuencia acumulada	$x_i \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$	$(x_i - \bar{x})^3 \cdot n_i$
0 - 100	50	20	20	1.000	342.011,83	–44.724.624,50
100 - 150	125	40	60	5.000	124.408,28	–6.938.154,30
150 - 210	180	42	102	7.560	24,85	–19,12
210 - 500	355	28	130	9.940	849.978,11	148.092.339,33
Total		130		23.500	1.316.423,08	96.429.541,42

a) La media viene dada por la siguiente expresión:

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{23.500}{130} = 180,769$$

El promedio o media aritmética de las compras es, pues, de 180.769 euros por empresa suministradora; para analizar su representatividad es necesario obtener alguna medida de dispersión; calculemos la desviación típica:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{1.316.423,08}{130}} \approx 100,63$$

Para relativizar la desviación típica y conocer su incidencia sobre la media, se calcula el coeficiente de desviación de Pearson, definido como el cociente entre la desviación típica y la media:

$$CV = \frac{\sigma_x}{\bar{x}} = \frac{100,63}{180,77} = 0,557$$

Este coeficiente es algo mayor, aunque no mucho mayor que 0,5; en consecuencia puede afirmarse que la media tiene una representatividad relativa, y, aunque no es totalmente descartable como medida de análisis de la distribución, hay que poner mucho cuidado en su interpretación.

b) El valor más frecuente de la distribución es la Moda; para el cálculo de este estadístico debe tenerse en cuenta que los intervalos son de diferente amplitud. Es necesario, en consecuencia, generar una columna de densidad del intervalo h_i.

Compras en miles de €	c_i (amplitud del intervalo)	n_i (número de empresas)	$h_i = n_i / c_i$
0 - 100	100	20	0,2
100 - 150	50	40	0,8
150 - 210	60	42	0,7
210 - 500	290	28	0,097

En la columna de densidad del intervalo h_i, puede observarse que el intervalo de mayor frecuencia relativa es el segundo; en consecuencia el intervalo modal es el intervalo 100-150.

Para el cálculo del punto modal se utiliza la expresión:

$$M_o = L_{\text{inf}} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot c_i \quad \Rightarrow \quad M_o = 100 + \frac{0,7}{0,2 + 0,7} \cdot 50 \cong 138.888 \text{ €}$$

c) Para estudiar la simetría de la serie calculamos el coeficiente de asimetría de Fisher, que viene dado por la expresión:

$$g_1 = \frac{\frac{1}{N}\sum_{i=1}^{r}\left(x_i - \bar{x}\right)^3 n_i}{\left(\sum_{i=1}^{r}\frac{\left(x_i - \bar{x}\right)^2 n_i}{N}\right)^{3/2}} = \frac{m_3}{\sigma_x^3} = \frac{\frac{1}{130} \cdot 96.429.541,42}{100,63^3} = \frac{741.765}{1.019.009} \cong 0,73$$

El momento de orden 3 respecto a la media (m_3) ha sido calculado a partir de las operaciones desarrolladas en la última columna de la tabla; en el denominador aparece la desviación típica calculada en un apartado anterior (100,63) elevada al cubo.

El valor 0,73 mayor que 0 indica que la distribución es asimétrica por la derecha, es decir que tiene más valores superiores a la media que los menores a la media.

d) Para obtener el porcentaje de las empresas que superaron los 145.000 euros de ventas podemos operar de la siguiente forma:

— Número de empresas que vende menos de 100.000 €: 20 = *n*.

— Número empresas que vende entre 100.000 y 150.000 €: 40

Si suponemos que estas 40 empresas están uniformemente distribuidas en el intervalo, podemos operar con una sencilla regla de tres para obtener las empresas que venden entre 100.000 y 145.000 euros; el razonamiento es el siguiente:

— Si en un intervalo de amplitud 50.000 hay 40 empresas...

— En un intervalo de amplitud 45.000 habrá X empresas.

De dónde

$$x = \frac{45.000 \cdot 40}{50.000} = 36 \text{ empresas}$$

El número total de empresas con menos de 145.000 € de ventas, es pues de 56 (20 +36); el porcentaje pedido será en consecuencia del 43,07% (56/130).

También se obtiene este resultado operando de la siguiente forma:

$$145 = 100 + \frac{k \cdot \frac{130}{100} - 20}{40} \cdot 50 \quad \Rightarrow k = 43{,}077$$

$100 - k = 56{,}923 \cong 56$.

e) De forma similar, para obtener el porcentaje de empresas que han vendido menos de 50.000 euros, trabajando ahora con el primer intervalo tendremos:

$$50 = 0 + \frac{k \cdot \frac{130}{100} - 0}{20} \cdot 100$$

$k = 7{,}6923$.

El 7,69% de las empresas vendieron menos de 50.000 €.

Ejercicio 4.14. *A continuación se presenta información referida a la distribución de sueldos del personal de una empresa según el tipo de empleado.*

Cuadro de distribución porcentual de Sueldos

Intervalos en euros	Personal Comercial	Personal Técnico	Total
0 - 400	14	0	14
400 - 800	17	8	25
800 - 1.200	18	18	36
1.200 - 1.600	13	14	27
1.600 - 2.000	8	14	22
2.000 - 2.400	14	19	33
2.400 - 2.800	9	18	27
2.800 - 10.000	7	9	16
Total	100	100	200

A partir de dicha información:

a) *¿Cuál es el importe promedio de las los sueldos? ¿Qué porcentaje ganan más, como media, cada uno de los grupos?*

b) *¿Es representativo el promedio total calculado?*

c) *¿Qué porcentaje de los empleados de cada grupo ganan entre 1.600 y 2.800 €?*

Respuesta

Construimos la tabla auxiliar:

$L_{i+1} - L_i$	Marca de clase x_i	c_i	$x_i \cdot n_{1i}$	$x_i \cdot n_{2i}$	$x_i \cdot n_i$
0 - 400	200	400	2.800	0	2.800
400 - 800	600	400	10.200	4.800	15.000
800 - 1.200	1.000	400	18.000	18.000	36.000
1.200 - 1.600	1.400	400	18.200	19.600	37.800
1.600 - 2.000	1.800	400	14.400	25.200	39.600
2.000 - 2.400	2.200	400	30.800	41.800	72.600
2.400 - 2.800	2.600	400	23.400	46.800	70.200
2.800 - 10.000	6.400	7.200	44.800	57.600	102.400
			162.600	213.800	376.400

En la que hemos calculado las marcas de clase de los intervalos y denominado x_1 y x_2 a los dos grupos de empleados (comercial y técnico).

a) Para obtener la media aritmética se tendrá que

$$\bar{x}_1 = \frac{\sum_{i=1}^{r} x_i n_{1i}}{N_1} = \frac{162.600}{100} = 1.626; \qquad \bar{x}_2 = \frac{\sum_{i=1}^{r} x_i n_{2i}}{N_2} = \frac{213.800}{100} = 2.138;$$

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{376.400}{200} = 1.882.$$

Como promedio, el personal técnico gana un 31,5% más que el personal comercial (2.138/1.626) y un 13% más que la media de todos los empleados de la compañía (2.138/1.882).

b) Para estudiar la representatividad de la media necesitamos conocer la desviación típica:

$L_{i-1} - L_i$	Marca de clase x_i	$x - \bar{x}$	n_i	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
0 - 400	200	–1.682	14	2.829.124	39.607.736
400 - 800	600	–1.282	25	1.643.524	41.088.100
800 - 1.200	1.000	–882	36	777.924	28.005.264
1.200 - 1.600	1.400	–482	27	232.324	6.272.748
1.600 - 2.000	1.800	–82	22	6.724	147.928
2.000 - 2.400	2.200	318	33	101.124	3.337.092
2.400 - 2.800	2.600	718	27	515.524	13.919.148
2.800 - 10.000	6.400	4.518	16	2.829.124	326.597.184
Total			200		458.975.200

$$\sigma = \sqrt{\frac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_i}{N}} = \sqrt{\frac{458.975.200}{200}} = 1.115$$

Para relativizar la desviación típica y conocer su incidencia sobre la media se calcula el coeficiente de variación de Pearson, definido como el cociente entre la desviación típica y la media:

$$CV = \frac{\sigma_x}{\bar{x}} = \frac{1.115}{1.882} = 0{,}8$$

Este coeficiente es mayor que 0,5; en consecuencia debe afirmarse que la media no tiene representatividad.

c) Como 1.600 y 2.800 son limites de intervalos, la respuesta se obtiene calculando las frecuencias relativas acumuladas entre los intervalos afectados.
Los resultados son:

— Personal comercial: 31%.
— Personal técnico: 51%.
— Total personal: 41%

Ejercicio 4.15. *En un aeropuerto se han tomado en promedio los retrasos en la partida de los últimos 800 vuelos. Calcule todas las medidas de dispersión a partir de la siguiente información:*

Retraso en minutos (x_i)	10	40	50	70
Nº de vuelos (n_i)	200	300	200	100

Respuesta

Construimos la tabla auxiliar:

x_i	n_i	N_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	$\lvert x_i - M_e \rvert n_i$	$\lvert x_i - \bar{x} \rvert n_i$
10	200	200	2.000	20.000	6.000	5.750
40	300	500	12.000	480.000	0	375
50	200	700	10.000	500.000	2.000	2.250
70	100	800	7.000	490.000	300	3.125
Suma	800		31.000	1.490.000	11.000	11.500

En la que se ha obtenido la Mediana a partir de la columna N_i de frecuencias absolutas acumuladas:

$$\frac{N}{2} = 400 \quad \Rightarrow \quad N_i > 400 \quad \Rightarrow \quad x_i = 40 \quad \Rightarrow \quad M_e = 40$$

Recordemos que para determinar la mediana en las distribuciones no unitarias se ordena la distribución, se obtiene la frecuencia absoluta acumulada N_i, y se calcula el valor $N/2$ (en nuestro caso 400). Si existe un N_i que supera al $N/2$ la mediana es el X_i que corresponde a dicho N.

La **media aritmética** viene dada por

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{31.000}{800} = 38,75$$

— **Rango**

$$R = X_{(n)} - X_{(1)} = 70 - 10 = 60$$

— **Recorrido intercuartílico**

$$R_1 = Q_3 - Q_1$$

$$\frac{qN}{4} = \frac{3 \cdot 800}{4} = 600 \quad \Rightarrow \quad Q_3 = 50$$

$$\frac{qN}{4} = \frac{1 \cdot 800}{4} = 200 \Rightarrow Q_1 = \frac{10 + 40}{2} = 25$$

$$R_1 = 50 - 25 = 25$$

— **Rango entre percentiles**

$$R_p = P_{90} - P_{10}$$

$$\frac{90 \cdot 800}{100} = 720 \quad \Rightarrow \quad P_{90} = 70$$

$$\frac{10 \cdot 800}{100} = 80 \quad \Rightarrow \quad P_{10} = 10$$

$$R_p = 70 - 10 = 60$$

— **Desviación absoluta media**

$$D_{\bar{x}} = \frac{\sum |x_i - \bar{x}| n_i}{N} \quad \Rightarrow \quad D_{\bar{x}} = \frac{11.500}{800} = 14,375$$

— **Desviación mediana**

$$D_{M_e} = \frac{\sum |x_i - M_e| n_i}{N} \quad \Rightarrow \quad D_{M_e} = \frac{11.000}{800} = 13,75$$

— **Varianza**

Aplicamos la fórmula de la varianza en función de los momentos respecto al origen:

$$\sigma^2 = m_2 = a_2 - a_1^2 = a_2 - \bar{x}^2$$

Conocemos que $a_1 = \bar{x} = 38,75$; necesitamos conocer a_2; en nuestro caso

$$a_2 = \sum_{i=1}^{r} \frac{x_i^2 n_i}{N} = \frac{1.490.000}{800} = 1.862,5$$

por lo que

$$\sigma_x^2 = 1.862,5 - (38,75)^2 = 360,9375$$

— **Desviación típica**

$$\sigma_x = \sqrt{\sigma_x^2} \quad \Rightarrow \quad \sigma_x = \sqrt{360,9375} = 18,998$$

— **Coeficiente de apertura**

$$C_{ap} = \frac{X_{(n)}}{X_{(1)}} \quad \Rightarrow \quad C_{ap} \frac{70}{10} = 7$$

— **Recorrido relativo**

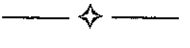
$$RR = \frac{R}{\bar{x}} \quad \Rightarrow \quad RR = \frac{60}{38,75} = 1,548$$

— **Recorrido semi-intercuartílico**

$$R_{si} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \quad \Rightarrow \quad R_{si} = \frac{50 - 10}{50 + 10} \cong 0{,}667$$

— **Coeficiente de variación de Pearson**

$$CV = \frac{\sigma_x}{\bar{x}} \quad \Rightarrow \quad CV = \frac{18{,}998}{38{,}75} \cong 0{,}49027$$



Ejercicio 4.16. *Se desea comparar dos métodos, A y B para enseñar a utilizar un nuevo sistema de producción en serie. Para llevar a cabo esta investigación, un especialista seleccionó a diez operarios con el mismo nivel de cualificación. A cada par de operarios se le asignó al azar un método (a uno el A y a otro al método B).*

Después de cuatro semanas cada operario se sometió a un examen, siendo los resultados obtenidos, los siguientes:

Par N.º	Método A	Método B
1	36	35
2	37	35
3	41	40
4	42	41
5	36	36
6	35	34
7	42	40
8	33	31
9	40	39
10	38	37

Si usted tuviera que recomendar uno de estos métodos, ¿cuál elegiría? Fundamente su respuesta.

Respuesta

	Método *A*	Método *B*
$\bar{x}$	38	36,8
σ_x	2,97	3,03
CV	7,82%	8,23%

Se recomendaría el método *A* pues, teniendo las variabilidades parecidas se observa un mayor promedio en las pruebas de rendimiento en los operarios que utilizaron ese método.

TABLA RESUMEN DEL CAPÍTULO 4:
Principales medidas de dispersión de una distribución unidimensional

Rango, Recorrido o Amplitud	Es la diferencia entre el mayor valor y el menor de una distribución.	$R = x_{(n)} - x_{(1)}$
Coeficiente de apertura	Otra forma de medir el rango.	$C_{ap} = \frac{X_{(n)}}{X_{(1)}}$
Rango *intercuartílico* o entrepercentiles	Diferencia entre cuartiles o perceptiles.	$R_i = Q_3 - Q_1$ $R_{ip} = P_{90} - P_{10}$
Recorrido relativo	Cociente entre el recorrido y la media aritmética.	$RR = \frac{R}{\bar{x}}$
Recorrido semi-*intercuartílico*	Cociente entre el recorrido *intercuartílico* y la suma del primer y tercer cuartil.	$R_{si} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$
Desviación media	Media de los valores absolutos de las desviaciones respecto a la media aritmética.	$D_{\bar{x}} = \sum_{i=1}^{r} \lvert x_i - \bar{x} \rvert \frac{n_i}{N}$
Desviación mediana	Media de los valores absolutos de las desviaciones respecto a la mediana.	$D_{\bar{x}} = \sum_{i=1}^{r} \lvert x_i - M_e \rvert \frac{n_i}{N}$
Varianza	Media aritmética de los cuadrados de las desviaciones respecto a la media.	$\sigma_x^2 = \sum_{i=1}^{r} (x_i - \bar{x})^2 \frac{n_i}{N}$
Desviación típica	Raíz cuadrada de la varianza.	$\sigma_x = \sqrt{\sigma_x^2}$
Cuasivarianza y cuasi-desviación típica	Varianza y desviación típica aplicados a datos muestrales; se divide por $n - 1$ en vez de por n.	$s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{n_i}{N-1}$ $s_x = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2 n_i}{N-1}}$
Coeficiente de Variación de Pearson	Cociente entre la desviación típica y la media aritmética.	$\gamma = \frac{\sigma}{\bar{X}}$

TABLA RESUMEN DEL CAPÍTULO 4:
Principales medidas de dispersión de una distribución unidimensional
(Continuación)

Índice de Gini	Mide el mayor o menor grado de equidad en el reparto de las variables y varía entre 0 y 1.	$I_G = \dfrac{\sum_{i=1}^{r-1}(p_i - q_i)}{\sum_{i=1}^{r-1} p_i}$
Coeficiente de asimetría de Fisher	Basado en el momento respecto a la media de tercer orden: Distribución simétrica $\Rightarrow$ $\Rightarrow g_1 = 0$. Si $g_1 < 0 \Rightarrow$ distribución asimétrica a la izquierda. Si $g_1 > 0 \Rightarrow$ distribución asimétrica a la derecha.	$g_1 = \dfrac{m_3}{\sigma^3} = \dfrac{\sum_{i=1}^{r}(x_i - \bar{x})^3 n_i}{N\sigma^3}$
Coeficiente de Curtosis de Fisher	Basado en el momento respecto a la media de cuarto orden: Si $g_2 = 0 \Rightarrow$ distribución normal. Si $g_2 < 0 \Rightarrow$ distribución Platicúrtica. Si $g_2 > 0 \Rightarrow$ distribución Leptocúrtica.	$g_2 = \dfrac{m_4}{\sigma_x^4} - 3$

Capítulo 5

DISTRIBUCIÓN DE FRECUENCIAS BIDIMENSIONALES

5.1. INTRODUCCIÓN

En la práctica es muy frecuente que en el estudio de una población estemos interesados en medir no sólo una, sino varias variables; *cuando estudiamos dos variables de una población tenemos una* ***distribución de frecuencias bidimensional***, sí estudiamos múltiples variables dispondremos de una ***distribución de frecuencias multidimensional.***

Ejemplo, en la caracterización de la población que visita una determinada ciudad, podemos estar interesados en conocer la nacionalidad y el nivel de renta de los visitantes (distribución bidimensional), pero también y además, el motivo de la visita, la duración de la estancia, el medio de locomoción empleado para llegar a la ciudad, el gasto realizado en su estancia, la distribución del gasto por conceptos, etc. (distribución multidimensional). Con lo estudiado en los capítulos anteriores, nos limitaríamos a analizar cada variable de forma aislada; en este capítulo se aborda el análisis conjunto de dos variables, poniendo especial hincapié en determinar la posible existencia de alguna relación de dependencia entre ellas.

Centrándonos en las distribuciones bidimensionales, tendremos, para cada individuo observado los valores correspondientes a dos variables o dos atributos, que denotamos por X e Y.

La posibilidad de que la información observada se corresponda con un valor o con un atributo, nos da varios tipos de distribuciones bidimensionales, a saber:

— Las dos variables son cualitativas (es decir, se trata de dos atributos, cada uno con sus diversas modalidades); medimos, por ejemplo, la nacionalidad y el motivo de la visita.
— Una de las dos variables es cuantitativa (ya sea discreta o continua) y otra cualitativa; medimos, por ejemplo, el gasto realizado y el medio de locomoción empleado.
— Las dos variables son cuantitativas (discretas o continuas); por ejemplo, el gasto realizado y la duración de la estancia.

5.2. CONSTRUCCIÓN DE TABLAS ESTADÍSTICAS BIDIMENSIONALES

Se llama distribución conjunta de frecuencias de las dos variables (x, y) a la tabla que representa los valores observados de ambas variables y sus frecuencias de aparición.

Cuando las variables son cuantitativas a las tablas de frecuencias se les denomina ***Tablas de Correlación*** y cuando se trata de atributos o variables cualitativas se las denomina ***Tablas de Contingencia.***

Las distribuciones bidimensionales adoptan el siguiente ***formato general***:

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

Dónde:

x_1, x_2 ....., x_r Son los r valores o modalidades que toma la variable x.

y_1, y_2 ....., y_s Son los s valores o modalidades que toma la variable y.

n_{i1}, n_{i2} ....., n_{is} Es la frecuencia en la que aparece el valor i de la variable X conjuntamente con cada valor 1, 2, .. s de la variable Y.

n_{1j}, n_{2j} ....., n_{rj} Es la frecuencia en la que aparece el valor j de la variable Y conjuntamente con cada valor 1, 2, .. r de la variable X.

$n_{i\cdot}$ Es la frecuencia total con la que aparece el valor i de la variable X.

$n_{\cdot j}$ Es la frecuencia total con la que aparece el valor j de la variable Y.

$N = n...$ Es la frecuencia total de la distribución.

Ejemplo 5.1. Se muestra en la siguiente distribución el número de empresas existentes en España según condición jurídica y estrato de asalariados a 1 de enero de 2009, según el Directorio Central de Empresas del Instituto Nacional de Estadística:

Estrato de asalariados	Sociedades anónimas	Sociedades limitadas	Comunidades de bienes	Personas físicas	Otros tipos	Total
Sin asalariados	25.680	358.446	64.759	1.217.370	101.215	1.767.470
De 1 a 9	41.938	664.012	48.604	569.557	78.885	1.402.996
De 10 a 19	15.334	71.951	1.168	5.943	7.205	101.601
De 20 a 49	14.825	34.264	254	1.027	5.271	55.641
De 50 a 99	5.581	7.212	33	0	2.249	15.075
De 100 a 499	4.958	4.527	12	0	1.753	11.250
De 500 o más	1.014	408	1	0	374	1.797
Total	109.330	1.140.820	114.831	1.793.897	196.952	3.355.830

Fuente: DIRCE. INE.

X_i Es el atributo Estrato de asalariados, que toma las modalidades x_1 = Sin asalariados, x_2 = De 1 a 9 ....., x_r = De 500 o más.

Y_j Es el atributo Condición o Forma Jurídica de la empresa, que toma las modalidades y_1 = Sociedades anónimas, y_2 = Sociedades limitadas ....., y_s = Otros tipos.

$n_{11}, n_{12}..., n_{1s}$ Es la frecuencia en la que aparece el atributo 1 de la variable X (Sin asalariados) conjuntamente con cada uno de los atributos de y (n_{11} = 25.680 (25.680 empresas que operan bajo la forma de sociedades anónimas, n_{12} ....., n_{1s} = 374 (374 empresas con forma jurídica otros tipos).

$n_{11}, n_{21}..., n_{r1}$ Es la frecuencia en la que aparece el atributo 1 de Y (Sociedades Anónimas) conjuntamente con los distintos atributos 1, 2, .. s de la variable X (n_{11} = 25.680 n_{21} = 41.938, ..., n_{r1} = 1.014, de 500 o más asalariados con forma jurídica sociedad anónima).

$n_{i\cdot}$ Es la frecuencia total con la que aparece la modalidad i de la variable X ($n_{1\cdot}$ = 1.767.470, es decir que el número total de empresas sin asalariados fue de 1.767.470).

$n_{\cdot j}$ Es la frecuencia total con la que la modalidad j de la variable Y ($n_{\cdot 2}$ = 1.140.820, indicativa de que el número de empresas con forma jurídica de sociedad limitada fue de 1.140.820 empresas).

$N = n_{\cdot\cdot}$ Es la frecuencia total de la distribución, en el ejemplo 3.355.830 empresas.

Las ***distribuciones marginales*** aparecen cuando se estudian aisladamente cada una de las variables (con independencia de la otra o, sí se trata de distribuciones mul-

tidimensionales, del resto de las variables). Así, al igual que en las distribuciones unidimensionales, podemos definir la ***frecuencia relativa*** de un elemento (x_i, y_j) en base a la relación:

$$f_{ij} = \frac{n_{ij}}{n..}$$

Como puede comprobarse, se verifica que la suma de todas las frecuencias relativas es igual a 1, es decir:

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

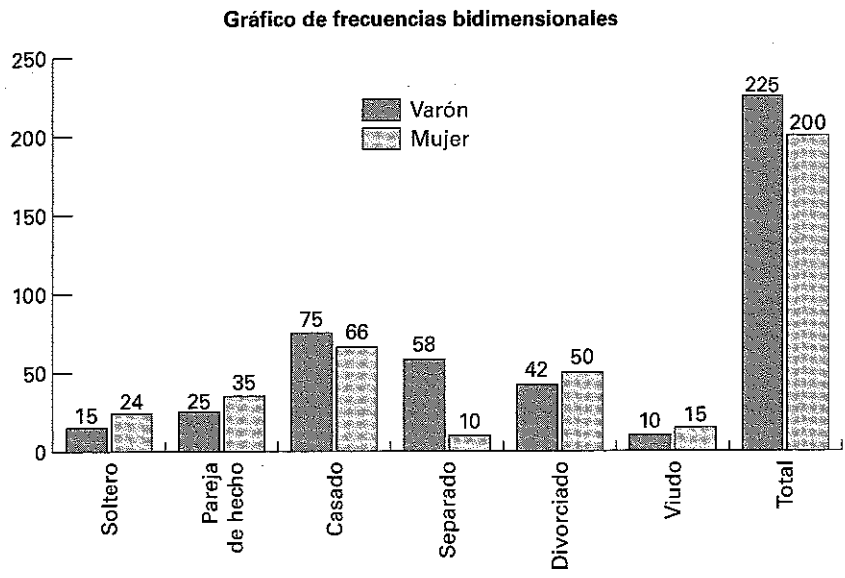
5.3. REPRESENTACIÓN GRÁFICA DE LAS DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES

Al igual que en lo descrito para las distribuciones unidimensionales las distribuciones bidimensionales pueden adoptar múltiples representaciones gráficas; dependiendo de los datos disponibles algunas pueden ser más plásticas que otras; veamos algunos ejemplos:

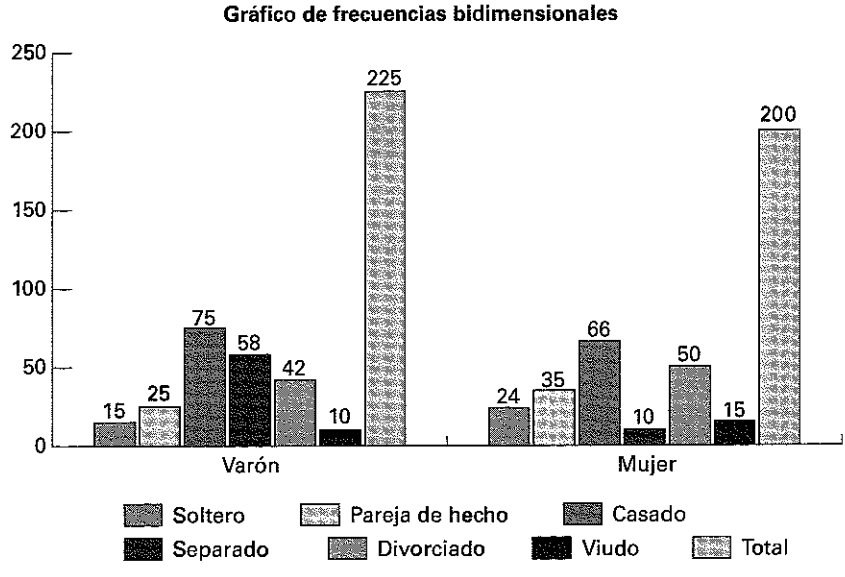
Ejemplo 5.2. Representar gráficamente la siguiente distribución bidimensional de frecuencias sobre el sexo y el estado civil de 425 personas.

Sexo / Estado civil	Varón	Mujer
Soltero	15	24
Pareja de hecho	25	35
Casado	75	66
Separado	58	10
Divorciado	42	50
Viudo	10	15
Total	225	200

Podemos representarlos gráficamente en cualquiera de las siguientes formas:



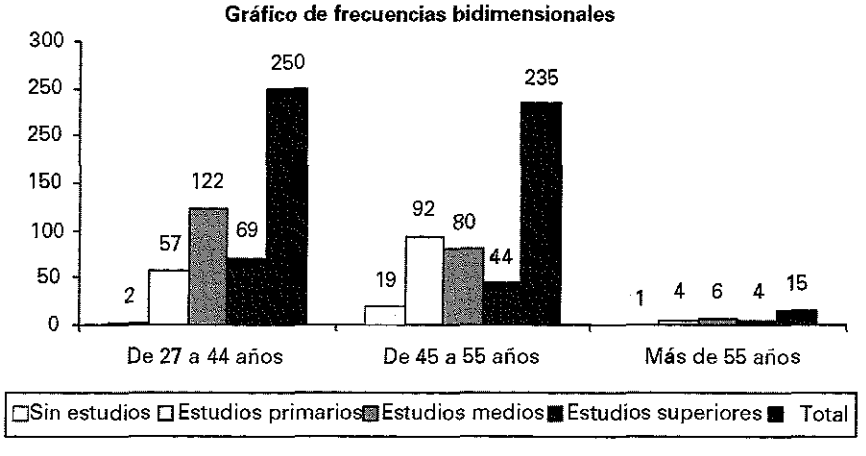
O, a la inversa:



Ejemplo 5.3. Representar gráficamente la siguiente distribución bidimensional con tramos de edad y nivel de estudios en una población de 500 individuos:

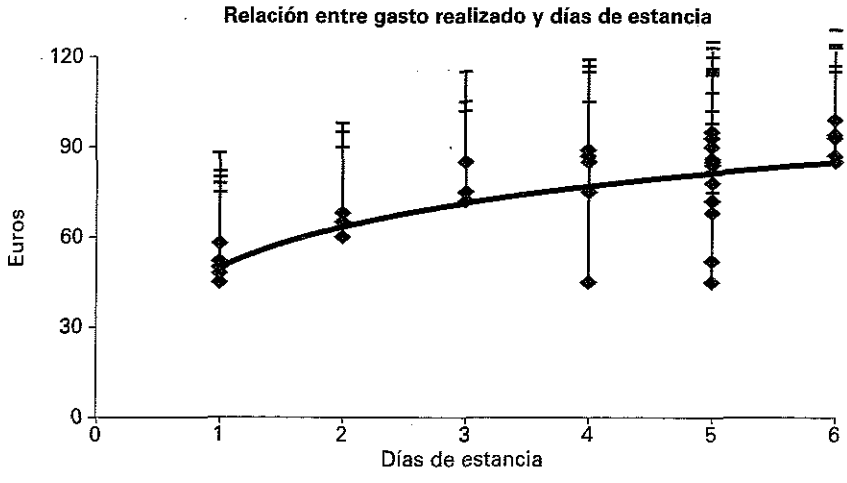
Nivel de estudios / Edad	Sin estudios	Estudios primarios	Estudios medios	Estudios superiores	Total
De 27 a 44 años	2	57	122	69	250
De 45 a 55 años	19	92	80	44	235
Más de 55 años	1	4	6	4	15
Total	22	153	208	117	500

Uno de los posibles gráficos puede ser el siguiente:



Ejemplo 5.4. Representar gráficamente la distribución del gasto realizado por 23 clientes de un hotel en relación con el número de días de estancia en el mismo.

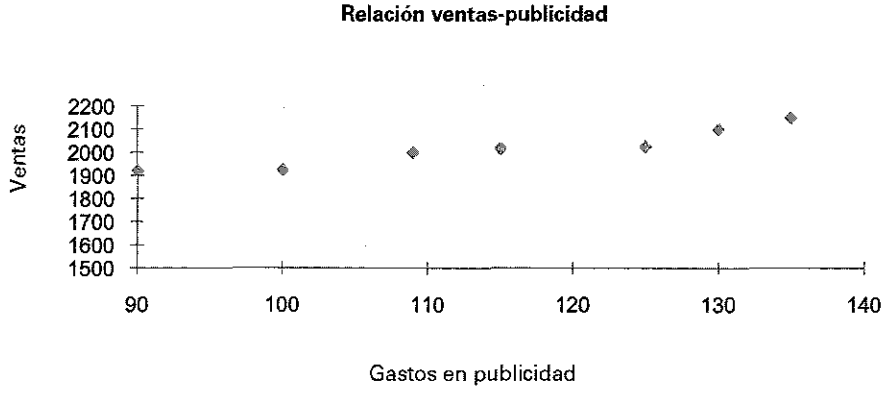
N.º de Cliente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Días de estancia	1	2	1	1	2	3	5	4	6	5	4	1	4	3	5	4	6	5	4	6	5	1	2
Gasto en €	50	60	45	48	65	85	95	89	99	86	75	58	85	75	90	87	93	85	75	94	84	52	68



Nótese que en el gráfico de puntos, se ha incorporado, lo que más adelante denominaremos, una ***línea de tendencia*** que aproxima los puntos observados.

Ejemplo 5.5. Un tipo de gráficos muy habitualmente empleado en la representación de las distribuciones bidimensionales es el denominado «gráfico de dispersión»; este tipo de gráficos se utilizan para analizar visualmente la relación que existe entre las dos variables, representándose los valores de la variable X en el eje de ordenadas y los de la variable Y en el de abscisas.

En el siguiente ejemplo se muestra la relación entre gastos en publicidad y ventas en una empresa durante siete años.



5.4. EL CÁLCULO DE LAS MEDIDAS DE POSICIÓN Y DE DISPERSIÓN EN LAS DISTRIBUCIONES MARGINALES DE FRECUENCIAS

Como se ha indicado anteriormente el concepto de distribución marginal de frecuencias en el ámbito de las distribuciones bidimensionales surge cuando se estudia aisladamente una de las variables, es decir, surgen del estudio de una variable de forma independiente a la otra.

Sí estamos en una distribución bidimensional, tendremos que considerar los pares:

— $(x_i, n_{i\cdot})$ donde i varía entre 1 y r.
— $(y_j, n_{\cdot j})$ donde j varía entre 1 y s.

Obteniendo por tanto las siguientes expresiones para la media aritmética y para la varianza:

Estadísticos	Variable x	Variable y
Media aritmética	$\bar{x} = \dfrac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N}$	$\bar{y} = \dfrac{\sum_{j=1}^{s} y_j n_j}{N}$
Varianza	$\sigma_x^2 = \dfrac{\sum_{i=1}^{r} (x_i - \bar{x})^2 n_{i\cdot}}{N}$	$\sigma_y^2 = \dfrac{\sum_{j=1}^{s} (y_j - \bar{y})^2 n_j}{N}$

Al igual que en las distribuciones unidimensionales pueden definirse los ***momentos respecto al origen y respecto a la media***; la única diferencia en este caso es que se les denota con doble subíndice:

Momentos respecto al origen:

Se obtienen mediante la siguiente expresión genérica:

$$a_{hk} = \sum_{i=1}^{r} \sum_{j=1}^{s} x_i^h \, y_j^k \, \frac{n_{ij}}{N}$$

Los más importantes son:

Estadísticos	Variable x	Variable y
Momentos respecto al origen de primer orden	$a_{10} = \dfrac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N}$	$a_{01} = \dfrac{\sum_{j=1}^{s} y_j n_{\cdot j}}{N}$
Momentos respecto al origen de segundo orden	$a_{20} = \dfrac{\sum_{i=1}^{r} x_i^2 n_{i\cdot}}{N}$	$a_{02} = \dfrac{\sum_{j=1}^{s} y_j^2 n_{\cdot j}}{N}$

Y el denominado momento producto, que es nuevo para este tipo de distribuciones y que se denota con a_{11}:

$$a_{11} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{ij}}{N}$$

Momentos respecto a la media

Se obtienen mediante la expresión general:

$$m_{hk} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})^h (y_j - \bar{y})^k \frac{n_{ij}}{N}$$

Dado que, como puede fácilmente comprobarse, $m_{10} = m_{01} = 0$, los principales momentos respecto a la media son los de segundo orden, es decir las varianzas marginales y lo que se conoce con el nombre de covarianza:

$$m_{20} = \sum_{i=1}^{r}(x_i - \bar{x})^2 \frac{n_{i\bullet}}{N}; \qquad m_{02} = \sum_{j=1}^{s}(y_j - \bar{y})^2 \frac{n_{\bullet j}}{N}$$

La covarianza se define como:

$$m_{11} = S_{xy} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} x_i y_j}{N} - \bar{x}\cdot\bar{y}$$

Al igual que en las distribuciones unidimensionales (véase el epígrafe 3.9), los momentos respecto a la media pueden ponerse en relación con los momentos respecto al origen, con las siguientes relaciones[6]:

$$S_x^2 = m_{20} = a_{20} - a_{10}^2$$

$$S_y^2 = m_{02} = a_{02} - a_{01}^2$$

$$\text{Cov}(xy) = S_{xy} = m_{11} = a_{11} - a_{10}a_{01}$$

Ejemplo 5.6. Calcule los principales estadísticos de las siguientes tablas bidimensionales de frecuencias referidas a salarios, categorías profesionales y años de antigüedad de 50 trabajadores de una determinada empresa:

TABLA 5.5.1

Años de antigüedad / Salarios mensuales en euros (€)	1	2	3	4	Más de 4	Total
1.000-1.100	2	5	2	1	5	15
1.101-1.250	2	4	1	11	2	20
1.251-1.500	3	2	4	1	3	13
1.501-2.500	0	0	0	0	2	2
Total	7	11	7	13	12	50

TABLA 5.5.2

Categorías / Salarios mensuales en euros (€)	Operarios/ trabajadores de producción	Administrativos	Técnicos y Oficiales	Comerciales	Otros	Total
1.000-1.100	2	5	2	5	1	15
1.101-1.250	3	4	1	10	2	20
1.251-1.500	3	2	1	1	6	13
1.501-2.500	1	0	1	0	0	2
Total	9	11	5	16	9	50

[6] Puede verse la demostración en Casas Sánchez y Santos Peñas, *Introducción a la Estadística para Administración y Dirección de Empresas*. Editorial CEURA. 2002.

TABLA 5.5.3

Categorías / Años de antigüedad	Operarios/ trabajadores de producción	Administrativos	Técnicos y Oficiales	Comerciales	Otros	Total
1	2	1	2	1	1	7
2	3	1	1	4	2	11
3	3	2	1	1	0	7
4	1	4	1	4	3	13
Más de 4	0	3	0	6	3	13
Total	9	11	5	16	9	50

La obtención de las medidas de posición de las distribuciones marginales de cada una de las variables se ajustaría a los mismos procedimientos descritos en el análisis de las distribuciones unidimensionales; así para la variable X (salarios), para la variable Y (categorías) y para la variable Z (años de antigüedad), tendríamos los siguientes parámetros:

Salarios mensuales en euros (↔)	Marca de clase x_i	N.° de trabajadores n_i	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
1.000-1.100	1.050	15	15.750	–172,5	29.756,25	446.343,75
1.101-1.250	1.175	20	23.500	-47,5	2.256,25	45.125
1.251-1.500	1.375	13	17.875	152,5	23.256,25	302.331,25
1.501-2.500	2.000	2	4.000	777,5	604.506,3	1.209.012,5
Total		50	61.125		659.775	2.002.812,5
Media aritmética			1.222,5			
Varianza						40.056,25
Desviación típica						200,14
Coeficiente de Variación de Pearson						0,16

El alumno puede obtener, como práctica los valores de la mediana, moda, cuartiles, coeficientes de asimetría y curtosis, etc. de la variable unidimensional X.

En el caso de la variable Y, al ser cualitativa, además de las frecuencias relativas, sólo indicaremos la moda.

Categoría profesional	Frecuencia n_j	Frecuencia relativa f_j (× 100)
Operarios/trabajadores de producción	9	18
Administrativos	11	22
Técnicos y Oficiales	5	10
Comerciales	16	32
Otros	9	18
Total	50	100

Siendo en este caso la moda la categoría correspondiente a los comerciales, acumulando el 32% de la frecuencia total.

En el caso de la variable Z (años de antigüedad) deberemos proceder a la estimación de una marca de clase para el último intervalo (intervalo abierto) antes de proceder al cálculo de los parámetros unidimensionales; si suponemos una marca de clase de 6 años para este intervalo, obtendremos los siguientes valores:

Media aritmética	3,48
Varianza	2,97
Desviación típica	1,72
Coeficiente de Variación de Pearson	0,49

Para el cálculo del parámetro bidimensional de la covarianza nos quedaremos con el par de variables X (salarios), Z (años de antigüedad), ambas de naturaleza cuantitativa, procediendo de la siguiente forma:

1. Obtención de los valores medios (media aritmética) univariantes para X y Z; estos valores calculados son $\overline{x} = 1.222{,}5$ y $\overline{z} = 3{,}48$.
2. Utilizando los valores medios calculamos las columnas $x_i - \overline{x}$ y $z_k - \overline{z}$, operando, a partir de las marcas de clase de los intervalos, de la siguiente forma:

	$x_i - \overline{x}$
Para $x = 1.050$	–172,5
Para $x = 1.175$	–47,5
Para $x = 1.375$	152,5
Para $x = 2.000$	777,5

	$z_k - \overline{z}$
Para $z = 1$	–2,48
Para $z = 2$	-1,48
Para $z = 3$	-0,48
Para $z = 4$	0,52
Para $z = 6$	2,52

3. A partir de los valores de las n_{ij} en la tabla 5.5.1 construimos la siguiente tabla:

	$x_i - \bar{x}$	$z_k - \bar{z}$	n_{ik}	$(x_i - \bar{x})(z_k - \bar{z})\, n_{ik}$
Para (x = 1.050; z = 1)	–172,5	–2,48	2	855,6
Para (x = 1.050; z = 2)	–172,5	–1,48	5	1.276,5
Para (x = 1.050; z = 3)	–172,5	–0,48	2	165,6
Para (x = 1.050; z = 4)	–172,5	0,52	1	–89,7
Para (x = 1.050; z = 6)	–172,5	2,52	5	–2.173,5
Para (x = 1.175; z = 1)	–47,5	–2,48	2	235,6
Para (x = 1.175; z = 2)	–47,5	–1,48	4	281,2
Para (x = 1.175; z = 3)	–47,5	–0,48	1	22,8
Para (x = 1.175; z = 4)	–47,5	0,52	11	–271,7
Para (x = 1.175; z = 6)	–47,5	2,52	2	–239,4
Para (x = 1.375; z = 1)	152,5	–2,48	3	–1.134,6
Para (x = 1.375; z = 2)	152,5	–1,48	2	–451,4
Para (x = 1.375; z = 3)	152,5	–0,48	4	–292,8
Para (x = 1.375; z = 4)	152,5	0,52	1	79,3
Para (x = 1.375; z = 6)	152,5	2,52	3	1.152,9
Para (x = 2.000; z = 1)	777,5	–2,48	0	0
Para (x = 2.000; z = 2)	777,5	–1,48	0	0
Para (x = 2.000; z = 3)	777,5	–0,48	0	0
Para (x = 2.000; z = 4)	777,5	0,52	0	0
Para (x = 2.000; z = 6)	777,5	2,52	2	3.918,6
Total / Suma				3.335,0

En la que tenemos:

— En la segunda columna los valores $x_i - \bar{x}$,
— En la tercera columna los valores $z_k - \bar{z}$,
— En la cuarta los valores de las frecuencias n_{ik} ($n_{11} = 2$, ya que hay dos trabajadores con un salario de entre 1.000 y 1.100 ⇐mensuales y 1 año de antigüedad; $n_{33} = 4$, ya que hay 4 trabajadores con un salario comprendido entre 1.251 y 1.500 ⇐mensuales y 3 años de antigüedad, etc.).
— En la quinta columna el producto de las 3 anteriores, y
— En la última fila el sumatorio de la última columna.

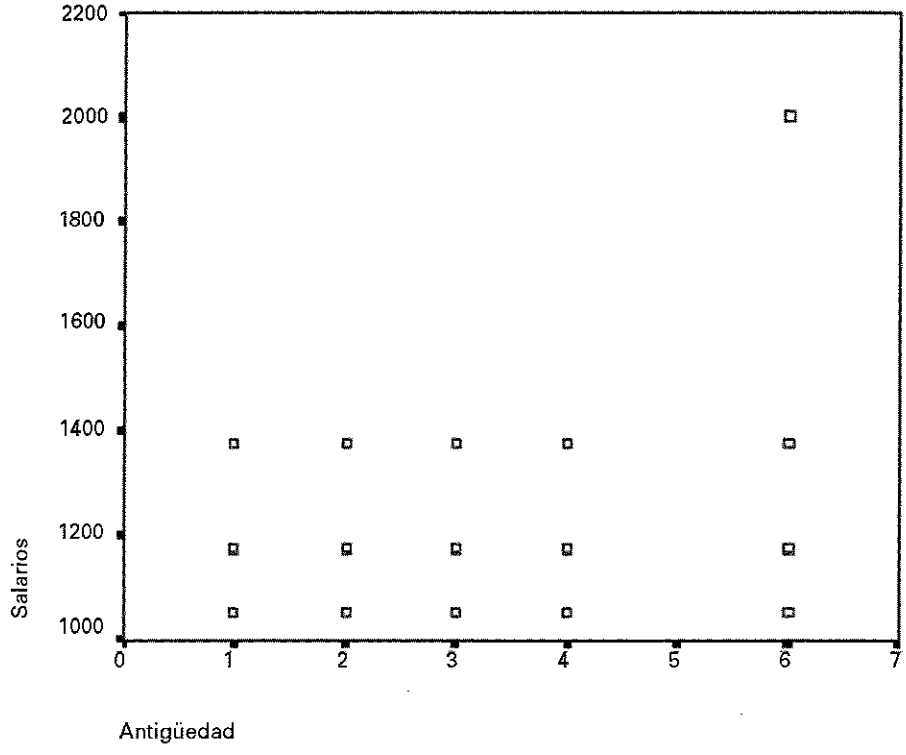
Aplicando la expresión de la covarianza, tenemos:

$$Cov(x,z) = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} n_{ik}\left(x_i - \bar{x}\right)\left(z_k - \bar{z}\right)}{N} = \frac{3.335}{50} = 66,7$$

Este valor representa una relación positiva entre los salarios (X) y los años de antigüedad (Z), indicándonos que, estadísticamente, a medida que aumenta la antigüedad de los trabajadores aumenta su salario.

Representación gráfica

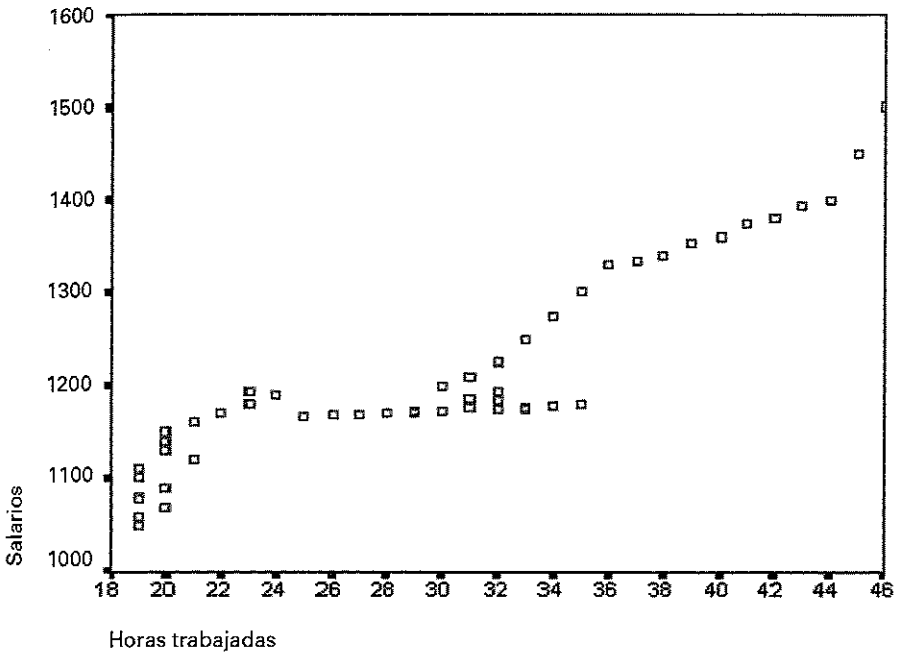
Mediante un diagrama de dispersión tendríamos el siguiente gráfico de salarios en relación con la antigüedad:



5.5. LA DEPENDENCIA ESTADÍSTICA ENTRE DOS O MÁS VARIABLES

El gráfico anterior parece indicar la inexistencia de una clara relación entre las dos variables consideradas, ya que algunas personas con una antigüedad de 6 años en la empresa ganan menos que otras con un año de antigüedad y, sin embargo, otras ganan mucho más.

Hemos generado otro gráfico correspondiente a una distribución entre dos nuevas variables para 50 trabajadores de la misma categoría que trabajan en una empresa; las variables consideradas han sido el salario mensual y las horas de trabajo, con el siguiente resultado:



En este diagrama de dispersión la relación parece más evidente, de forma que la nube de puntos tiene una cierta tendencia hacia arriba, indicativa de que al aumentar las horas trabajadas aumenta el salario recibido.

Puede resultarnos interesante estudiar esta dependencia con objeto de verificar si existe relación entre dos o más pares de variables (salario y horas trabajadas, salarios y categorías profesionales, salarios y antigüedad en una empresa, número de visitantes y plazas hoteleras existentes en una ciudad, número de clientes y precio de las habitaciones de un hotel, etc.)

El estudio de estas relaciones de dependencia estadística es realizado por la ***Teoría de la Correlación***, *que determina si existe una cierta covariación o variación conjunta entre dos variables y refleja numéricamente dicha dependencia.*

Esta relación o posible relación de dependencia puede adoptar tres resultados:

- ***Independencia funcional o correlación nula***: cuando no existe ninguna relación entre las variables, como parecía deducirse del ejemplo 5.5.
- ***Dependencia funcional o correlación funcional***: cuando existe una función tal que todos los valores de la variable la satisfacen (a cada valor de X le corresponde uno solo de Y o a la inversa).

  Sería el caso de la correlación que se da, por ejemplo, entre los ingresos de una empresa y el número de productos vendidos, cuando no se hace ningún tipo de descuento por compras.

Unidades de producto vendido	Ingresos de la empresa
1	100 €
2	200 €
3	300 €
4	400 €

— ***Dependencia aleatoria o dependencia estadística parcial***: se da cuando los puntos del diagrama se ajustan en alguna medida a una función (recta o curvilínea); esta situación es la más habitual en Estadística, ya que, cuando se comparan dos variables, lo normal es que exista algún grado de relación entre ellas, incluso en el caso de que no exista razón aparente alguna que explique dicha relación.

Este es el caso de la relación entre salarios y horas trabajadas del ejemplo anterior, dónde como puede comprobarse, cuando se examina un conjunto más o menos amplio de trabajadores, no siempre a un mismo número de horas trabajadas le corresponde el mismo salario, ni viceversa; no obstante, si existe una relación indicativa de que a más horas trabajadas se percibe mayor salario mensual. Probablemente la relación no es funcional, es decir, matemática, por cuánto intervienen otras variables como la categoría de los empleados, la antigüedad, alguna prima especial por conocimiento de idiomas, horas extras nocturnas o en días festivos percibidas a precio diferente del resto, etc.

Las relaciones de dependencia estadística pueden ser positivas (directas) o negativas (inversas).

En los casos en que exista dependencia estadística es interesante disponer de algún instrumento que nos permita medir el grado de dependencia o correlación existente y decidir conocer la función (recta, parábola, aproximación logarítmica, etc.) que explica mejor la dependencia.

Como veremos más tarde, el primer aspecto, es decir el signo y la intensidad de la dependencia, es estudiado por la ***Teoría de la Correlación,*** mientras que el segundo, el estudio de las funciones que relacionan las variables, es abordado por la ***Teoría de la Regresión***.

5.6. CASUALIDAD, CAUSALIDAD Y ESPECIFICACIÓN DE MODELOS

Antes de estudiar la correlación y la regresión de variables es interesante hacer una reflexión sobre las características de estas relaciones.

Digamos en primer lugar que el principal objetivo de estas relaciones es la explicación y en su caso la predicción de una variable, que llamaremos ***variable dependiente o endógena*** con otra u otras variables, las ***variables independientes o exógenas***.

Se trataría, pues de poder explicar el comportamiento de una variable endógena, como por ejemplo la cifra de negocio de una determinada empresa, en función de ciertas variables exógenas, como los precios finales de los productos fabricados, el gasto en publicidad o en I+D+i, la puesta en marcha de una nueva estrategia de venta, etc.

Es importante hacer una consideración al término de ***causalidad*** de las relaciones; supongamos que una empresa durante un período de estudio ha aplicado diferentes precios de venta de uno de sus productos y que dispone de una tabla en la que aparecen relacionados los distintos precios y las ventas obtenidas con cada precio aplicado; es bastante probable que comparando ambas series, pueda establecer una relación, más o menos compleja, que le determine, con algún grado de aproximación, la cifra de ventas que obtendrá sí aplicase un cierto precio a su producto.

En este caso existe una *causalidad* derivada de la *teoría general de la demanda*, que nos indica que, normalmente, y con las demás condiciones constantes, a medida que se aumenta el precio se genera una menor demanda y viceversa

Es posible, no obstante, que si comparamos la serie de las ventas con alguna otra serie de datos que nada tuviese que ver con la empresa estudiada, número de nacimientos, automóviles matriculados en China o cantidad vendida de ordenadores, encontremos alguna relación estadística más o menos intensa; en estos casos tendríamos que hablar de relaciones de ***casualidad***, ya que no es nada probable que las causas de la evolución de la demanda de nuestra empresa tengan algo que ver con la variación de la natalidad de determinado país o con la venta de ordenadores, al menos no es posible construir un modelo lógico, amparado en alguna teoría sustentable, que establezca dicha relación.

Hay que tener en cuenta, sin embargo, que algunas variables, puedan estar indirectamente relacionadas a través de terceras variables que sí podrían haberse incluido en el modelo teórico; así, por ejemplo, un aumento de las matrículas de automóviles o de las ventas de ordenadores puede derivarse de un incremento en el nivel de vida que a su vez sería también una buena variable explicativa del aumento de las ventas de nuestra empresa.

Al estudiar una relación entre variables, es importante, pues, la ***especificación previa de un modelo teórico*** que recoja las principales relaciones de causalidad; así, por ejemplo, como nos dice la teoría de la demanda, la variable endógena ventas de la empresa en cuestión, podría venir explicada por los precios de venta de su producto, por la evolución de la renta disponible en su espacio comercial, por los precios aplicados por sus competidores, etc.

Como es lógico, en la explicación de nuestra variable de interés es seguro que intervienen otras múltiples variables, ratios de calidad-precio con los productos competidores, capacidad logística, gastos en publicidad, incentivos de la red comercial, etc.; sin embargo, no siempre es posible tener datos de todas las variables influyentes, ni tampoco es siempre aconsejable incluirlas en un modelo que se haría excesivamente complejo, por lo que se tiende a la simplificación.

Estos modelos son conocidos como ***modelos econométricos*** y todos ellos incluyen una variable adicional, que trata de recoger el efecto conjunto de múltiples variables irrelevantes o escasamente relevantes y que se denomina *«perturbación aleatoria» (v)*; por incluir esta variable los modelos se denominan *estocásticos* y adoptan una forma multiecuacional del tipo:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... a_n x_n + v$$

A través de los parámetros a_i, estos modelos permiten cuantificar la influencia de cada una de las variables exógenas (x_i) en la variable endógena y efectuar predicciones

o previsiones sobre la evolución de esta variable a medida que varían las variables exógenas.

El modelo anterior corresponde a la más simple de las relaciones, la relación lineal, que en el caso de dos variables, adopta la ecuación de una recta ($Y = a + b\,X$); más adelante estudiaremos otras relaciones más complejas.

En todo caso, y como colofón a este epígrafe, el alumno debe tener siempre claro que en Estadística en general y en Economía en particular, *los estudios de regresión, correlación y dependencia estadística deben, ajustarse siempre a un modelo causal más o menos complejo.*

5.7. CORRELACIÓN O GRADO DE DEPENDENCIA LINEAL ENTRE DOS VARIABLES

Cuando tenemos dos variables cuantitativas[7] una medida de la asociación o correlación entre ellas viene dada por el ***Coeficiente de Correlación Lineal de Pearson***, que se define como

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

Expresión en la que:

S_{xy} se corresponde con la covarianza entre X e Y.
S_x se corresponde a la desviación típica de X.
S_y se corresponde a la desviación típica de Y.

El valor de este parámetro está siempre comprendido entre –1 y +1 ($-1 \leq r_{xy} \leq 1$), lo que nos permite y facilita la interpretación de las relaciones en la siguiente forma:

— Cuando $|r_{xy}| = 1$ se tiene una relación lineal perfecta entre las variables X e Y, por lo que podemos calcular exactamente el valor de Y asociado con cada uno de los valores de X o viceversa. Si $r_{xy} = 1$ la relación es positiva o directa, si $r_{xy} = -1$, la relación es negativa o inversa. El signo positivo o negativo se lo da el valor de la covarianza.
— Cuando $r_{xy} = 0$ indica que no existe ninguna relación de tipo lineal entre las variables. Ello no es óbice para que exista otra dependencia no lineal (cuadrática, por ejemplo).
— Cuando $-1 < r_{xy} < 1$ existe dependencia estadística; en general suele aceptarse la siguiente clasificación:

[7] Existen también diversos coeficientes o indicadores que miden la correlación entre variables cualitativas y entre variables cualitativas y variables cuantitativas, pero no los estudiaremos en este texto; pueden citarse al efecto los coeficientes de Spearman, de Goodman y Kruskall, de Kendall, de Yule, etc...; los interesados en esta materia pueden ver el libro Santos, Muñoz, Juez *Diseño de Encuestas para Estudios de Mercado. Técnicas de Muestreo y Análisis Multivariante.* Ed. Ramón Areces. 2003.

- Valor de r de 0 a 0,25 implica que no existe *correlación suficiente* entre ambas variables.
- Valor de r de 0,25 a 0,50 implica una *correlación baja* a moderada.
- Valor de r de 0,50 a 0,75 implica *correlación moderada* a buena.
- Valor de r de 0,75 o mayor, implica una muy buena a *excelente correlación*.

Estos rangos de valores se pueden extrapolar también, lógicamente, a las correlaciones negativas.

Señalemos, por último, que, aunque viene expresado en términos numéricos, este coeficiente tiene carácter cualitativo, es decir que si en un caso se obtiene un r = 0,3 y en otro un $r = 0,6$, sólo podemos afirmar que en el segundo caso la intensidad de la relación es mayor que en el primero, pero no que es el doble que en el primero.

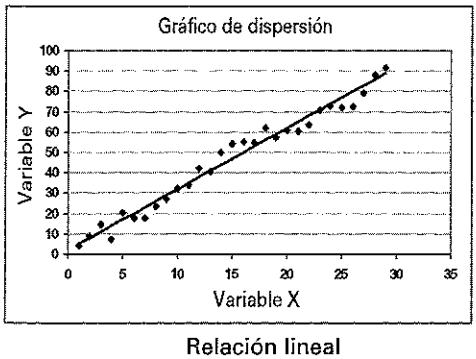
5.8. REGRESIÓN LINEAL SIMPLE

La simple constatación de la existencia de una asociación entre dos variables no permite realizar predicciones sobre los valores que adoptará una variable al asignar valores a la otra. Para ello, y una vez decidido si existe o no esa dependencia entre las variables es importante saber si podemos encontrar una función (con forma de recta, parábola, etc.) que nos dé una buena aproximación de la nube de puntos y que nos sirva, por tanto, para hacer predicciones; esta función matemática se denomina ***ecuación de regresión***.

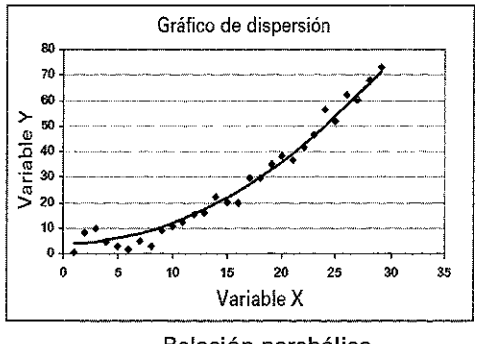
La regresión consiste en ajustar lo más posible la nube de puntos de un diagrama de dispersión a una función. Cuando la función es una recta obtenemos la recta de regresión lineal, cuando es una parábola, una regresión parabólica, cuando es una exponencial, una regresión exponencial, etc.

La regresión de dos variables debe afrontar, pues, dos tipos de problemas: decidir qué función se ajusta mejor a los datos disponibles y realizar dicho ajuste.

Para afrontar el primer problema una forma muy útil es acudir a la representación gráfica de los datos; así, se comprueba, por ejemplo, que en la primera de las siguientes distribuciones, la figura que mejor se ajusta a los datos disponibles es una recta, mientras que la segunda queda bastante mejor explicada con una parábola.



Relación lineal



Relación parabólica

Si se dispone de los programas informáticos más utilizados para realizar regresiones estadísticas puede optarse por realizar distintas simulaciones comprobando la bondad de los distintos ajustes, mediante el análisis de determinados parámetros que veremos más adelante.

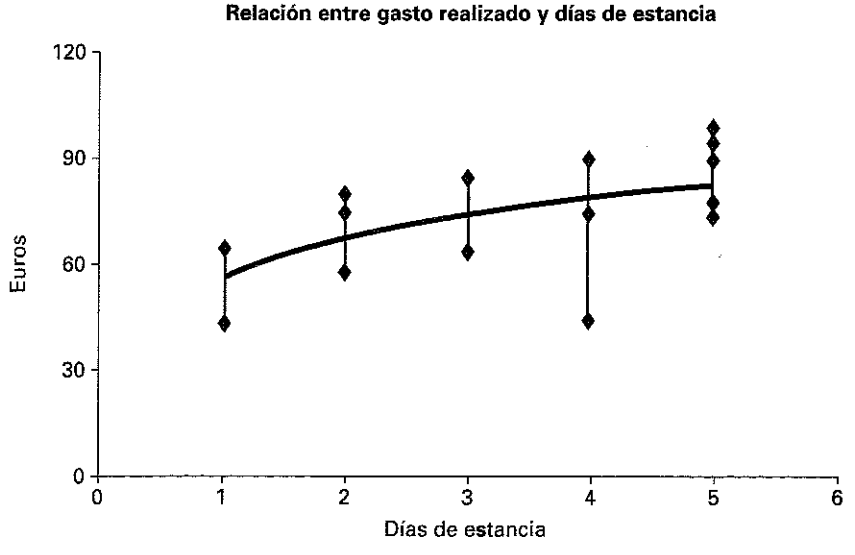
Veamos como afrontar el segundo de los problemas trabajando con la relación funcional más simple, desde el punto de vista del análisis matemático, es decir, con la **relación *lineal*.** Como es sabido, este tipo de relaciones vienen dadas por ecuaciones del tipo:

$$y = a + b\,x,$$

Que se corresponden con la ecuación general de una recta en dos dimensiones.

Se trata en definitiva de ajustar una recta de forma que se aproxime el máximo posible a la nube de puntos de la distribución, determinando los parámetros a y b de la ecuación; para ello pueden utilizarse diversos procedimientos o *métodos de ajuste*; el más utilizado es el denominado ***Método o Criterio de Ajuste por Mínimos Cuadrados.***

Este procedimiento está basado en admitir que la representación más adecuada de la dependencia entre dos variables es aquella función que hace mínima la suma de las diferencias al cuadrado entre los valores reales y los valores teóricos obtenidos a partir de la función ajustada (distancias verticales de las observaciones o puntos de información previa de que disponemos a la recta). Supongamos como ejemplo las variables gasto realizado por los clientes de un hotel y días de estancia en el mismo; se muestran en el siguiente gráfico las distancias que se hacen mínimas al estimar la regresión.



Luego dadas dos variables cuantitativas X, Y, se trata de encontrar una ecuación del tipo:

$$Y = a + b\,X$$

Que nos permita aproximar los valores de Y a partir de los de X.

En principio existen infinitas soluciones capaces de satisfacer la condición expresada en el párrafo anterior, ya que sobre el plano podemos trazar infinitas rectas más o menos próximas entre sí; para elegir una es necesario añadir alguna restricción adicional que permita obtener una solución perfectamente determinada. Con esta finalidad, exigimos también que el error cometido al realizar la predicción de Y sea mínimo.

Si llamamos y_i al valor de la variable Y que tiene asociado el valor x_i de la variable X, e y_i^* al valor que resultará de calcular la ecuación $Y = a + b\,X$, con el valor de x_i, es decir: $y_i^* = a + bx_i$, entonces el error cometido en la predicción será:

$$e_i = y_i - y_i^* \text{ (la distancia vertical en el gráfico)}$$

De forma que cada punto puede escribirse como:

$$y_i = y_i^* + e_i$$

O, lo que es lo mismo:

$$y_i = a + b\,x_i + e_i$$

$$e_i = y_i - a - b \cdot x_i$$

Estos errores son unos positivos y otros negativos, dependiendo de que la y_i observada esté por debajo o por encima de la recta y_i^* estimada. Se elevan al cuadrado mediante la expresión $\bullet\ e_i^2$, con objeto de que la suma algebraica de los errores no sea nula.

Se trata de hacer mínimo $S = \sum e_i^2$, es decir:

$$S = \sum_{i=1}^{n} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2$$

Para que esta expresión sea un mínimo, la condición necesaria es que se anulen las derivadas parciales respecto a las incógnitas, que son los coeficientes de regresión lineal a y b.

Derivando respecto de los parámetros a y b e igualando los términos a cero, se obtiene el siguiente sistema de ecuaciones:

$$\frac{\partial}{\partial a}\sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = 0 \Rightarrow -2\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0 \Rightarrow \sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = 0 \Rightarrow -2\sum_{i=1}^{N}x_i\left(y_i - a - bx_i\right) = 0 \Rightarrow \sum_{i=1}^{N}x_i\left(y_i - a - bx_i\right) = 0$$

Operando queda el siguiente ***sistema de ecuaciones normales***:

$$\begin{aligned} \sum_{i=1}^{N} y_i &= Na + b\sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} y_i x_i &= a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 \end{aligned} \qquad [5.1]$$

En el caso de que queramos hacer la regresión de X sobre Y, las ecuaciones serían:

$$\begin{aligned} \sum_{i=1}^{N} x_i &= Na' + b'\sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i &= a'\sum_{i=1}^{N} y_i + b'\sum_{i=1}^{N} y_i^2 \end{aligned} \qquad [5.2]$$

Operando adecuadamente podríamos también llegar a otras expresiones que nos permiten el cálculo de los parámetros a y b de la ecuación de regresión; así, por ejemplo, puede demostrarse que:

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{S_{xy}}{S_x^2}$$

$$a = \bar{y} - b\bar{x}$$

Luego la recta de regresión puede obtenerse mediante la expresión:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \qquad [5.3]$$

Del mismo modo podríamos habernos planteado la ecuación de regresión de x sobre y. En este caso, si suponemos que $x = a' + b'\,y$, tendríamos:

$$b' = \frac{S_{xy}}{S_y^2}$$

$$a' = \bar{x} - b'\bar{y} \qquad [5.4]$$

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

Expresando las medias, covarianzas y varianzas en momentos respecto al origen y respecto a la media y teniendo en cuenta que:

$\bar{x} = a_{10}$	$\bar{y} = a_{01}$	
$S_{xy} = m_{11}$	$S_x^2 = m_{20}$	$s_y^2 = m_{02}$

Las ecuaciones normales de regresión se pueden expresar como:

— **Regresión de *y* sobre *x***

$$b = \frac{m_{11}}{m_{20}}$$

$$a = a_{01} - ba_{10}$$

$$y - a_{01} = \frac{m_{11}}{m_{20}}(x - a_{10}) \qquad [5.5]$$

— **Regresión de *x* sobre *y***

$$b' = \frac{m_{11}}{m_{02}}$$

$$a' = a_{10} - ba_{01}$$

$$x - a_{10} = \frac{m_{11}}{m_{02}}(y - a_{01}) \qquad [5.6]$$

El alumno puede elegir en cada caso las fórmulas de cálculo que considere más cómodas (ecuaciones normales o método de los momentos).

Ejemplo 5.7. Una empresa ha comprobado experimentalmente que sus ventas (en miles de euros por semana) están relacionadas con el número de trabajadores disponibles para atender a la clientela; dispone a tal fin de los siguientes datos:

N.º de trabajadores (x_i)	Ventas (y_i)
5	20
6	25
7	29
8	33
9	37
10	41

Ajústese la función lineal que mejor exprese la relación entre ambas variables.

SOLUCIÓN

Utilizamos el sistema de ecuaciones normales, para cuyo planteamiento necesitamos los valores de las columnas 3.ª y 4.ª de la siguiente tabla:

	y_i	x_i	$x_i \cdot y_i$	x_i^2
	20	5	100	25
	25	6	150	36
	29	7	203	49
	33	8	264	64
	37	9	333	81
	41	10	410	100
Total	185	45	1.460	355

De la expresión [5.1] se desprende que el sistema de ecuaciones normales

$$\sum y_i = Na + b\sum x_i$$

$$\sum y_i x_i = a\sum x_i + b\sum x_i^2$$

Tomará los valores:

$$185 = 6a + 45b$$
$$1.460 = 45a + 355b$$

Cuya solución es[8]:

$$a = -0{,}238$$

$$b = 4{,}143$$

Con lo que la recta ajustada será:

$$y_j = -0{,}238 + 4{,}143\, x_i$$

Veamos ahora en qué medida esta ecuación se ajusta a los valores existentes, lo que nos dará un indicio de su capacidad para predecir otros valores futuros.

	Valor estimado	Valor Real	Diferencia Real-Estimado	% diferencia
Para $x = 5$	$y = 20{,}48$	20	–0,48	–2,38%
Para $x = 6$	$y = 24{,}62$	25	0,38	1,52%
Para $x = 7$	$y = 28{,}76$	29	0,24	0,82%
Para $x = 8$	$y = 32{,}90$	33	0,10	0,29%
Para $x = 9$	$y = 37{,}05$	37	–0,05	–0,13%
Para $x = 10$	$y = 41{,}19$	41	–0,19	–0,46%

Las diferencias porcentuales entre el valor real y el estimado son relativamente pequeñas (promedio inferior a –0,06%), por lo que parece que la función obtenida se ajusta bastante a la nube de puntos de la distribución.

Si ello es así, parece posible que la recta estimada permita situar otros posibles pares de valores entre las dos variables investigadas, es decir que conociendo el valor de una de las variables (la que denominamos independiente) podamos «estimar» el valor de la otra (la variable dependiente) o lo que es lo mismo «predecir» un hipotético resultado; en nuestro caso, por ejemplo, predecir cuales serían las ventas sí aumentamos la plantilla hasta 11 ó 15 empleados.

Veremos este aspecto en el siguiente apartado.

[8] Aplicando, por ejemplo, la regla de Kramer:

$$a = \frac{\begin{vmatrix} 185 & 45 \\ 1460 & 355 \end{vmatrix}}{\begin{vmatrix} 6 & 45 \\ 45 & 355 \end{vmatrix}} = -0{,}23 \qquad b = \frac{\begin{vmatrix} 6 & 185 \\ 45 & 1460 \end{vmatrix}}{\begin{vmatrix} 6 & 45 \\ 45 & 355 \end{vmatrix}} = 4{,}14$$

5.9. BONDAD DEL AJUSTE Y PREDICCIONES

En el apartado anterior hemos construido una nueva variable: la diferencia entre el valor observado o valor disponible con anterioridad y el valor estimado; esta diferencia se denota ***e*** y representa el error cometido en cada predicción, de forma que $e = y - y^*$.

Siempre que se efectúa una regresión es necesario estimar algunas medidas de dispersión que nos valoren el grado en el que la función estimada puede sustituir a las observaciones de las que se obtuvo; estas medidas de dispersión también nos puede proporcionar información sobre el grado de dependencia entre las variables regresadas y se conocen como *medidas de bondad del ajuste*.

Una primera medida de dispersión, conocida como *varianza residual o varianza de los errores o residuos*, viene dada por la media aritmética de los errores al cuadrado, es decir:

$$S_e^2 = \frac{\sum_{i=1}^{N}\left(y_i - y_i^*\right)^2}{N} = \frac{\sum_{i=1}^{N} e_i^2}{N}$$ (dado que la media de e_i por definición es igual a 0).

Cuando esta medida es alta quiere decir que los residuos son grandes y, en consecuencia, que en general, la función estimada se aleja bastante de los valores originales (nube de puntos) y es poco representativa de los mismos. Por el contrario, si la expresión da valores pequeños será indicativo de que existe bastante representatividad.

A partir de este concepto se construye el denominado ***Coeficiente de Determinación*** que es la medida de bondad del ajuste habitualmente más utilizada.

Para ello hacemos el siguiente razonamiento:

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

- Es decir, la varianza total de nuestra variable dependiente *Y* tiene dos componentes: uno debido a la relación entre las variables y que está contenida en el término S_y^2, que llamamos varianza explicada por la regresión; y
- Otro que es la *varianza residual* S_e^2 y que contiene la variabilidad que no es capaz de explicar el modelo lineal.

A partir de esta expresión definimos el *coeficiente de determinación* R^2 como el grado de participación de la varianza explicada en la varianza total de la variable observada ***y***; este coeficiente vendrá definido por la expresión:

$$R^2 = r_{xy}^2 = \frac{S_{y^*}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2} \qquad [5.7]$$

En el caso de la regresión lineal, puede comprobarse que la varianza de la variable y^* puede calcularse como:

$$S_{y^*}^2 = \frac{\sum_{i=1}^{N}\left(y_i^* - \bar{y}\right)^2}{N} = \frac{\sum_{i=1}^{N}\left(a + bx_i - \bar{y}\right)^2}{N} = \frac{S_{xy}^2}{S_x^2}$$

Obteniendo por tanto, la siguiente expresión alternativa, que como vemos, *coincide con el Coeficiente de Correlación Lineal de Pearson elevado al cuadrado*:

$$R^2 = r_{xy}^2 = \frac{S_{Y^*}^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} \qquad [5.8]$$

Pudiendo expresarse también en relación con los momentos respecto a la media mediante la siguiente expresión:

$$R^2 = \frac{m_{11}^2}{m_{02} \cdot m_{20}} \qquad [5.9]$$

Para su cálculo el alumno puede utilizar la expresión que le resulte más cómoda.

Este coeficiente por su definición es genérico y sirve para cualquier tipo de regresión, ya sea lineal, cuadrática o de cualquier otra naturaleza[9].

El coeficiente de determinación R^2 tiene un valor comprendido entre 0 y 1 (0 y 100%); cuando su valor es cero indica una nula representatividad de la ecuación de regresión y cuando su valor es 1 indica un ajuste perfecto entre la ecuación estimada y la nube de puntos; los valores intermedios indican mayor o menor representatividad o ajuste de la función, recomendándose valores altos (superiores a 0,85) para dar por representativa la ecuación estimada.

De lo anterior también se desprende que, a la hora de estimar los parámetros del modelo, resultará de vital importancia que el término de error no ejerza ninguna influencia determinante en la explicación del comportamiento de la variable dependiente. Aunque estos temas se estudian en un curso superior y particularmente en la asignatura de Econometría, baste aquí indicar que, cuando se aplica el método de mínimos cuadrados ordinarios, se deben realizar hipótesis de comportamiento sobre el término de error, las cuales afectan a la forma de su distribución. Sin entrar en detalle, avanzaremos que los errores deben presentar un comportamiento aleatorio y no presentar correlación con la/s variables/s explicativa/s del modelo.

[9] Existe algún caso, como la denominada regresión logística, caracterizada por ser la variable dependiente cualitativa, en que este coeficiente no es calculable, si bien existen aproximaciones, entre las cuales citar el R cuadrado de Cox y Shell y el R cuadrado de Nagelkerke.

Una vez estimada una ecuación de regresión podemos emplearla para obtener datos que se encuentren en el mismo rango que los estudiados (***interpolación***) o para obtener valores ajenos a los inicialmente disponibles (***extrapolación***). Así en el ejemplo anterior, con la ecuación estimada podríamos estimar las ventas de otra empresa con 5, 6, 7, 8, 9 ó 10 trabajadores o los de la propia empresa u otra similar que tuviese 20 trabajadores; lógicamente en el primer caso (interpolación) se obtendrán valores mas fiables que en el segundo, ya que en nuestros datos de partida no hemos incorporado datos experimentales de una empresa de este nuevo tamaño y desconocemos si a partir de cierto tamaño se mantendrán los mismos términos de progresión de las ventas al aumentar el número de trabajadores.

Ejemplo 5.8. Una determinada empresa dispone en sus registros de los siguientes datos que relacionan el gasto semanal en publicidad con la cifra de ventas de un determinado período (en ambos casos los datos se presentan en miles de euros):

Período	Gasto en publicidad (x_i)	Ventas (y_j)
Semana 1	20	200
Semana 2	25	240
Semana 3	24	400
Semana 4	23	350
Semana 5	20	200
Semana 6	20	250
Semana 7	10	80

Obtener una recta de regresión que permita predecir las ventas futuras en función del gasto realizado en publicidad y valorar la calidad de dichas predicciones.

SOLUCIÓN

Vamos a operar por el método de los momentos; procedemos a elaborar la siguiente tabla:

	x_i	y_j	$x_i \cdot y_j$	x_i^2	y_j^2
	20	200	4.000	40.000	400
	25	240	6.000	57.600	625
	24	400	9.600	160.000	576
	23	350	8.050	122.500	529
	20	200	4.000	40.000	400
	20	250	5.000	62.500	400
	10	80	800	6.400	100
Total	142	1.720	37.450	489.000	3.030

Recordemos las expresiones que permiten el cálculo de los momentos respecto a la media y de los momentos respeto al origen:

	Variable *x*	**Variable *y***
Momentos respecto al origen de primer orden	$a_{10} = \dfrac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N}$	$a_{01} = \dfrac{\sum_{i=1}^{s} y_j n_{\cdot j}}{N}$
Momentos respecto al origen de segundo orden	$a_{20} = \dfrac{\sum_{i=1}^{r} x_i^2 n_{i\cdot}}{N}$	$a_{02} = \dfrac{\sum_{i=1}^{s} y_j^2 n_{\cdot j}}{N}$
Momento producto respeto al origen	$a_{11} = \dfrac{\sum_{i=1}^{r}\sum_{i=1}^{s} x_i y_j n_{ij}}{N}$	
Momento producto respecto a la media (covarianza)	$m_{11} = \sigma_{xy} = \dfrac{\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N} = \dfrac{\sum_{i=1}^{r}\sum_{j=1}^{s} x_i y_j n_{ij}}{N} - \bar{x}\cdot\bar{y}$	
Momentos respecto a la media en relación con momentos respecto al origen	$s_x^2 = m_{20} = a_{20} - a_{10}^2$ $s_y^2 = m_{02} = a_{02} - a_{01}^2$ $\mathrm{Cov}(xy) = s_{xy} = m_{11} = a_{11} - a_{10}a_{01}$	

Calculamos:

	Momentos respecto al origen de primer orden	**Momentos respecto al origen de segundo orden**
Variable *x*	$a_{10} = \dfrac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N} = \dfrac{142}{7} = 20{,}29 = \bar{x}$	$a_{20} = \dfrac{\sum_{i=1}^{r} x_i^2 n_{i\cdot}}{N} = \dfrac{3.030}{7} = 432{,}86$
Variable *y*	$a_{01} = \dfrac{\sum_{i=1}^{s} y_j n_{\cdot j}}{N} = \dfrac{1.720}{7} = 245{,}71 = \bar{y}$	$a_{02} = \dfrac{\sum_{i=1}^{s} y_j^2 n_{\cdot j}}{N} = \dfrac{489.000}{7} = 69.857{,}14$
Momento producto respeto al origen	$a_{11} = \dfrac{\sum_{i=1}^{r}\sum_{i=1}^{s} x_i y_j n_{ij}}{N} = \dfrac{37.450}{7} = 5.350$	

	Momentos respecto a la media de segundo orden
Variable x	$S_x^2 = m_{20} = a_{20} - a_{10}^2 = 432{,}85 - 20{,}29^2 = 21{,}35$
Variable y	$S_y^2 = m_{02} = a_{02} - a_{01}^2 = 69.857{,}14 - 245{,}71^2 = 9.481{,}63$
Momento producto respecto a la media	$S_{xy} = m_{11} = a_{11} - a_{10}a_{01} = 5.350 - 20{,}29 \cdot 245{,}71 = 365{,}51$

Aplicando las relaciones

$$b = \frac{m_{11}}{m_{20}} = \frac{365{,}51}{21{,}35} = 17{,}12$$

$$a = a_{01} - ba_{10} = -101{,}63$$

Nos queda

$$y - \bar{y} = \frac{m_{11}}{m_{20}}(x - \bar{x}) \quad \Leftrightarrow \quad y - a_{01} = \frac{m_{11}}{m_{20}}(x - a_{10})$$

$$\Rightarrow \quad y - 245{,}71 = 17{,}12\,(x - 20{,}29) \quad \Rightarrow \quad y = 17{,}12x - 101{,}63$$

El coeficiente de determinación valdría:

$$R^2 = \frac{m_{11}^2}{m_{02} \cdot m_{20}} = \frac{365{,}51^2}{9.481{,}63 \cdot 21{,}34} = 0{,}66$$

Indicativo de que la recta de regresión no tiene un alto poder predictivo; de hecho, con esta ecuación, hubiésemos obtenido los siguientes valores de y_j.

Período	Gasto en publicidad (x_i)	Ventas reales (y_j)	Ventas estimadas (y_j^*)
Semana 1	20	200	$y_1 = 17{,}12 \cdot 20 - 101{,}63 = 240{,}8$
Semana 2	25	240	$y_2 = 17{,}12 \cdot 25 - 101{,}63 = 326{,}4$
Semana 3	24	400	$y_3 = 17{,}12 \cdot 24 - 101{,}63 = 309{,}3$
Semana 4	23	350	$y_4 = 17{,}12 \cdot 23 - 101{,}63 = 292{,}2$
Semana 5	20	200	$y_5 = 17{,}12 \cdot 20 - 101{,}63 = 240{,}8$
Semana 6	20	250	$y_6 = 17{,}12 \cdot 20 - 101{,}63 = 240{,}8$
Semana 7	10	80	$y_7 = 17{,}12 \cdot 10 - 101{,}63 = 69{,}6$

Estos valores son algo acertados en algunos casos (la sexta semana nos da un valor relativamente aproximado) y bastante alejados en otros, ya que por, ejemplo, para la última semana, es decir, para un gasto en publicidad de 10 hubiésemos previsto unas ventas de 69.600 euros, cuando en realidad las ventas fueron de 80.000 euros; parece, en consecuencia, que la regresión lineal no es suficientemente buena para predecir la serie; es decir, que la línea recta no se ajusta adecuadamente a la nube de puntos que presentan los datos de partida.

Convendría por ello buscar otra regresión no lineal como las que estudiamos en el siguiente apartado.

Ejemplo 5.9. En una distribución de frecuencias para dos variables X e Y se ha obtenido la siguiente tabla de correlaciones:

X \ Y	2	3	4	*Total*
4	5	6	2	*13*
5	4	4	2	*10*
6	3	6	8	*17*
Total	*12*	*16*	*12*	*40*

Obtener la regresión lineal simple que une ambas variables (x sobre y; y sobre x).
Construimos la tabla auxiliar

x_i \ y_j	2	3	4	$n_{i\cdot}$	$x_i \cdot n_{i\cdot}$	$x_i^2 \cdot n_{i\cdot}$	$\sum_j y_j n_{ij}$	$x_i \sum_j y_j n_{ij}$
4	5	6	2	13	52	208	36	144
5	4	4	2	10	50	250	28	140
6	3	6	8	17	102	612	56	336
$n_{\cdot j}$	12	16	12	40	204	1.070	120	620
$y_j \cdot n_{\cdot j}$	24	48	48	120				
$y_j^2 \cdot n_{\cdot j}$	48	144	192	384				

A partir de esta información obtenemos los momentos respecto al origen a_{hk} y respecto a la media m_{hk} y sabiendo que los momentos respecto a la media pueden expresarse en función de los momentos respecto al origen, tendremos:

$$a_{10} = \frac{204}{40} = 5{,}1 \qquad a_{01} = \frac{120}{40} = 3$$

$$a_{20} = \frac{1.070}{40} = 26,75 \qquad a_{02} = \frac{384}{40} = 9,6$$

$$a_{11} = \frac{620}{40} = 15,5 \qquad m_{11} = a_{11} - a_{10}a_{01} = 15,5 - 5,1 \cdot 3 = 0,2$$

$$m_{20} = a_{20} - a_{10}^2 = 26,75 - 5,1^2 = 0,74 \qquad m_{02} = a_{02} - a_{01}^2 = 9,6 - 3^2 = 0,6$$

Dado que:

$$b = m_{11}/m_{20} \qquad a = a_{01} - ba_{10} \qquad y - a_{01} = m_{11}/m_{20}(x - a_{10})$$

Se tiene que

$$b = 0,27 \qquad a = 1,62 \qquad y - 3 = 0,27 \cdot (x - 5,1)$$

Operando se obtiene la siguiente recta de regresión de y sobre x: $y = 1,62 + 0,27x$.

El valor del coeficiente de determinación indica, sin embargo, que el ajuste realizado no es adecuado.

$$R^2 = \frac{m_{11}^2}{m_{02} \cdot m_{20}} = \frac{0,2^2}{0,74 \cdot 0,6} = 0,09$$

Por lo que en principio no existe relación de dependencia entre ambas variables o debe estudiarse otro tipo de ajuste no lineal.

5.10. REGRESIÓN NO LINEAL

Los modelos de regresión no siempre son lineales, es decir, no siempre vienen expresados por la ecuación de una recta. Existen también modificaciones de esta ecuación de tal manera que se pueden practicar análisis de regresión cuadrática, cúbica, logarítmica, logística, etc. Además en la regresión pueden intervenir dos o más variables independientes, constituyendo lo que se denomina *análisis de regresión multivariante*.

Los principales modelos no lineales que se utilizan en estadística son:

La función polinómica:

$$y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$$

Aunque no lo tratamos en este nivel, puede resolverse aplicando el criterio de mínimos cuadrados de forma similar a la ecuación lineal.

La función potencial:

$$y = ax^b$$

En estas funciones puede tomarse logaritmos neperianos de forma que:

$$ln(y) = ln(a) + b \cdot ln(x)$$

Si en esta transformación se hace un cambio de variables tal que $y' = ln(y)$ y $x' = ln(x)$ se estará ante una regresión lineal del tipo: $y' = a' + b'x'$. Una vez determinadas a' y b', se calcula a y b tomando antilogaritmos.

La función exponencial:

$$y = ab^x$$

Operando como en el caso anterior se transforma en: $ln(y) = ln(a) + x \cdot ln(b)$. Haciendo, en este caso, $y' = log(y)$ calculamos a' y b' de modo que $y' = a' + b'x$. Una vez calculados obtenemos los parámetros originales a y b.

La función logarítmica:

$$y = a + b \cdot \log(x)$$

Basta con hacer el cambio $x' = ln(x)$ y tratarlo como una ecuación lineal.

Ejemplo 5.10. Dados los siguientes datos correspondientes a dos variables

y_i	x_i
8	2
10	3
12	4
16	6
21	10

Ajústese una parábola que exprese la relación entre ambas variables.

SOLUCIÓN

En este caso la curva seleccionada es $y_j^* = a + bx_i + cx_i^2$

	y_j	x_i	x_i^2	x_i^3	x_i^4	$x_i \cdot y_j$	$x_i^2 \cdot y_j$
	8	2	4	8	16	16	32
	10	3	9	27	81	30	90
	12	4	16	64	256	48	192
	16	6	36	216	1.296	96	576
	21	10	100	1.000	10.000	210	2.100
Suma	67	25	165	1.315	11.649	400	2.990

El sistema de ecuaciones normales, obtenido derivando los residuos en función de los parámetros a, b y c será:

$$\sum_{i=1}^{N} y_i = Na + b\sum_{i=1}^{N} x_i + c\sum_{i=1}^{N} x_i^2$$

$$\sum_{i=1}^{N} y_i x_i = a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} x_i^2 + c\sum_{i=1}^{N} x_i^3 \quad \rightarrow \quad \begin{aligned} 67 &= 5a + 25b + 165c \\ 400 &= 25a + 165b + 1.315c \\ 2.990 &= 165a + 1.315b + 11.649c \end{aligned}$$

$$\sum_{i=1}^{N} y_i x_i^2 = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i^3 + c\sum_{i=1}^{N} x_i^4$$

Cuya solución es:

$a = 2{,}8566$
$b = 2{,}6737$
c $= -0{,}0856$

Con lo que la función ajustada será

$$y^* = 2{,}8566 + 2{,}6737\, x_i - 0{,}0856\, x_i^2$$

Se muestra en la siguiente tabla los valores estimados y la comparación con los valores observados:

	Valor estimado	Valor Real	% diferencia
Para $x = 2$	$y = 7{,}8615$	8	1,73%
Para x = 3	$y = 10{,}1072$	10	–1,07%
Para $x = 4$	$y = 12{,}1816$	12	–1,51%
Para $x = 6$	$y = 15{,}8169$	16	1,14%
Para $x = 10$	$y = 21{,}0328$	21	–0,16%

Ejemplo 5.11. La relación entre renta familiar disponible y consumo aparente de un determinado grupo de familias es la siguiente (en miles de euros anuales).

Consumo (C)	*Renta (R)*
2	3
3	6
4	9
5	12
6	15
7	20

Ajustar una función de consumo potencial

SOLUCIÓN

La función a ajustar es del tipo $C = A \cdot R^b$
Tomando logaritmos se tiene $\ln(C) = \ln(A) + b \cdot \ln(R)$
Llamando $c = \ln(C)$ $r = \ln(R)$ $a = \ln(A)$
La función a ajustar es $c = a + b \cdot r$; es decir un ajuste lineal.
Formamos la siguiente tabla:

Consumo (C)	Renta (R)	ln C	ln R	ln C · ln R	$(\ln R)^2$
2	3	0,6931	1,0986	0,7615	1,2069
3	6	1,0986	1,7918	1,9684	3,2104
4	9	1,3863	2,1972	3,0460	4,8278
5	12	1,6094	2,4849	3,9993	6,1748
6	15	1,7918	2,7081	4,8522	7,3335
7	20	1,9459	2,9957	5,8294	8,9744
		8,5252	13,2763	20,4569	31,7279

El sistema de ecuaciones normales será:

$$8{,}5252 = 6\,a + 13{,}2763\,b$$
$$20{,}4569 = 13{,}2763\,a + 31{,}7279\,b$$

Cuya solución es:

$a = -0{,}0784$

$b = 0{,}6776$

A = antilogaritmo de $a = 0{,}9246$

Con lo que la función ajustada será

$$C = 0{,}9246\, R^{0{,}6776}$$

El valor de consumo esperado para los distintos niveles de renta será:

	Valor estimado	Valor real	% diferencia
Para Renta = 3	$C = 1{,}9464$	2	2,68%
Para Renta = 6	$C = 3{,}1132$	3	–3,77%
Para Renta = 9	$C = 4{,}0975$	4	–2,44%
Para Renta = 12	$C = 4{,}9793$	5	0,41%
Para Renta = 15	$C = 5{,}7920$	6	3,47%
Para Renta = 20	$C = 7{,}0386$	7	–0,55%

Ejemplo 5.12. Dada la siguiente distribución bidimensional

x_i	y_i
5	70
6	75
7	80
11	85
12	90

Ajustar una función exponencial del tipo $y = a\, b^x$

Linealizamos mediante logaritmos $\ln(y) = \ln(a) + x \ln(b)$

Y formamos la siguiente tabla para obtener la recta $Y = A + Bx$

Llamando a: $Y = \ln(y)$; $B = \ln(b)$; $A = \ln(a)$;

	x_i	y_i	$\ln y_i$	x_i^2	$x_i \cdot \ln y_i$
	5	70	4,2485	25	21,2425
	6	75	4,3175	36	25,9049
	7	80	4,3820	49	30,6742
	11	85	4,4427	121	48,8692
	12	90	4,4998	144	53,9977
Total	41		21,8905	375	180,6885

El sistema de ecuaciones normales será

$$21{,}8905 = 5\,A + 41\,B$$
$$180{,}6885 = 41\,A + 375\,B$$

Cuya solución es:

$A = 4{,}1273$

$B = 0{,}0306$

a = antilogaritmo de A ($a = e^A$) = 62,0113

b = antilogaritmo de B ($b = e^B$) = 1,0311

Con lo que la función ajustada será

$$y = 4{,}1273 + 0{,}0306\,x$$

Y la función exponencial será

$$y^* = 62{,}0113 \cdot 1{,}0311^x$$

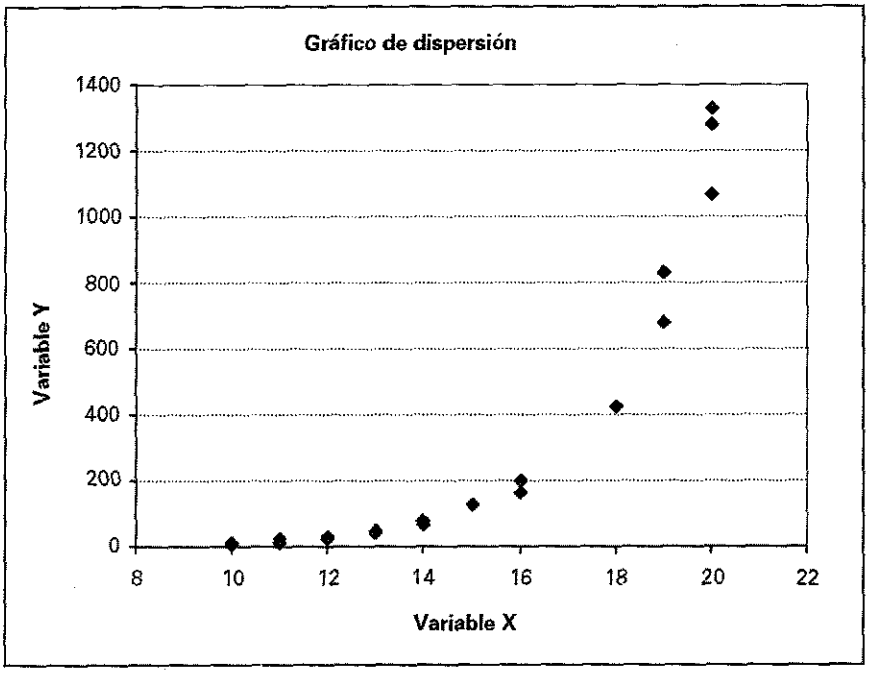
Se muestra a continuación los valores estimados y su comparación con los observados.

	Valor estimado	Valor real	% diferencia
Para $x = 5$	$y = 72{,}08$	70	–3,12%
Para $x = 6$	$y = 74{,}50$	75	0,67%
Para $x = 7$	$y = 76{,}81$	80	4,15%
Para x = 11	$y = 86{,}81$	85	–2,09%
Para $x = 12$	$y = 89{,}51$	90	0,55%

Ejemplo 5.13. Con los datos que aparecen en la siguiente tabla, estimar la ecuación de regresión que mejor describa la variable Y en función de la X, y compare los resultados obtenidos con respecto a un ajuste lineal.

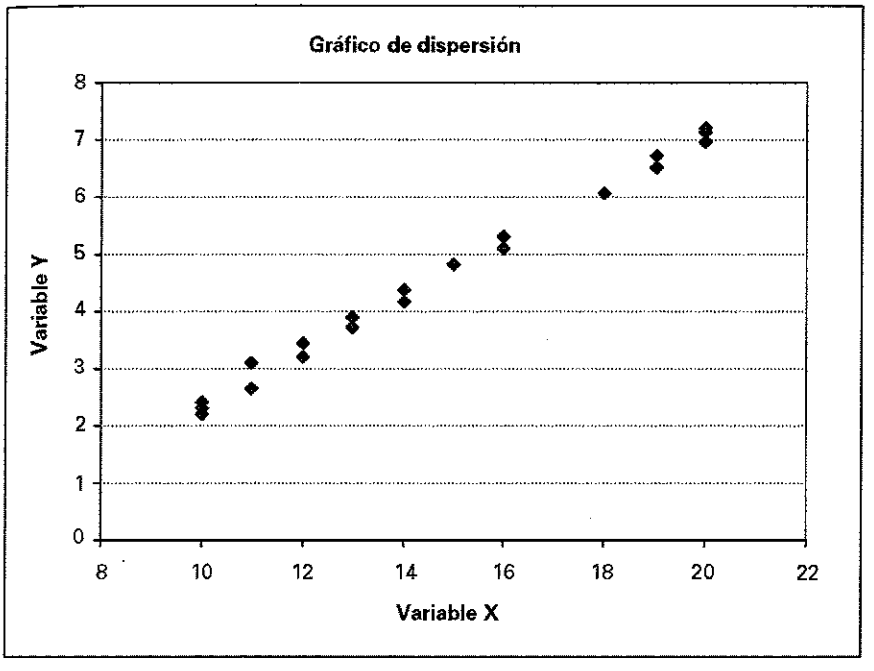
i	X	Y	i	X	Y
1	10	9	16	14	64
2	10	10	17	14	79
3	10	10	18	14	64
4	10	11	19	15	127
5	11	22	20	16	166
6	11	14	21	16	202
7	12	25	22	18	425
8	12	31	23	19	831
9	12	31	24	19	678
10	13	41	25	19	830
11	13	50	26	19	678
12	13	41	27	20	1.330
13	14	79	28	20	1.066
14	14	79	29	20	1.279
15	14	64	Suma	422	8.336

Se muestra a continuación el gráfico de dispersión de ambas variables. Tal y como se aprecia, la relación no es lineal y, por lo tanto, no será éste el mejor ajuste.



Si la relación funcional fuera la exponencial, X y ln(Y) se ajustarían adecuadamente a través de una línea recta. Para comprobar esto, se realiza la representación gráfica de ambas variables donde puede apreciarse una clara relación lineal.

Análogamente, en el caso de una relación potencial, se observará una relación lineal entre ln (X) y ln (Y), y si la función más adecuada fuera la logarítmica dicha relación lineal aparecerá entre Y y ln (X).



Se muestra a continuación la tabla con los cálculos necesarios para la estimación de la regresión lineal y la exponencial.

i	x_i	y_i	$ln(y_i)$	x_i^2	$x_i y_i$	$x_i \cdot ln(y_i)$
1	10	9	2,1972	100	90	21,9722
2	10	10	2,3026	100	100	23,0259
3	10	10	2,3026	100	100	23,0259
4	10	11	2,3979	100	110	23,9790
5	11	22	3,0910	121	242	34,0015
6	11	14	2,6391	121	154	29,0296
7	12	25	3,2189	144	300	38,6265
8	12	31	3,4340	144	372	41,2078
9	12	31	3,4340	144	372	41,2078
10	13	41	3,7136	169	533	48,2764
11	13	50	3,9120	169	650	50,8563
12	13	41	3,7136	169	533	48,2764
13	14	79	4,3694	196	1.106	61,1723
14	14	79	4,3694	196	1.106	61,1723
15	14	64	4,1589	196	896	58,2244
16	14	64	4,1589	196	896	58,2244
17	14	79	4,3694	196	1.106	61,1723
18	14	64	4,1589	196	896	58,2244
19	15	127	4,8442	225	1.905	72,6628
20	16	166	5,1120	256	2.656	81,7918
21	16	202	5,3083	256	3.232	84,9323
22	18	425	6,0521	324	7.650	108,9376
23	19	831	6,7226	361	15.789	127,7300
24	19	678	6,5191	361	12.882	123,8638
25	19	830	6,7214	361	15.770	127,7071
26	19	678	6,5191	361	12.882	123,8638
27	20	1330	7,1929	400	26.600	143,8587
28	20	1066	6,9717	400	21.320	139,4334
29	20	1279	7,1538	400	25.580	143,0767
Suma	422	8.336	131,0587	6.462	155.828	2.059,5332
Media	14,55	287,45 =	4,52 =	222,83	5.373,38 =	71,02 =

La estimación de los parámetros asociados a la regresión lineal sería:

$$b_1 = \frac{m_{11}^{(1)}}{m_{20}} = \frac{a_{11} - a_{10}a_{01}^{(1)}}{a_{20} - a_{10}^2} = \frac{5.373{,}38 - 14{,}55 \cdot 287{,}45}{222{,}83 - 14{,}55^2} = 107{,}50$$

$$a_1 = a_{01}^{(1)} - b_1 a_{10} = 287{,}45 - 107{,}50 \cdot 14{,}55 = -1.276{,}81$$

Y por tanto la recta de regresión será: $\boxed{y_{1i} = -1.276{,}81 + 107{,}50 \cdot x_i}$

Análogamente, la estimación de los parámetros asociados a la regresión exponencial será:

$$b_2 = \frac{m_{11}^{(2)}}{m_{20}} = \frac{a_{11} - a_{10}a_{01}^{(2)}}{a_{20} - a_{10}^2} = \frac{71{,}02 - 14{,}55 \cdot 4{,}52}{222{,}83 - 14{,}55^2} = 0{,}4745$$

$$a_2 = a_{01}^{(2)} - b_2 a_{10} = 4{,}52 - 0{,}4745 \cdot 14{,}55 = -2{,}3858$$

Y por tanto la recta de regresión será: $\ln(y_{2i}) = -2{,}3858 + 0{,}4745 \cdot x_i$
Deshaciendo la transformación, la ecuación finalmente estimada será:

$$y_{2i} = \exp(-2{,}3858)\exp(0{,}4745)^{x_i} \Rightarrow y_{2i} = 0{,}0920 \cdot 1{,}6072^{x_i}$$

Se muestra en la siguiente tabla los valores predichos para las dos ecuaciones lineales calculadas, los errores obtenidos y el valor predicho y error derivado de la ecuación exponencial.

i	x_i	y_i	$\ln(y_i)$	y_{1i}	$\ln(y_{2i})$	$e_{1i} = y_i - y_{1i}$	$e_{2i} = \ln(y_i) - \ln(y_{2i})$	y_{2i}	$y_i - y_{2i}$
1	10	9	2,1972	-201,84	2,3594	210,84	-0,1622	10,58	-1,58
2	10	10	2,3026	-201,84	2,3594	211,84	-0,0568	10,58	-0,58
3	10	10	2,3026	-201,84	2,3594	211,84	-0,0568	10,58	-0,58
4	10	11	2,3979	-201,84	2,3594	212,84	0,0385	10,58	0,42
5	11	22	3,0910	-94,35	2,8339	116,35	0,2571	17,01	4,99
6	11	14	2,6391	-94,35	2,8339	108,35	-0,1948	17,01	-3,01
7	12	25	3,2189	13,15	3,3084	11,85	-0,0895	27,34	-2,34
8	12	31	3,4340	13,15	3,3084	17,85	0,1256	27,34	3,66
9	12	31	3,4340	13,15	3,3084	17,85	0,1256	27,34	3,66
10	13	41	3,7136	120,64	3,7829	-79,64	-0,0694	43,95	-2,95
11	13	50	3,9120	120,64	3,7829	-70,64	0,1291	43,95	6,05
12	13	41	3,7136	120,64	3,7829	-79,64	-0,0694	43,95	-2,95
13	14	79	4,3694	228,14	4,2575	-149,14	0,1120	70,63	8,37
14	14	79	4,3694	228,14	4,2575	-149,14	0,1120	70,63	8,37
15	14	64	4,1589	228,14	4,2575	-164,14	-0,0986	70,63	-6,63
16	14	64	4,1589	228,14	4,2575	-164,14	-0,0986	70,63	-6,63
17	14	79	4,3694	228,14	4,2575	-149,14	0,1120	70,63	8,37
18	14	64	4,1589	228,14	4,2575	-164,14	-0,0986	70,63	-6,63
19	15	127	4,8442	335,64	4,7320	-208,64	0,1122	113,52	13,48
20	16	166	5,1120	443,13	5,2065	-277,13	-0,0945	182,45	-16,45
21	16	202	5,3083	443,13	5,2065	-241,13	0,1018	182,45	19,55
22	18	425	6,0521	658,12	6,1555	-233,12	-0,1035	471,32	-46,32
23	19	831	6,7226	765,62	6,6301	65,38	0,0926	757,53	73,47
24	19	678	6,5191	765,62	6,6301	-87,62	-0,1109	757,53	-79,53
25	19	830	6,7214	765,62	6,6301	64,38	0,0914	757,53	72,47
26	19	678	6,5191	765,62	6,6301	-87,62	-0,1109	757,53	-79,53
27	20	1330	7,1929	873,12	7,1046	456,88	0,0884	1217,54	112,46
28	20	1066	6,9717	873,12	7,1046	192,88	-0,1329	1217,54	-151,54
29	20	1279	7,1538	873,12	7,1046	405,88	0,0492	1217,54	61,46

Como puede observarse los errores cometidos con la ecuación exponencial son mucho menores que los asociados a la regresión lineal (7ª y 10ª columna). Para comprobar la bondad del ajuste obtenido en ambas regresiones calcularemos el R^2.

El R^2 de la regresión lineal obtenido será:

$$R_1^2 = 1 - \frac{S_{e_1}^2}{S_y^2} = 1 - \frac{\sum_{i=1}^{29} e_{1i}^2}{\sum_{i=1}^{29} (y_i - \bar{y})^2} = 1 - \frac{1.038.083,94}{4.749.373,17} = 0,781$$

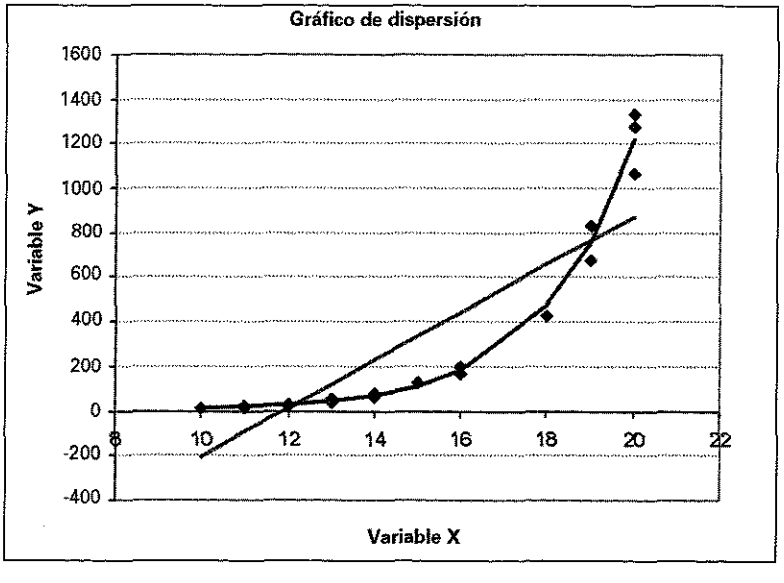
siendo, por tanto, el coeficiente de correlación lineal igual a:

$$r = \sqrt{0,781} = 0,884$$

El R^2 asociado a la ecuación exponencial será:

$$R_1^2 = 1 - \frac{S_{e_1}^2}{S_y^2} = 1 - \frac{\sum_{i=1}^{29} e_{1i}^2}{\sum_{i=1}^{29} (y_i - \bar{y})^2} = 1 - \frac{0,3829}{72,7012} = 0,995 \quad \text{, siendo } z_i = \ln(y_i).$$

Si no hubiésemos comprobado previamente la distribución del diagrama de dispersión, hubiésemos podido dar por buena la relación lineal, con un $R^2 = 0,781$, pero como vemos, la función exponencial ajusta mucho mejor los datos. Se muestra en el siguiente gráfico el diagrama de dispersión de X e Y, y los valores predichos por ambas regresiones.



5.11. INTRODUCCIÓN A LA REGRESIÓN MÚLTIPLE

Pasamos a continuación a generalizar el modelo anterior al caso de un modelo con varias variables exógenas, de tal forma que se trata de determinar la relación que existe entre la variable endógena Y y las variables exógenas, X_1, X_2, X_k. Dicho modelo se puede formular matricialmente de la siguiente manera:

$$Y = X \cdot \beta + e \Rightarrow Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + e_i \qquad i = 1, 2, \ldots, n$$

Donde:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}$$ Es el vector de observaciones de la variable endógena.

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix} = [X_1\ X_2\ \ldots\ X_k]$$ Es la matriz de observaciones de las variables exógenas.

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K \end{pmatrix}$$ Es el vector de coeficientes que pretendemos estimar.

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}$$ Es el vector de términos de error.

Si en la expresión anterior se considerara que existe término independiente, , la matriz X quedaría como:

$$X = \begin{pmatrix} 1 & X_{12} & \ldots & X_{1k} \\ 1 & X_{22} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & X_{n2} & \ldots & X_{nk} \end{pmatrix} = [1\ \ X_2\ \ X_3\ \ \ldots\ \ X_k]$$

Y el modelo quedaría así:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ik} + u_i \qquad i = 1, 2, ..., n$$

El problema a resolver nuevamente es la minimización de la suma de los cuadrados de los términos de error tal que:

$$Min \quad \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - Y_i^*)^2 = \sum_{i=1}^{N} (Y_i - \beta X_i)^2$$

Desarrollando dicho cuadrado y derivando respecto a cada β_i obtenemos el siguiente sistema de ecuaciones expresado en notación matricial:

$$X' X \beta = X' Y$$

En donde basta con despejar β premultiplicando ambos miembros por la inversa de la matriz $(X' X)$ para obtener la estimación de los parámetros del modelo tal que:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Donde:

$$X'X = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1} X_{i2} & \cdots & \sum_{i=1}^{n} X_{i1} X_{ik} \\ \sum_{i=1}^{n} X_{i2} X_{i1} & \sum_{i=1}^{n} X_{i2}^2 & \cdots & \sum_{i=1}^{n} X_{i2} X_{ik} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \sum_{i=1}^{n} X_{ik} X_{i1} & \sum_{i=1}^{n} X_{ik} X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{pmatrix} \qquad X'Y = \begin{pmatrix} \sum_{i=1}^{n} X_{i1} Y_i \\ \sum_{i=1}^{n} X_{i2} Y_i \\ \cdots\cdots \\ \sum_{i=1}^{n} X_{ik} Y_i \end{pmatrix}$$

Si en el modelo existiera término independiente, α, las matrices anteriores serían:

$$X'X = \begin{pmatrix} n & \sum_{i=1}^{n} X_{i1} & \cdots & \sum_{i=1}^{n} X_{ik} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \cdots & \sum_{i=1}^{n} X_{i1} X_{ik} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{ik} X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{pmatrix} \qquad X'Y = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1} Y_i \\ \cdots\cdots \\ \sum_{i=1}^{n} X_{ik} Y_i \end{pmatrix}$$

El resultado de multiplicar dichas matrices conduce a la obtención de la estimación de los parámetros β_i del modelo:

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ \sum_{i=1}^{n} X_{i2}X_{i1} & \sum_{i=1}^{n} X_{i2}^2 & \cdots & \sum_{i=1}^{n} X_{i2}X_{ik} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \\ \cdots\cdots \\ \sum_{i=1}^{n} X_{ik}Y_i \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdots \\ \hat{\beta}_k \end{pmatrix}$$

Cada uno de los coeficientes estimados, $\hat{\beta}_i$, representan la variación que experimenta la variable dependiente Y cuando una variable independiente X_i varía en una unidad y todas las demás permanecen constantes (supuesto *ceteris paribus*).

Además de los supuestos comentados en la regresión lineal simple, en este caso se presupone que las variables independientes no están correlacionadas entre sí. Si una de las variables X_i pudiera expresarse en función de las demás (correlación perfecta) la matriz X'X no sería invertible y no podrían calcularse los parámetros (existirían infinitas soluciones). Los efectos del incumplimiento de ésta hipótesis de partida pueden ser muy graves. No obstante, existen procedimientos para minimizar estos efectos, entre ellos, la correcta selección de las variables explicativas, eliminando aquellas que resulten redundantes. Otros procedimientos serían la regresión ridge o contraída y la regresión en componentes principales, cuyo estudio se aborda en textos más avanzados; en la asignatura de Econometría se profundizará sobre estos aspectos.

Ejemplo 5.14. *Una empresa quiere medir el impacto que tienen en las ventas el gasto en publicidad y la política de precios aplicada. Los datos de ventas, gasto en publicidad y evolución del precio medio de sus productos se recogen en la siguiente tabla.*

Año	Ventas (miles de €) Y	Gasto en publicidad (miles de €) X_1	Evolución precio X_2
1995	5.710	24	100
1996	5.700	21	102
1997	5.760	26	103
1998	5.720	20	102
1999	5.790	29	100
2000	5.790	30	100
2001	5.700	20	102
2002	5.750	24	102
2003	5.770	20	100
2004	5.720	18	96
2005	5.760	22	97
2006	5.780	24	98
2007	5.830	28	95
2008	5.870	30	94
2009	5.840	28	98

Considerando el término independiente, la matriz X será igual a:

1	24	100
1	21	102
1	26	103
1	20	102
1	29	100
1	30	100
1	20	102
1	24	102
1	20	100
1	18	96
1	22	97
1	24	98
1	28	95
1	30	94
1	28	98

Según hemos visto, el sistema de ecuaciones normales se obtiene a partir de la expresión:

$$X'X \cdot \beta = X'Y$$

Que en nuestro caso sería:

$$\begin{pmatrix} 15 & 364 & 1.489 \\ 364 & 9.062 & 36.086 \\ 1.489 & 36.086 & 147.919 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 86.490 \\ 2.101.170 \\ 5.584.250 \end{pmatrix}$$

Despejando el vector de parámetros en la ecuación, la solución sería

$$\hat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} 6.383.937 \\ 8.539 \\ -8.312 \end{pmatrix}$$

Luego la ecuación finalmente obtenida sería: $Y = 6.383{,}937 + 8{,}539\ X1 - 8{,}312\ X2$, la cual nos indica que un incremento de 1.000 euros en gastos de publicidad, manteniendo el precio constante, se traduce en un aumento de 8.539 euros en las ventas. Análogamente, si aumentamos el precio en una unidad, las ventas disminuirán en 8.312 euros.

Se muestra en la siguiente tabla los valores de ventas estimados, los errores cometidos en la estimación, y su cuadrado.

Año	Ventas (miles de €) Y	Ventas estimadas (miles de €) Y^*	Errores $e = Y - Y^*$	Errores² $e^2 = (Y - Y^*)^2$
1995	5.710	5.757,63	–47,63	2268,36
1996	5.700	5.715,39	–15,39	236,76
1997	5.760	5.749,77	10,23	104,71
1998	5.720	5.706,85	13,15	172,97
1999	5.790	5.800,32	–10,32	106,51
2000	5.790	5.808,86	–18,86	355,65
2001	5.700	5.706,85	–6,85	46,90
2002	5.750	5.741,00	9,00	80,95
2003	5.770	5.723,47	46,53	2.164,76
2004	5.720	5.739,65	–19,65	385,94
2005	5.760	5.765,49	–5,49	30,11
2006	5.780	5.774,25	5,75	33,04
2007	5.830	5.833,34	–3,34	11,18
2008	5.870	5.858,73	11,27	126,95
2009	5.840	5.808,41	31,59	998,16

Para valorar el grado de ajuste de la ecuación obtenida, calculamos el coeficiente de determinación

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{\sum_{i=1}^{15} e_i^2}{\sum_{i=1}^{15} (y_i - \bar{y})^2} = 1 - \frac{7.122,93}{38.160} = 0,8133$$

Es decir, el 81,33% de la variabilidad de las ventas es explicada por la ecuación de regresión obtenida.

5.12. ESTUDIO DE LA ASOCIACIÓN ENTRE VARIABLES CUALITATIVAS

Para saber si existe o no relación entre variables de tipo cualitativo se utilizan las tablas de contingencia. Este tipo de variables pueden ser nominales (por ejemplo el sexo de los encuestados), de atributos (marcas de un producto) u ordinales (por ejemplo la medición del grado de satisfacción de los encuestados en una determinada escala). El empleo de las tablas de contingencia está especialmente indicado si las variables son de tipo nominal.

Una tabla de contingencia se utiliza para mostrar la existencia de relaciones entre dos variables en una encuesta estadística. También, mediante una tabla de contingencia podemos establecer una medición del grado de relación que se da entre ambas variables.

Ejemplo 5.15. *Supongamos que mediante una encuesta estadística estamos estudiando determinado atributo de la población (opina a favor o en contra), y deseamos saber si existen diferencias en las respuestas de los encuestados en función de su sexo.*

Para ello, realizamos una tabla cruzada de doble entrada en donde resumimos los resultados obtenidos en la encuesta:

Opiniones a favor y en contra en función del sexo

	Varón	Mujer	Total
A favor	32	10	42
En contra	11	27	38
Total	43	37	80

Las tablas de la forma del ejemplo reciben el nombre de Tablas de Contingencia, y sobre ellas se analizará la dependencia entre las respuestas dadas a las preguntas realizadas en relación con el sexo, utilizando el estadístico χ^2, y podemos evaluar el grado de relación que se da entre las opiniones y el sexo a partir de diferentes coeficientes de asociación como la Odds Ratio, el coeficiente de contingencia, el coeficiente V de Crammer o la Q de Yule.

Estadístico $\chi 2$

El estadístico χ^2 se calcula del siguiente modo:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Siendo:

— r el número de filas.
— s el número de columnas.
— Oij (frecuencia observada) el número de casos observados clasificados en la fila i de la columna j.
— Eij (frecuencia esperada) el número de casos esperados, en el supuesto de independencia, correspondientes a la fila i de la columna j.

Se define la *frecuencia esperada* como aquella frecuencia que se daría si los sucesos fueran independientes, es decir, hombres y mujeres manifestarían idénticas opiniones. Para calcular la frecuencia esperada o teórica de cada casilla (E_{ij}), se multiplican los dos totales marginales (fila y columna) y se divide este producto por el número total de casos.

Este estadístico toma valores mayores o iguales a cero, siendo cero en el caso de independencia absoluta y aumenta su valor en función del grado de dependencia entre ambas variables y del número de filas y columnas (o atributos de las variables).

Ejemplo 5.16. En el ejemplo anterior, calcular la tabla de frecuencias esperadas y el estadístico χ^2:

Frecuencias esperadas para las opiniones a favor y en contra en función del sexo

	Varón	Mujer	Total
A favor	23	19	42
En contra	20	18	38
Total	43	37	80

Calculándose del siguiente modo:

$$E_{11} = \frac{O_{1\cdot} O_{\cdot 1}}{O_{\cdot\cdot}} = \frac{42 \cdot 43}{80} = 23$$

$$E_{12} = \frac{O_{1\cdot} O_{\cdot 2}}{O_{\cdot\cdot}} = \frac{42 \cdot 37}{80} = 19$$

$$E_{21} = \frac{O_{2\cdot} O_{\cdot 1}}{O_{\cdot\cdot}} = \frac{38 \cdot 43}{80} = 20$$

$$E_{21} = \frac{O_{2\cdot} O_{\cdot 2}}{O_{\cdot\cdot}} = \frac{38 \cdot 37}{80} = 18$$

El valor de χ^2 para el ejemplo es:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \frac{(32-23)^2}{23} + \frac{(10-19)^2}{19} + \frac{(11-20)^2}{20} + \frac{(27-18)^2}{18} = 17{,}91$$

Lo que parece indicar que el sexo tiene influencia a la hora de estar a favor o en contra. Para establecer la intensidad de dicha dependencia, se utilizan las medidas de asociación, las cuales se estudian a continuación.

Medidas de asociación

La *Odds Ratio* se define como el cociente de las siguientes probabilidades:

$$OR = \frac{\frac{O_{11}}{O_{1\cdot}} / \frac{O_{12}}{O_{1\cdot}}}{\frac{O_{21}}{O_{2\cdot}} / \frac{O_{22}}{O_{2\cdot}}} = \frac{O_{11}O_{22}}{O_{12}O_{21}}$$

Si $OR > 1$ entonces la probabilidad de «a favor» es mayor en los hombres que en las mujeres, si $OR = 1$ ambas probabilidades son iguales (independencia en las opiniones de hombres y mujeres) y si $OR < 1$ la probabilidad de «a favor» es menor en los hombres que en las mujeres.

El valor de esta medida está comprendido en el intervalo $(0; \infty)$.

Las propiedades más relevantes son las siguientes:

1. Es invariante ante los cambios de escala en filas y columnas, o tan sólo en filas o en columnas.
2. Alcanza sus valores extremos, 0 e ∞, bajo asociación perfecta.
3. OR y 1/OR indican igual intensidad de la asociación, pero en direcciones opuestas.

Con objeto de lograr una interpretación más fácil, se define la siguiente medida:

$$OR' = \ln (OR)$$

La cual es una medida *simétrica* cuyo rango de variación es $(-\infty, +\infty)$, tomando el valor 0 en el caso de independencia y $-\infty$ o $+\infty$ en el caso de asociación perfecta.

El *coeficiente de contingencia C* es una medida del grado de asociación entre dos conjuntos de atributos, estén ordenados o no, e independiente de la naturaleza de la variable (continua o discreta). Es un estadístico que se obtiene de la tabla de contingencia mediante la siguiente fórmula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

El valor de este coeficiente está entre 0 y 1, los valores más próximos a 1 indicarían un mayor grado de interdependencia entre variables. Lógicamente, este coeficiente nunca puede alcanzar el valor 1, aunque haya completa asociación.

El *coeficiente V de Cramer*, es otro estadístico que se obtiene a partir de la χ^2. Su valor oscila entre 0 y 1, siendo 0 cuando la independencia es completa y 1 cuando se da una completa asociación. Se obtiene a partir de:

$$V = \sqrt{\frac{\chi^2}{n * \min(r-1, s-1)}}$$

Otra medida de asociación es la *Q de Yule* que se calcula sobre las diferencias entre las frecuencias observadas (O_{ij}) y esperadas (E_{ij}). En una tabla 2×2 la medida Q de Yule se calcula a través de la siguiente expresión:

$$Q = \frac{nD_{11}}{O_{11}O_{22} - O_{12}O_{21}}$$

donde $$D_{11} = O_{11} - E_{11}$$

La Q de Yule está comprendida entre -1 y 1, siendo los criterios interpretativos:

— $Q = 0$ independencia.
— $Q > 0$ asociación positiva.
— $Q < 0$ asociación negativa.

Ejemplo 5.17. Calcular las medidas de asociación para la tabla de contingencia planteada.

Diferencias entre las frecuencias observadas y esperadas para las opiniones a favor y en contra en función del sexo

	Nivel 1	**Nivel 2**
Nivel 1	$D_{11} = O_{11} - E_{11} = 9$	$D_{12} = O_{12} - E_{12} = -9$
Nivel 2	$D_{21} = O_{21} - E_{21} = -9$	$D_{22} = O_{22} - E_{22} = 9$

En el ejemplo, *OR* y *OR'* toman los siguientes valores:

$$OR = \frac{32 \cdot 27}{10 \cdot 11} = 7{,}85 \qquad OR' = \ln\left(\frac{32 \cdot 27}{10 \cdot 11}\right) = 2{,}06$$

lo cual quiere decir que los hombres muestran una opinión más favorable que las mujeres.

El valor del coeficiente de contingencia es igual a:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{17{,}91}{17{,}91 + 80}} = 0{,}4277$$

lo cual indica un grado de asociación medianamente alto.

El valor *V de Cramer* será de:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(k-1, r-1)}} = \sqrt{\frac{17{,}91}{80}} = 0{,}4732$$

Y por último, la *Q de Yule* toma el valor:

$$Q = \frac{nD_{11}}{O_{11}O_{22} - O_{12}O_{21}} = \frac{80 \cdot 9}{32 \cdot 27 + 10 \cdot 11} = 0{,}74$$

5.13. EL TRATAMIENTO DE LAS DISTRIBUCIONES BIDIMENSIONALES Y DE LA REGRESIÓN EN HOJAS DE CÁLCULO EXCEL Y EN SPSS

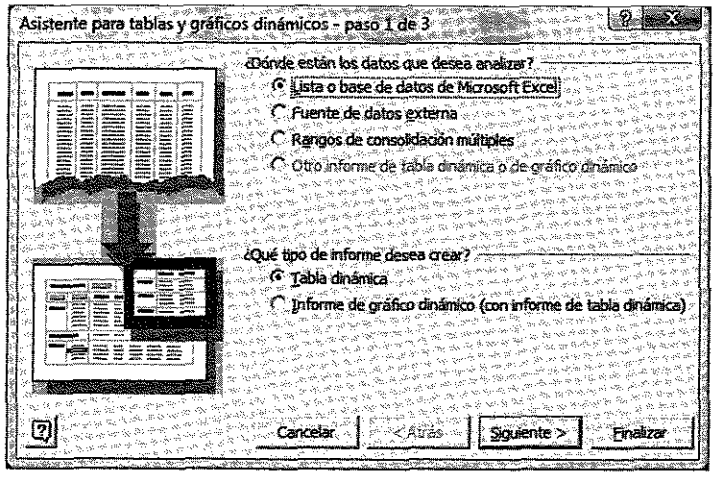
5.13.1. El tratamiento de las distribuciones bidimensionales y de la regresión en Excel

Para el tratamiento de las distribuciones bidimensionales, la hoja de cálculo Excel cuenta, entre otros instrumentos, con la herramienta denominada «Tablas Dinámicas», cuyo funcionamiento el alumno debe conocer.

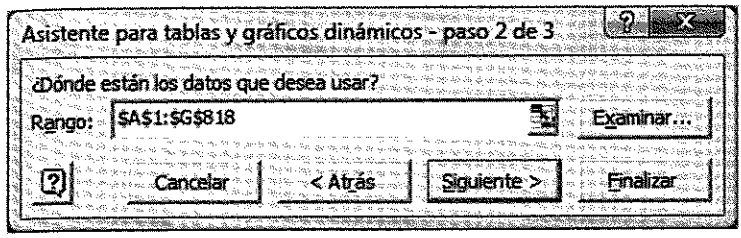
Se trata de unos tipos especiales de tablas que resumen información por campos de una lista o base de datos. Con esta opción pueden realizarse tablas cruzadas de datos, en donde aparecen por filas los valores que toma un rango y por columnas los de otro rango. El calificativo de dinámicas se debe a que una vez hecha una tabla hay diversas posibilidades de modificarla. Se utiliza el *Asistente para Tablas Dinámicas* para la creación de una tabla.

El *Asistente de Tablas* guía paso a paso la creación de la Tabla mediante cuadros de diálogo en 3 pasos sucesivos.

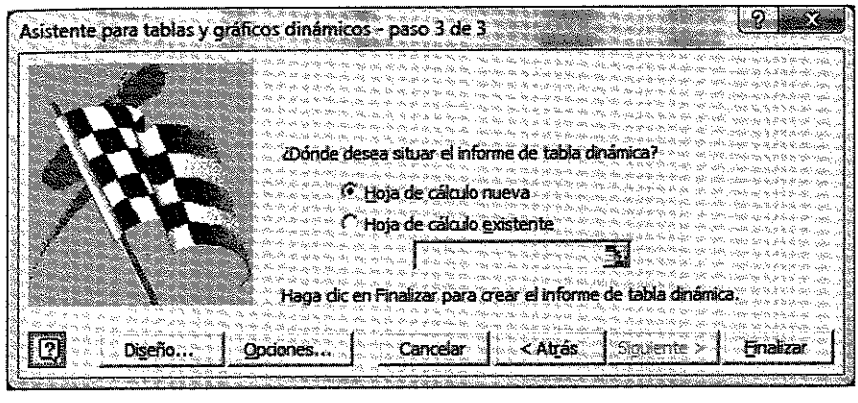
Paso 1. Se indica la fuente de datos que se va a utilizar para crear la tabla.



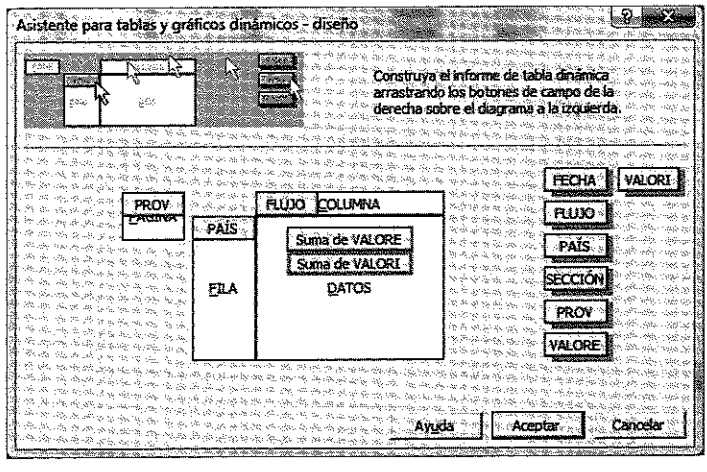
Paso 2. El rango de la ubicación de los datos fuente.



Paso 3. Se especifica la ubicación del informe o tabla dinámica y el diseño y opciones.



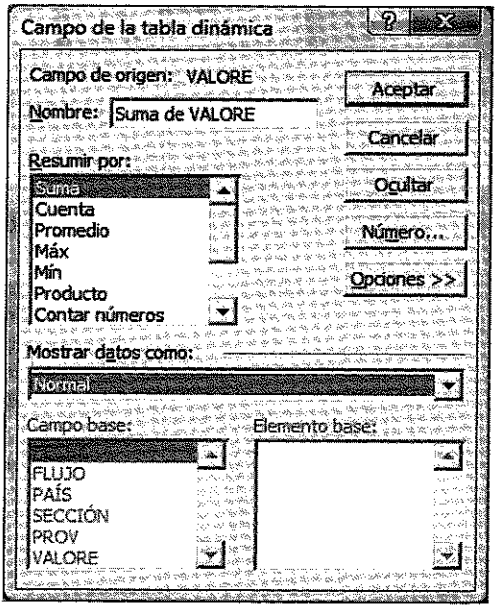
Respecto al diseño, se organiza la tabla arrastrando los campos de la lista a las filas, columnas y página.



Destacar que la opción *Área Página* nos permite seleccionar un campo en el que para cada dato la tabla mostrará el resultado del cruce de datos según los criterios especificados en filas y columnas, pero sólo uno cada vez.

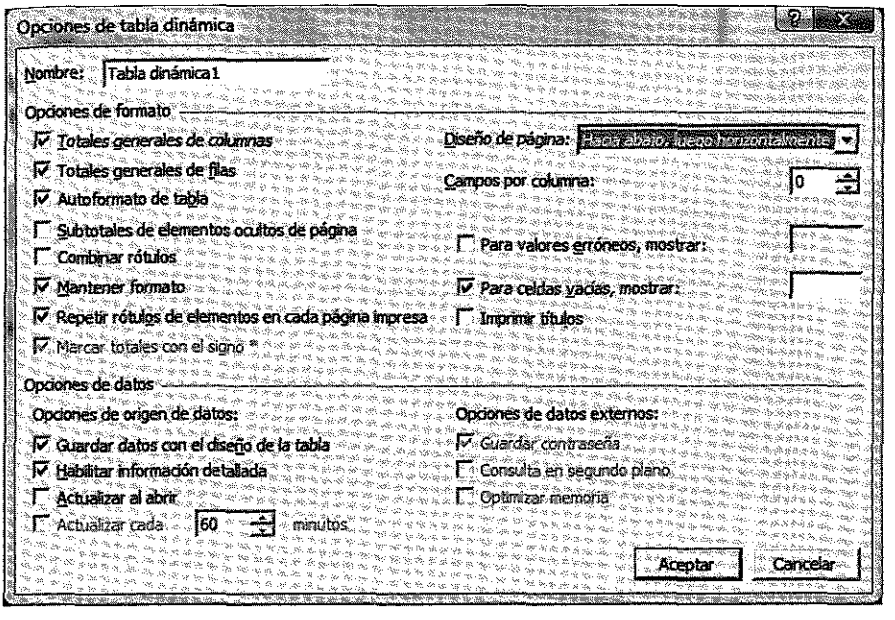
Haciendo doble click en los campos de datos podemos modificar los tipos de análisis que realizamos con los datos, donde las opciones disponibles son:

— Suma.
— Cuenta.
— Promedio.
— Máximo.
— Mínimo.
— Producto.
— Desviación típica.
— Varianza.



En *Opciones* se puede elegir una presentación en forma de valor, de porcentaje de fila, de columna o del total, o bien como diferencia respecto a un elemento base del campo.

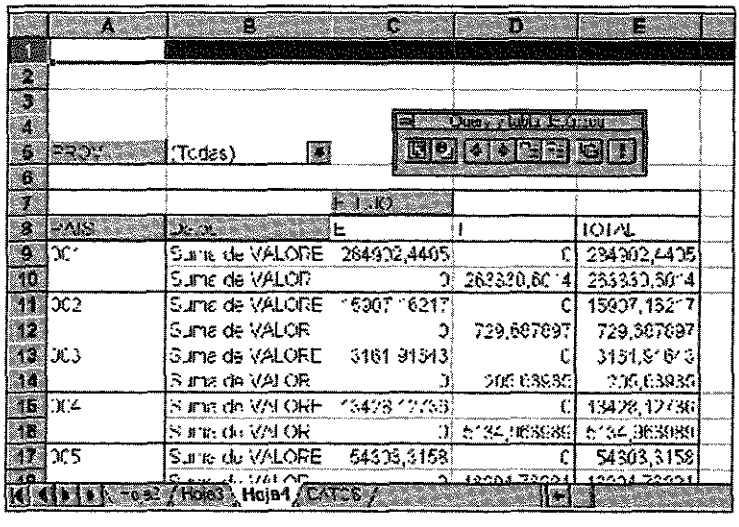
Se muestran a continuación las opciones de presentación disponibles.



Una vez creada la tabla dinámica, se puede modificar con las siguientes acciones:

— Introducir campos de datos adicionales.
— Eliminar un campo.
— Reorganizar campos y elementos de la tabla.
— Cambiar la forma en que se calculan los datos.
— Modificar los nombres de los campos y elementos.
— Cambiar el formato.
— Ocultar y mostrar elementos de los campos.
— Ocultar y mostrar los datos detallados.
— Agrupar elementos.
— Ordenar los elementos de un campo.
— Calcular totales o subtotales.

Destacar por último, que las tablas dinámicas una vez realizadas permiten desplegar una barra de herramientas desde donde se realizan las acciones señaladas. Este menú aparece de forma automática cuando se crea una tabla.



Por su parte, para el **cálculo de covarianzas y de coeficientes de correlación y de regresión se utilizan** las siguientes funciones:

=COVAR(matriz1;matriz2)
=COEF.DE.CORREL(matriz1;matriz2)
=ESTIMACION.LINEAL(conocido_y;conocido_x;constante;estadística)
=ESTIMACION.LOGARITMICA(conocido_y;conocido_x;constante;estadística)
=TENDENCIA(conocido_y;conocido_x;nueva_matriz_x; constante)
=PENDIENTE(conocido_y;conocido_x)
=INTERSECCION.EJE(conocido_y;conocido_x)

Veamos con mayor detalle la forma en la que se realiza una estimación lineal.

La función ESTIMACION.LINEAL calcula las estadísticas de una regresión lineal utilizando el método de «mínimos de cuadrados» y devuelve una matriz que describe la línea. Puesto que esta función devuelve una matriz de valores, debe introducirse como fórmula matricial.

La ecuación para la línea es: $Y = m_1X_1 + m_2X_2 + \ldots + m_nX_n + b$

La matriz que devuelve ESTIMACION.LINEAL es $\{m_n, m_{n-1}, \ldots, m_1, b\}$, aunque ESTIMACION.LINEAL también puede devolver estadísticas de regresión adicionales, entre ellas el R cuadrado.

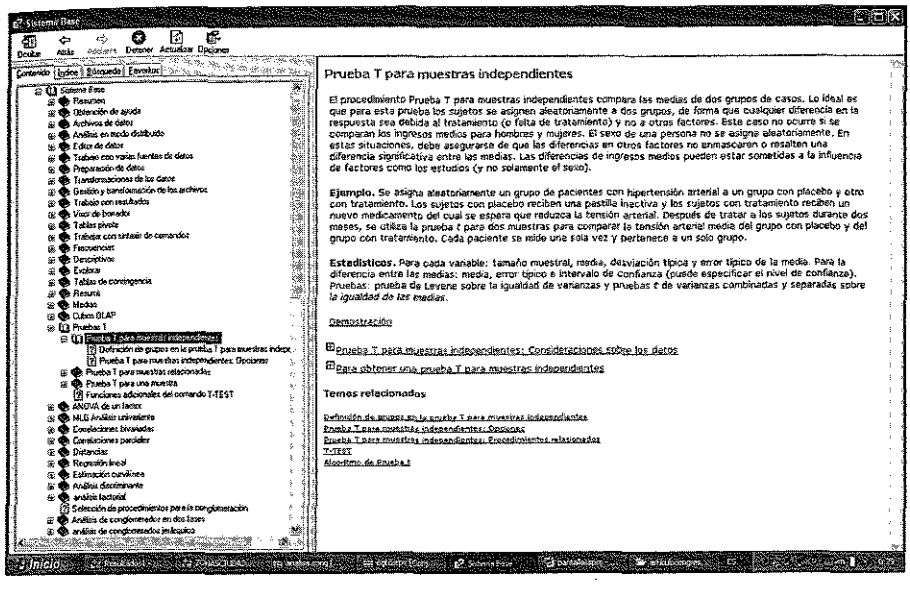
5.13.2. El tratamiento de las distribuciones bidimensionales y de la regresión en SPSS

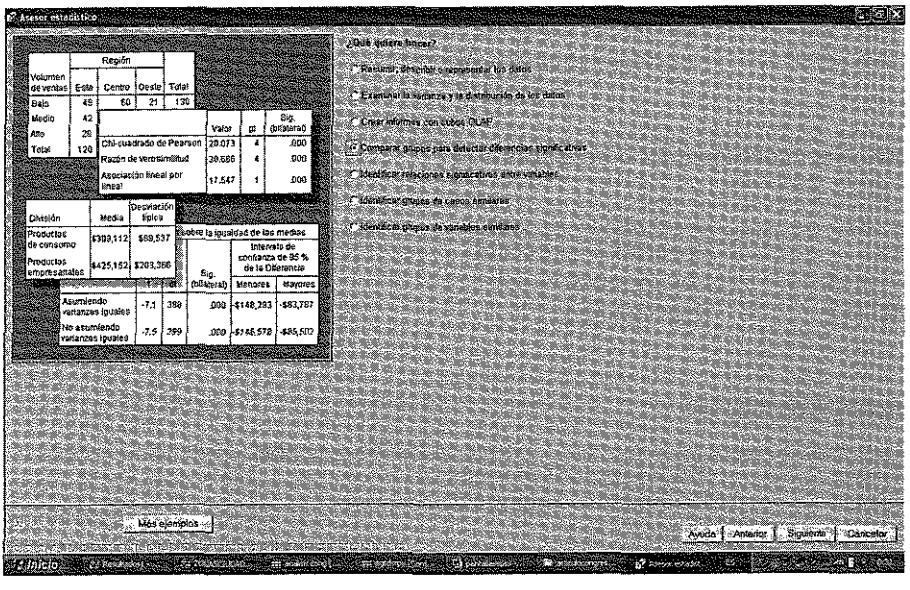
Entre las múltiples posibilidades que ofrece el SPSS para el tratamiento de variables bidimensionales, haremos mención de tres herramientas: el asesor de estadística, las tablas personalizadas y la regresión.

El Asesor de Estadística

SPSS dispone de una ayuda especial que denomina *El Asesor de Estadística* que nos puede ser útil. El Asesor estadístico muestra una serie de preguntas diseñadas para encontrar el procedimiento adecuado. La primera pregunta es simplemente «¿Qué quiere hacer?».

Por ejemplo, si desea realizar un contaste de la T para muestras independientes el Asesor de Estadística le presenta la siguiente información:





El programa SPSS, como otros muchos programas, puede transformar y generar datos utilizando operadores y funciones.

Los operadores de los que dispone son de tres tipos: aritméticos, relacionales y lógicos. Algunas de los operadores que pueden ser interesantes son:

Operadores aritméticos

x+y	Suma
x-y	Resta
x*y	Multiplicación
x/y	División
x**y	Potencia (Exponenciación)

Operadores relacionales

x = y ó x EQ y	Igualdad
x <> y o x = y O x NE y	Desigualdad
x > y o x GT y	Mayor que
x ≥ y o x GE y	Mayor o igual que
x < y o x LT y	Menor que
x ≤ y o x LE y	Menor o igual que

Operadores lógicos

x & y ó x AND y	Conjunción booleana (AND)
x / y ó x OR y	Disyunción booleana (OR)
≂x ó NOT x	Negación booleana (NOT)

Funciones exponenciales y logarítimicas

EXP(x)	Exponencial de base e (2.71828)
LN(x)	Logaritmo neperiano
LG 10(x)	Logaritmo decimal
SQRT(x)	Raíz cuadrada

Funciones numéricas

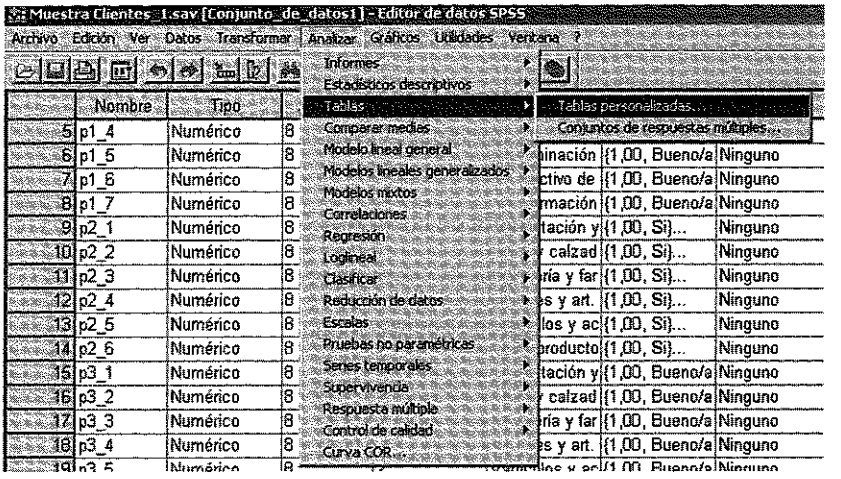
ABS(x)	Valor absoluto (o módulo)
RND(x)	Redondeo al entero más cercano
MOD(x;y)	Resto de X/Y
TRUNC(x)	Halla el menor entero mayor o igual que X

Funciones de estadística descriptiva

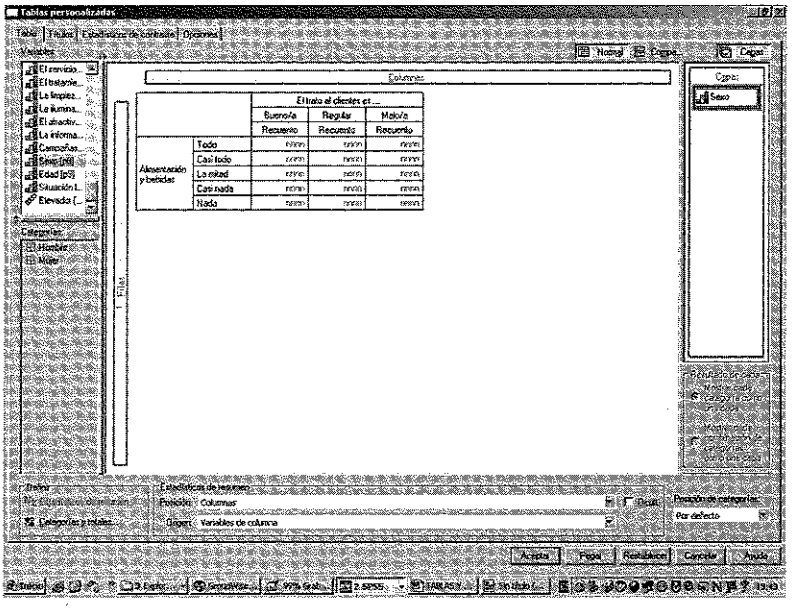
SUM (A)	Halla la suma de las observaciones de la variable A
SUM (A, B, C ...)	Halla el vector de las sumas de las observaciones de las variables A, B, C...
MEAN(A)	Halla la media de la variable A
MEAN(A, B, C, ...)	Halla el vector de las medias de las variables A, B, C, ...
SD(A)	Halla la desviación típica de la variable A
SD(A, B, C, ...)	Halla el vector de las desviaciones típicas de las variables A, B, C, ...
VARIANCE(A)	Halla la varianza de la variable A.
VARIANCE(A, B, C,)	Halla el vector de las varianzas de las variables A, B, C,
CFVAR(A)	Halla el coeficiente de variación de la variable A.
CFVAR(A, B, C, ...)	Halla el vector de coeficientes de variación de las variables A, B, C, ...
MAX(A)	Halla el máximo de las observaciones de la variable A.
MAX(A, B, C, ...)	Halla el vector de los máximos de las observaciones de las variables A, B, C, ...
MIN(A)	Halla el mínimo de las observaciones de la variable A.
MIN(A, B, C, ...)	Halla el vector de los mínimos de las observaciones de las variables A, B, C, ...
LAG(numvar;n)	Desplaza el comienzo de la variable numérica numvar n posiciones hacia adelante y sustituye las n primeras posiciones por valores desaparecidos. se trata de la típica variable retardo de orden N.

Las Tablas Personalizadas

En este apartado se muestra lo sencillo que es la realización de tablas en SPSS. Se trabaja con el fichero «**MuestraClientes_1.sav**». Una vez cargado el fichero seleccionamos **Analizar/Tablas/Tablas Personalizadas** tal y como se observa en la figura de la página siguiente.



La pantalla a la que se accede es la que se muestra a continuación. En este ejemplo se ha elegido de la variable **P4,** el sector Alimentación y bebidas para las filas de la tabla y la variable **P1,** sobre el trato del cliente, para visualizarla en las columnas de la tabla.



En el apartado de **Capas** se ha incluido la variable **Sexo** para poder obtener las tres tablas que se muestran a continuación:

Sexo Total

		El trato al cliente es		
		Bueno/a	Regular	Mal/a
		Recuento	Recuento	Recuento
Alimentación y bebidas	Todo	78	5	1
	Casi todo	57	13	0
	La mitad	26	9	0
	Casi nada	11	0	0
	Nada	0	0	0

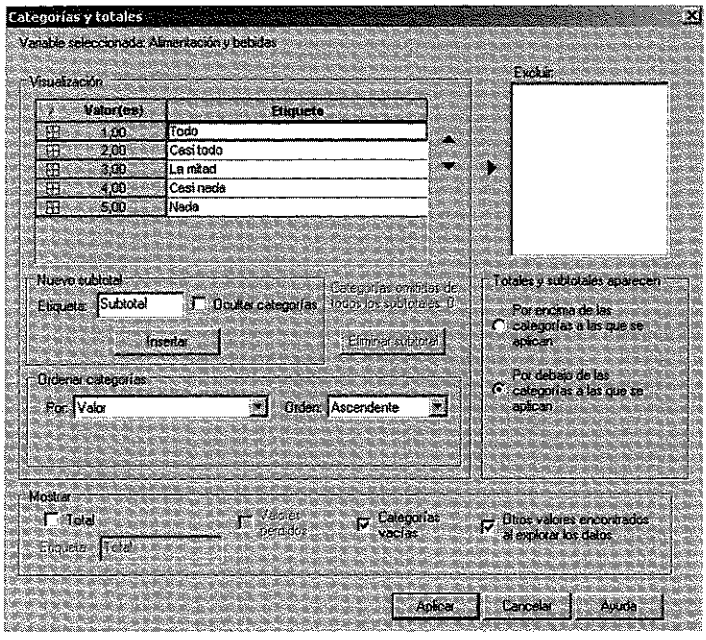
Sexo Hombre

		El trato al cliente es		
		Bueno/a	Regular	Mal/a
		Recuento	Recuento	Recuento
Alimentación y bebidas	Todo	33	2	0
	Casi todo	34	4	0
	La mitad	13	5	0
	Casi nada	6	0	0
	Nada	0	0	0

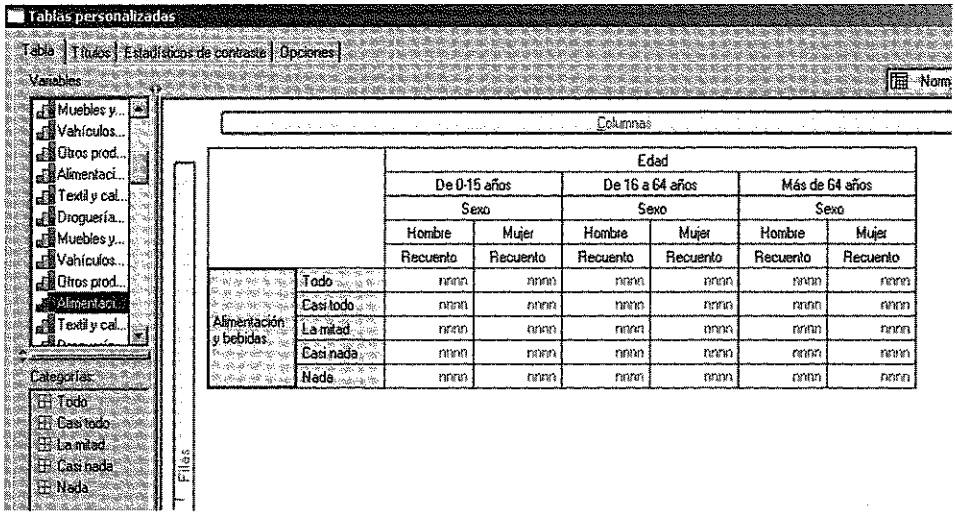
Sexo Mujer

		El trato al cliente es		
		Bueno/a	Regular	Mal/a
		Recuento	Recuento	Recuento
Alimentación y bebidas	Todo	39	3	1
	Casi todo	23	9	0
	La mitad	13	4	0
	Casi nada	6	0	0
	Nada	0	0	0

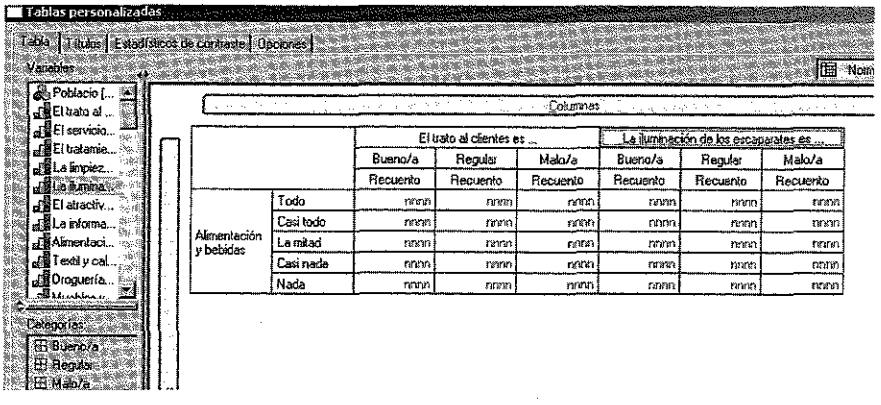
Si accedemos al menú de **Categorías y totales** podemos mostrar subtotales, ocultar u ordenar las categorías, poner totales y otras operaciones que pueden observarse en el dibujo siguiente:



Podemos obtener tablas con datos anidados o, si lo queremos, podemos sacar diferentes tablas juntas. Para ver esto utilizamos la variable *Edad* y la variable *Sexo* para tablas anidadas. Estas variables las trasladamos a la parte de las Columnas. De forma gráfica la forma de proceder y los resultados se muestran en las dos imágenes siguientes:

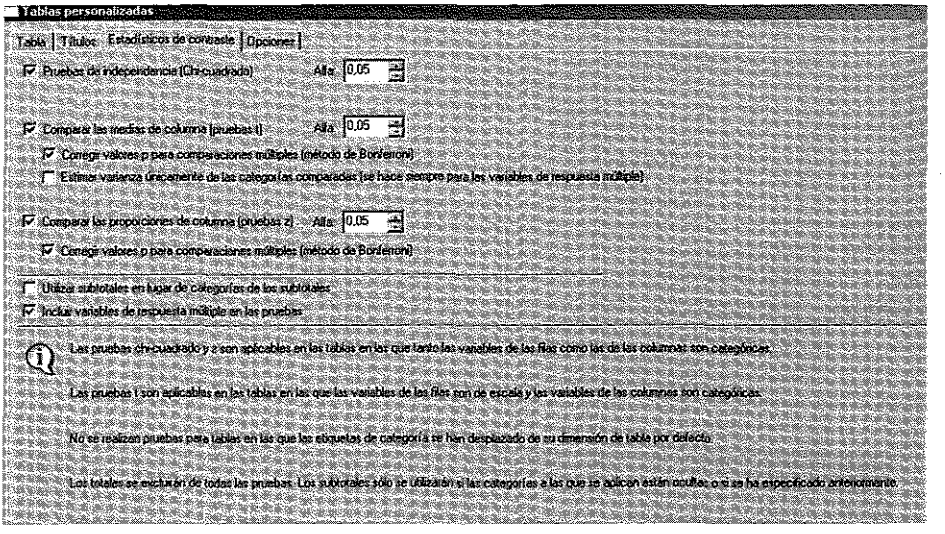


		Edad					
		De 0-15 años		De 16-64 años		Más de 64 años	
		Sexo		Sexo		Sexo	
		Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
		Recuento	Recuento	Recuento	Recuento	Recuento	Recuento
Alimenta-	Todo	0	3	28	27	13	13
ción	Casi todo	4	5	29	20	5	7
y bebidas	La mitad	0	3	15	12	3	2
	Casi nada	1	4	5	1	0	0
	Nada	0	0	0	0	0	0



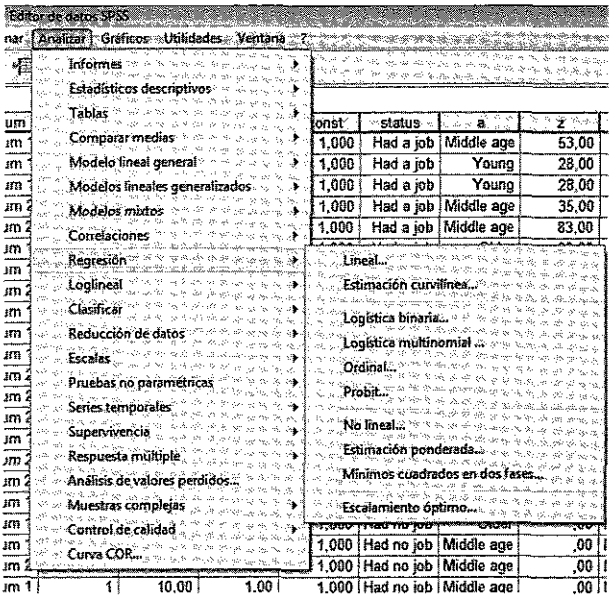
		El trato al cliente es			La iluminación de los escaparates es		
		Hombre	Mujer	Hombre	Mujer	Hombre	Mujer
		Recuento	Recuento	Recuento	Recuento	Recuento	Recuento
Alimenta-	Todo	78	5	1	64	18	0
ción	Casi todo	57	13	0	44	17	6
y bebidas	La mitad	26	9	0	11	12	12
	Casi nada	11	0	0	5	0	6
	Nada	0	0	0	0	0	0

También pueden llevarse a cabo diferentes contrastes estadísticos como la comparación de medias, proporciones y el contraste de independencia, tal y como se muestra en la siguiente pantalla:

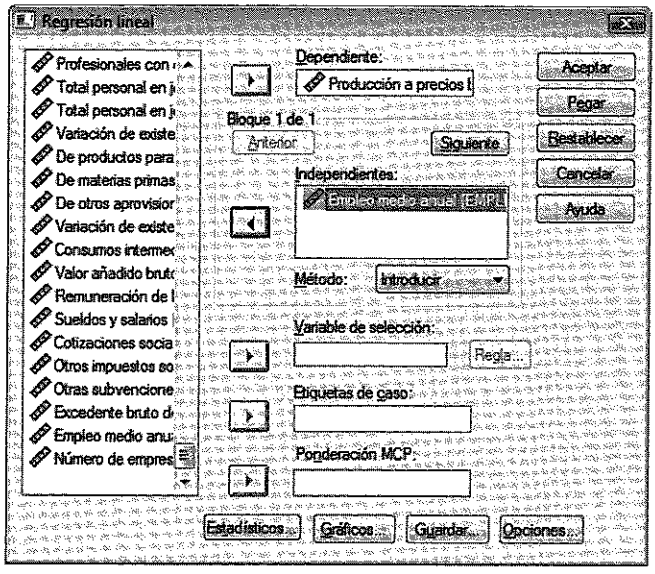


La regresión

Para la realización de análisis de regresión el SPSS dispone de un menú específico, el cual, cuenta con las opciones que se muestran en la figura siguiente.



Como ejemplo, disponemos de los datos de producción y empleo de un conjunto de empresas. Arrastrando estas variables, o a través de los botones, especificamos la variable dependiente e independiente.



Pulsando en aceptar, obtenemos los siguientes resultados.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregido	Error tip. de la estimación
1	394[a]	,155	,155	1223.204.25

[a] Varibles predictoras (Constante). Empleo medio anual.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4,20E + 009	1	4,2E+009	2803.928	,000[a]
	Residual	2,28E + 010	15233	1496228,6		
	Total	2,70E + 010	15234			

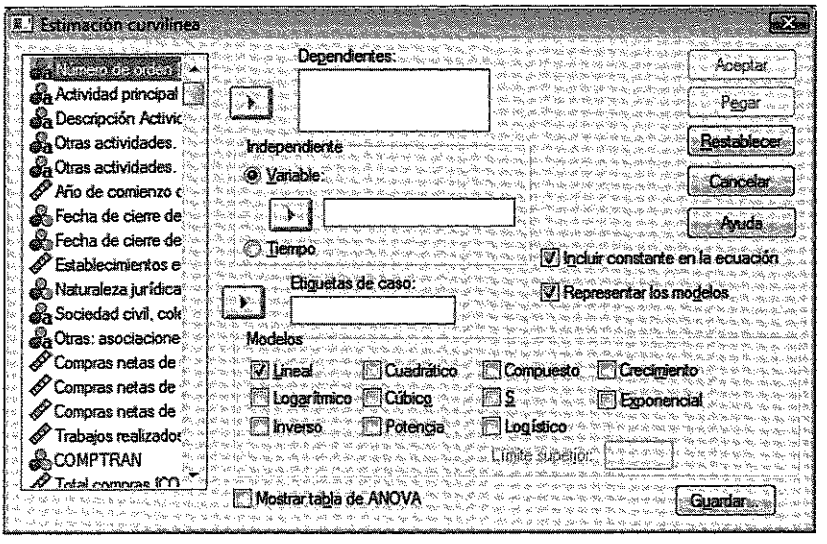
[a] Variables predictoras (Constante). Empleo medio anual.
[b] Variable dependiente. Producción a precios básicos.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error tip.	Beta		
1	(Constante	−2786.593	57.451		−48.504	,000
	Empleo medio anual	2930.304	55.339	,394	52.952	,000

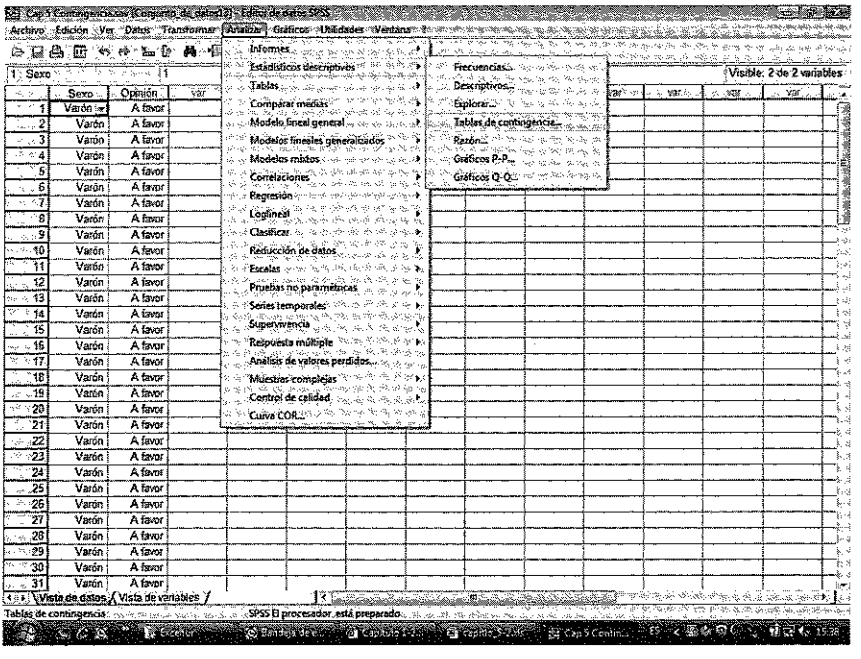
[b] Variable dependiente. Producción a precios básicos.

Por su parte, la estimación curvilínea, dispone de las siguientes opciones.

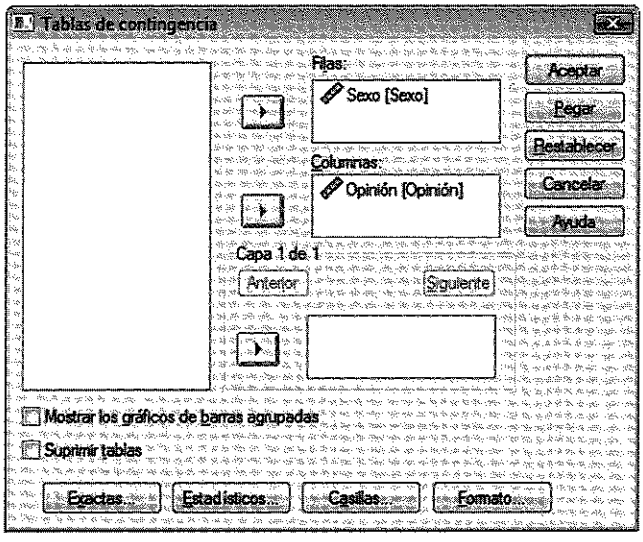


Tablas de contingencia

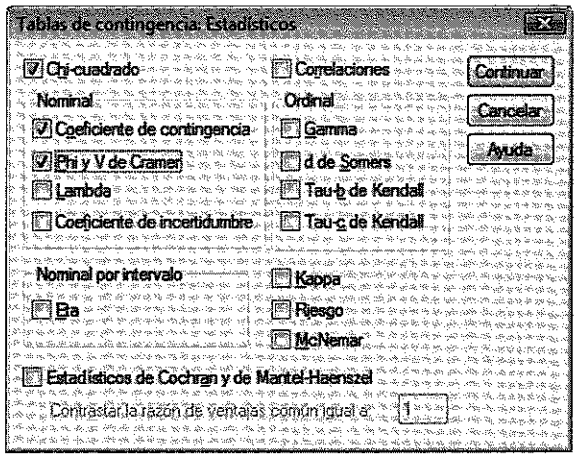
El SPSS dispone de un submenú específico para la elaboración de tablas de contingencia, accesible a través del menú Analizar->Estadísticos descriptivos.



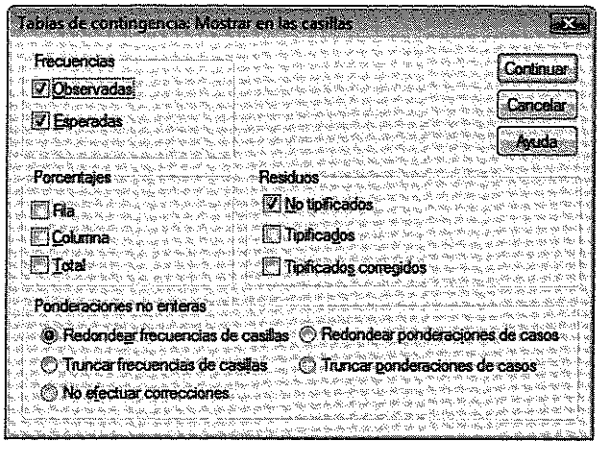
Tomando como referencia los datos del ejemplo 5.16, en primer lugar especificamos la variable fila y columna.



Pulsando en el menú *Estadísticos* señalamos las siguientes opciones.



Por último, en *Casillas* solicitamos además de las frecuencias observadas, las esperadas y la diferencia entre ambas.



Los resultados obtenidos serían los siguientes:

Tabla de contingencia Sexo * Opinión

			Opinión		Total
			A favor	En contra	
Sexo	Varón	Recuento	32	11	43
		Frecuencia esperada	22,6	20,4	43,0
		Residuo	9,4	−9,4	
	Mujer	Recuento	10	27	37
		Frecuencia esperada	19,4	17,6	37,0
		Residuo	−9,4	9,4	
Total		Recuento	42	38	80
		Frecuencia esperada	42,0	38,0	80,0

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	17.911[b]	1	,000		
Corrección por continuidad[a]	16.061	1	,000		
Razón de verosimilitudes	18.620	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	17,688	1	,000		
N de casos válidos	80				

[a] Calculado solo para una tabla de 2 × 2.

[b] 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 17,58.

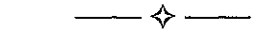
Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,473	,000
	V de Cramer	,473	,000
	Coeficiente de contingencia	,428	,000
N de casos válidos		80	

[a] Asumiendo la hipótesis alternativa.
[b] Empleando el error típico asintótico basado en la hipótesis nula.

5.14. EJERCICIOS

De distribuciones bidimensionales



Ejercicio 5.1. *Represente gráficamente las siguientes distribuciones de frecuencias bidimensionales:*

a) *En la tabla siguiente se indican los resultados de una encuesta realizada a 500 personas a las que se les ha preguntado el sexo y la línea aérea de la cual son usuarios.*

	Línea A	Línea B	Línea C	Línea D	Total
Varón	60	75	110	55	300
Mujer	20	45	90	45	200
Total	80	120	200	100	500

b) *En la tabla siguiente se indican los resultados de una encuesta realizada a 300 personas a las que se les ha preguntado su edad y la ciudad en la que veranean.*

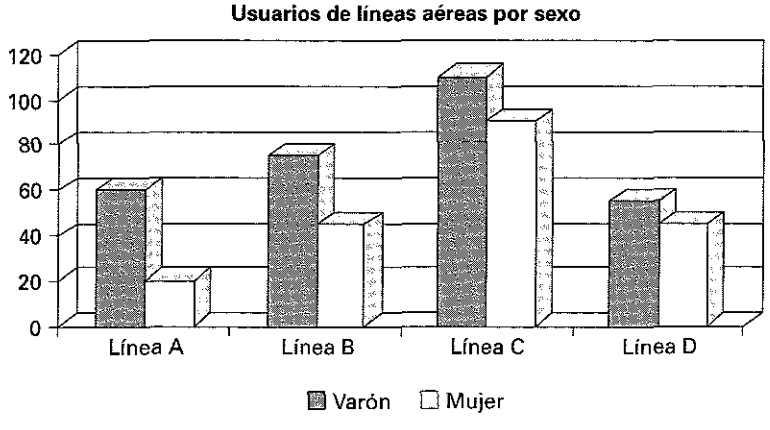
Edad	Marbella	Torremolinos	Mijas-Costa	Total
20-30	35	20	5	60
30-40	30	30	10	70
40-50	25	40	15	80
50-60	20	50	20	90
Total	110	140	50	300

c) *En la tabla siguiente se consignan los datos para determinar la relación entre los gastos de publicidad semanal y las ventas realizadas.*

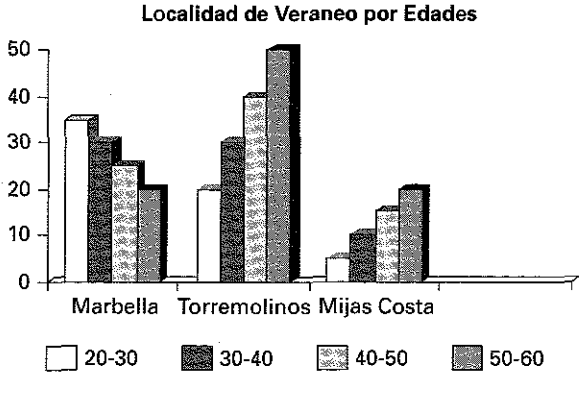
Coste de publicidad (€)	Ventas (€)	Coste de publicidad (€)	Ventas (€)
40	385	40	490
20	400	20	420
25	395	50	560
20	365	40	525
30	475	25	480
50	440	50	510

Respuesta

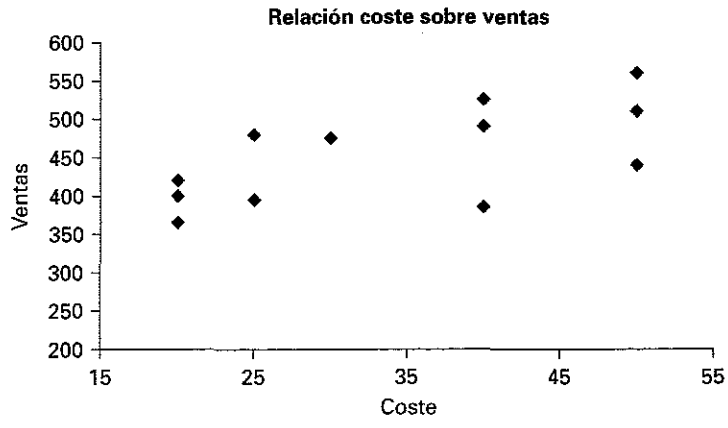
a)



b)



c)



Ejercicio 5.2. *Con los datos del ejercicio 5.1 b):*

a) *Obtenga una tabla de correlación de frecuencias relativas.*

b) *Halle la distribución de frecuencias marginales, la moda de la variable* y *(ciudad de veraneo) y la media de la variable* x *(edad).*

c) *Obtenga la distribución de* y *condicionada a* $x = 35$.

Respuesta

a) Sea x_i las marcas de clase. Las frecuencias relativas se obtienen haciendo $f_{ij} = \frac{n_{ij}}{N}$.

x_i	Marbella	Torremolinos	Mijas-Costa	$n_{i\cdot}$	$f_{i\cdot}$
25	0,12	0,07	0,02	60	0,2
35	0,1	0,1	0,03	70	0,23
45	0,08	0,13	0,05	80	0,27
55	0,07	0,17	0,07	90	0,3
$n_{\cdot j}$	110	140	50		
$f_{\cdot j}$	0,37	0,47	0,17		

b) Las distribuciones de frecuencias marginales son:

x_i	$n_{i\cdot}$
25	60
35	70
45	80
55	90

y_j	$n_{\cdot j}$
Marbella	110
Torremolinos	140
Mijas-Costa	50

La moda de y es Torremolinos y la media aritmética de x es:

$$\overline{x} = \frac{1}{N}\sum x_i\, n_{i\cdot} = 41{,}66 \text{ años}$$

c)

$Y = y_j / X = 35$	n_{2j}
Marbella	30
Torremolinos	30
Mijas-Costa	10

Ejercicio 5.3. *Dada la siguiente tabla:*

x_i \ Y_j	1	2	3	$n_{i\cdot}$
0	3	0	1	4
3	0	4	2	6
5	1	1	6	8
$n_{\cdot j}$	4	5	9	$N = 18$

a) *Calcule las distribuciones marginales de frecuencias.*

b) *Calcule las medias, desviaciones típicas marginales y la covarianza.*

c) *Calcule el coeficiente de correlación de Pearson.*

Respuesta

a)

x_i	0	3	5
$n_{i\cdot}$	4	6	8

y_j	1	2	3
$n_{\cdot j}$	4	5	9

b)

$$\bar{x} = a_{10} = \frac{\sum_1^3 x_i n_{i\cdot}}{N} = \frac{0\cdot 4 + 3\cdot 6 + 5\cdot 8}{18} = \frac{58}{18} = \frac{29}{9}$$

$$\bar{y} = a_{01} = \frac{\sum y_j n_{\cdot j}}{N} = \frac{1\cdot 4 + 2\cdot 5 + 3\cdot 9}{18} = \frac{41}{18}$$

$$s_x^2 = m_{20} = a_{20} - (a_{10})^2 = \frac{\sum x_i^2 n_{i\cdot}}{N} - \left(\frac{\sum x_i n_{i\cdot}}{N}\right)^2 =$$

$$= \frac{3^2\cdot 6 + 5^2\cdot 8}{18} - \left(\frac{29}{9}\right)^2 = 3{,}73$$

$$s_x = \sqrt{3{,}73} = 1{,}9132$$

$$s_y^2 = m_{02} = a_{02} - (a_{01})^2 = \frac{\sum y_j^2 n_{\cdot j}}{N} - \left(\frac{\sum y_j n_{\cdot j}}{N}\right)^2 =$$

$$= \frac{1^2\cdot 4 + 2^2\cdot 4 + 3^2\cdot 9}{18} - \left(\frac{41}{18}\right)^2 = 0{,}422939$$

$$s_y = \sqrt{0{,}422939} = 0{,}65033$$

$$\mathrm{Cov}(xy) = m_{11} = a_{11} - a_{10}\cdot a_{01} = \frac{\sum x_i y_j n_{ij}}{N} - (\bar{x}\cdot\bar{y}) =$$

$$= \frac{0\cdot 1\cdot 3 + 0\cdot 2\cdot 0 + 0\cdot 3\cdot 1 + 3\cdot 1\cdot 0 + 3\cdot 2\cdot 4 + 3\cdot 3\cdot 2 + 5\cdot 1\cdot 1}{18} + \frac{5\cdot 2\cdot 1 + 5\cdot 3\cdot 6}{18}$$

$$- \left(\frac{29}{9}\cdot\frac{41}{18}\right) = \frac{147}{18} - \frac{1.189}{162} = 0{,}82716$$

c)

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{m_{11}}{\sqrt{m_{20} \cdot m_{02}}} = \frac{0,82716}{1,9132 \cdot 0,65033} = 0,66$$

Ejercicio 5.4. *Una compañía quiere realizar un estudio sobre la influencia del gasto en I+D sobre sus ventas. Para ello dispone de los siguientes datos sobre los últimos años:*

Años	Gastos millones	Ventas millones
2006	3,0	130
2007	3,3	155
2008	3,8	175
2009	4,2	210

a) *Realice un gráfico de dispersión.*

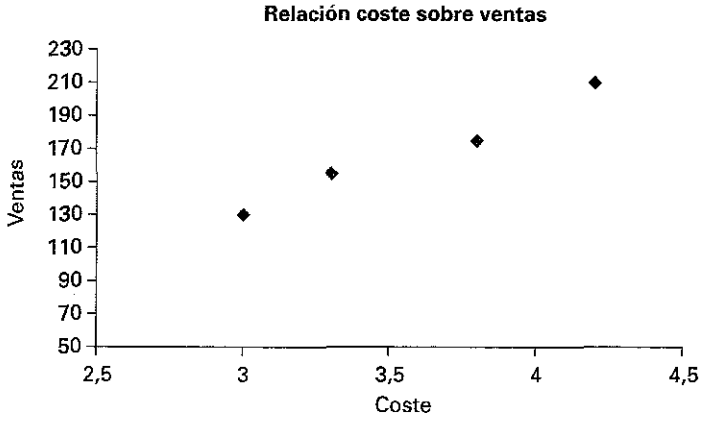
b) *Obtenga un modelo lineal que permita predecir las ventas a partir de los gastos en I+D. Comente los resultados.*

c) *Prediga las ventas del 2010 sabiendo que el gasto en I+D será de 4,5 millones.*

d) *Juzgue la bondad del modelo estimado.*

Respuesta

a)



b) Llamaremos x a los Gastos e y a las Ventas, por lo tanto ajustaremos una recta de Y sobre X.

x (Gastos)	y (Ventas)	x^2	y^2	$x \cdot y$
3	130	9,00	16.900	390
3,3	155	10,89	24.025	511,5
3,8	175	14,44	30.625	665
4,2	210	17,64	44.100	882
14,3	**670**	**51,97**	**115.650**	**2.448,5**

$$a_{10} = \frac{14,3}{5} = 2,86.$$

$a_{01} = 134.$

$a_{11} = 489,7.$

$a_{20} = 10,394.$

$a_{02} = 23.130.$

$m_{20} = 2,2144.$

$m_{02} = 5.174.$

$m_{11} = 106,46.$

$$b = \frac{106,46}{2,2144} = 48,076.$$

El coeficiente de regresión nos da la medida en que aumentarán las ventas al aumentar en un millón los gastos en I+D.

$$y - 134 = 48,076\ (x - 2,86) \Rightarrow y = 48,076\ x - 3,49736$$

Por lo tanto, la recta de regresión de y sobre x será:

c) Las ventas estimadas para 2010 son:

$$y_{2010} = -48,076\ (4,5) - 3,49736 = 212,84 \text{ millones}$$

d)

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = \frac{m_{11}^2}{m_{20} \cdot m_{02}} = \frac{(106,46)^2}{(2,2144 \cdot 5174)} = 0,9892$$

El 98,92% de la varianza de y está explicada por x, a través de la función ajustada y_{ti}, por lo que podemos decir que el modelo lineal es aceptable.

Ejercicio 5.5. *En Mercamadrid se ha observado durante un periodo de tiempo, las cantidades de Kg. vendidos de un producto y el precio correspondiente en euros, obteniéndose los siguientes resultados:*

$$\sum_1^6 y = 21; \quad \sum_1^6 x = 84; \quad \sum_1^6 xy = 273; \quad \sum_1^6 x^2 = 1.202; \quad \sum_1^6 y^2 = 91$$

Calcule:

a) *La recta de regresión de* Y *sobre* X.

b) *La varianza de* Y, *la varianza explicada por la regresión y la varianza residual.*

c) *El coeficiente de determinación.*

d) *¿Qué cantidad de producto se vendería a un precio de 10 €/Kg?*

Respuesta

a) $a_{01} = 3{,}5.$

$a_{10} = 14.$

$a_{11} = 45{,}5.$

$a_{20} = 200{,}\widehat{33}.$

$a_{02} = 15{,}1\hat{6}.$

$m_{20} = 4{,}\widehat{33}.$

$m_{02} = 2{,}9166.$

$m_{11} = -3{,}5.$

$$b = \frac{-3{,}5}{4{,}\widehat{33}} = -0{,}80798.$$

Esto significa que por cada euro por kilo que se aumente, las ventas se reducirán en 81 céntimos de euro.

$$y - 3{,}5 = -0{,}808(x - 14) \quad \Rightarrow \quad y = -0{,}808x + 14{,}814$$

b) $S_y^2 = 15{,}16 - (3{,}5)^2 = 2{,}9166.$

$$S_e^2 = m_{02}\frac{m_{11}^2}{m_{02}} = 0{,}08965.$$

$$S_{y^*}^2 = S_y^2 - S_e^2 = 2{,}82695.$$

c) La varianza de la variable endógena (y) y la varianza explicada por la regresión (y_{ti}) están muy aproximadas. Esto significa que al calcular el coeficiente de determinación el valor esté muy próximo a 1.

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{0{,}08965}{2{,}9166} = 0{,}9662$$

El 96,62% de la varianza de la variable endógena está explicada por la variable dependiente a través de la función ajustada. El modelo lineal es aceptable.

d) Si $x = 10$ €/Kg.

$$y = -0{,}808 \cdot (10) - 14{,}814 \quad \Rightarrow \quad y = 6{,}734 \text{ Kg.}$$

Ejercicio 5.6. *Se dispone de la siguiente información relativa a dos variables:*

$$\sum y_i = 186{,}95; \quad \sum x_i = 55; \quad \sum y_i^2 = 4.144{,}07; \quad \sum x_i^2 = 385;$$

$$\sum y_i x_i = 1.258{,}4.$$

a) *Ajuste los coeficientes de la recta de regresión utilizando las fórmulas mínimo cuadráticas.*

b) *¿Qué tanto por ciento de la variabilidad de* y *es explicada por la regresión?*

c) *Obtenga la predicción del valor de* y *para* x = *12.*

Respuesta

a) Ajustaremos la recta de regresión de Y sobre X

$$y - a_{01} = b(x - a_{10})$$

$$a_{10} = \frac{\sum x}{N}$$

$$a_{01} = \frac{\sum y}{N}$$

$$a_{11} = \frac{\sum xy}{N}$$

$$a_{20} = \frac{\sum x^2}{N}$$

$$a_{02} = \frac{\sum y^2}{N}$$

$$m_{11} = a_{11} - a_{10} \cdot a_{01}$$
$$m_{20} = a_{20} - (a_{10})^2$$
$$m_{02} = a_{02} - (a_{01})^2$$

$$b = \frac{m_{11}}{m_{20}}$$

Observamos que falta el dato de la frecuencia total, por lo tanto hagamos que N sea igual a *10*.

$$a_{10} = \frac{55}{10} = 5,5.$$

$$a_{01} = \frac{186,95}{10} = 18,695.$$

$$a_{11} = \frac{1.258,4}{10} = 125,84.$$

$$a_{02} = \frac{4.144,07}{10} = 414,407.$$

$$a_{20} = \frac{385}{10} = 38,5.$$

$$m_{11} = 125,84 - (5,5 \cdot 18,695) = 23,0175.$$
$$m_{20} = 38,5 - (5,5)^2 = 8,25.$$
$$m_{02} = 414,407 - (18,695)^2 = 64,904.$$

$$b = \frac{23,0175}{8,25} = 2,79.$$

$$y - 18,695 = 2,79(x - 5,5) \quad \Rightarrow \quad y = 2,79x + 3,35.$$

b) El coeficiente de determinación es:

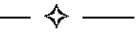
$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 1 - \frac{m_{02} - \dfrac{m_{11}^2}{m_{20}}}{m_{02}} = \frac{m_{02} - m_{02} + \dfrac{m_{11}^2}{m_{20}}}{m_{02}} = \frac{m_{11}^2}{m_{02} \cdot m_{20}} =$$

$$= \frac{(23{,}0175)^2}{64{,}904 \cdot 8{,}25} = 0{,}9894$$

El 98,94% de varianza es explicada por la regresión.

c) Si $x = 12$

$$y = 2{,}79 \cdot 12 + 3{,}35 = 36{,}83.$$



Ejercicio 5.7. *Una empresa dedicada al catering presenta los siguientes datos referidos a los valores mensuales de los costes totales en miles de euros y al total de bandejas producidas en miles de unidades durante los últimos doce meses:*

Suma del total de bandejas en miles	512
Suma de los costes en miles de Euros	2.275
Suma del cuadrado de la variable explicativa	24.290
Suma del producto entre la variable explicativa y la variable no explicada	106.941
Suma de la variación total no explicada	343,1729
Coeficiente de determinación	0,9989

A partir de dichos datos:

a) *Calcule el valor del coste en miles de euros cuando la cantidad de bandejas producida es de 25.000.*

b) *Comente la naturaleza de la relación, según señala la ecuación de regresión.*

Respuesta

A partir de los datos anteriores y sabiendo que la variable endógena es y, coste de producción, la cual viene explicada por el número de bandejas producido, x, tenemos que:

a)

$$b = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{(12 \cdot 106.941) - (512 \cdot 2.275)}{(12 \cdot 24.290) - (512)^2} = 4{,}039$$

$$a = \bar{y} - b\bar{x} = \frac{2.275}{12} - 4{,}039 \cdot \frac{512}{12} = 17{,}2527$$

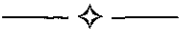
Luego la relación entre los costes de producción y las bandejas producidas es:

$$y = 17{,}2527 + 4{,}039 \cdot x$$

Por tanto, para $x = 25$ se obtiene:

$$17{,}2527 + 4{,}039 \cdot 25 = 118{,}2277$$

b) Se verifica una relación directa, es decir, a medida que aumenta la producción, los costes aumentan. En nuestro caso, el valor 4,039 para *b* significa que la cantidad de bandejas van a variar, según la estimación en 4,039 y en el mismo sentido que los costes.



Ejercicio 5.8. *Supongamos que se posee la siguiente información sobre el coste promedio anual por alumno universitario (y) en miles de euros y el número de alumnos (X) correspondiente a 20 universidades.*

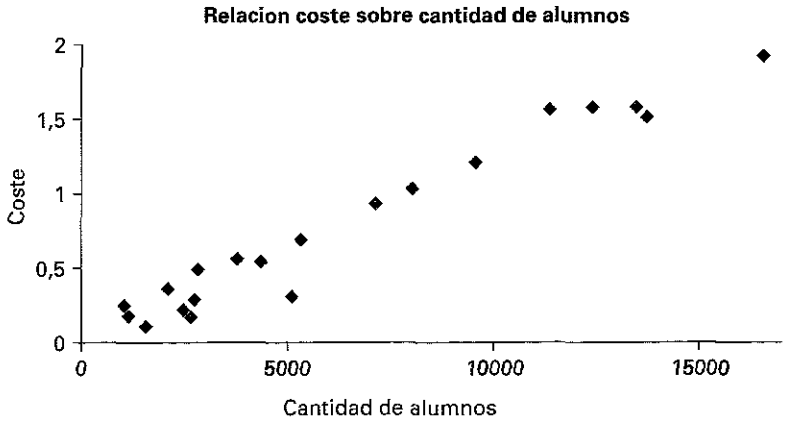
Universidad	Costes (*y*)	Cantidad de estudiantes (*x*)	Universidad	Costes (*y*)	Cantidad de estudiantes (*x*)
1	0,11	1.564	11	1,21	9.578
2	0,56	3.790	12	1,58	13.489
3	1,56	11.383	13	1,92	16.545
4	0,69	5.340	14	0,18	1.149
5	1,03	8.028	15	0,25	4.045
6	0,49	2.841	16	0,31	5.105
7	1,51	13.744	17	0,36	2.102
8	1,58	12.421	18	0,17	2.660
9	0,54	4.348	19	0,29	2.754
10	0,93	7.128	20	0,22	2.475

a) *Realice un gráfico de dispersión de los datos.*

b) *Estime la recta de regresión utilizando los datos del ejemplo referido a una muestra de 20 universidades.*

Respuesta

a)



b) En cada universidad se han registrado los valores de dos variables:

y = coste promedio mensual por alumno.
x = cantidad de alumnos inscriptos.

Vamos a construir una tabla apropiada para facilitar los cálculos.

y_i	x_i	y_i^2	x_i^2	$x_i \cdot y_i$
0,11	1.564	0,0121	2.446.096	172,04
0,56	3.790	0,3136	14.364.100	2.122,40
1,56	11.383	2,4336	129.572.689	17.757,48
0,69	5.340	0,4761	28.515.600	3.684,60
1,03	8.028	1,0609	64.448.784	8.268,84
0,49	2.841	0,2401	8.071.281	1.392,09
1,51	13.744	2,2801	188.897.536	20.753,44
1,58	12.421	2,4964	154.281.241	19.625,18
0,54	4.348	0,2916	18.905.104	2.347,92
0,93	7.128	0,8649	50.808.384	6.629,04
1,21	9.578	1,4641	91.738.084	11.589,38
1,58	13.489	2,4964	181.953.121	21.312,62
1,92	16.545	3,6864	273.737.025	31.766,40

(*Continúa*)

(*Continuación*)

y_i	x_i	y_i^2	x_i^2	$x_i \cdot y_i$
0,18	1.149	0,0324	1.320.201	206,82
0,25	4.045	0,0625	16.362.025	1.011,25
0,31	5.105	0,0961	26.061.025	1.582,55
0,36	2.102	0,1296	4.418.404	756,72
0,17	2.660	0,0289	7.075.600	452,20
0,29	2.754	0,0841	7.584.516	798,66
0,22	2.475	0,0484	6.125.625	544,50
15,49	130.489	18,5983	1.276.686.441	152.774,13

$\bar{y} = 0{,}7745.$

$\bar{x} = 6.524{,}45.$

$$S_y^2 = \frac{18{,}5983}{20} - (0{,}7745)^2 = 0{,}9299 - 0{,}5998 = 0{,}3301.$$

$$S_x^2 = \frac{1.276.686.441}{20} - (6.524{,}45)^2 =$$

$$= 63.834.322{,}05 - 4.268.447{,}80 = 21.265.874{,}25$$

$S_y = 0{,}5745.$

$S_x = 4.611{,}49.$

Calculamos la varianza entre x e y:

$$\operatorname{cov}(x, y) = \frac{\sum x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

$$\operatorname{cov}(x, y) = \frac{152.774{,}13}{20} - 0{,}7745 \cdot 6.524{,}45 =$$

$$= 7.638{,}7065 - 5.053{,}1865 = 2.585{,}52.$$

Con lo que ya podemos calcular el valor de b:

$$b = \frac{2.585{,}52}{2.126.587{,}25} = 0{,}0001216.$$

A continuación calculamos el valor de a:

$a = 0{,}7745 - 0{,}0001216 \cdot 6524{,}45 = 0{,}7745 - 0{,}793 = -0{,}0185.$

La recta de regresión estimada es, entonces:

$$\hat{y} = -0{,}0185 + 0{,}0001216x$$

Ejercicio 5.9. *Supongamos que un administrador de un complejo turístico en un balneario posee la siguiente información referente a 6 veranos consecutivos:*

x_i	y_i
25	6,5
27	7,0
30	9,0
28	8,5
31	9,0
30	8,2

Donde x *es la temperatura media durante el verano e* y *es el rendimiento del complejo en cientos de miles de euros.*

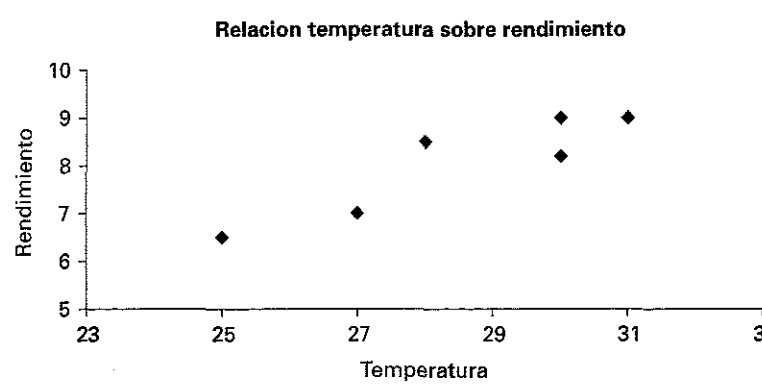
A partir de los datos:

a) *Grafique los datos utilizando un diagrama de dispersión.*

b) *Estime la recta de regresión poblacional.*

Respuesta

a)



b) $\bar{x} = \frac{1}{n}\sum x_i = 28{,}5.$

$\bar{y} = \frac{1}{n}\sum y_i = 8{,}03.$

$$\left.\begin{array}{l} \sum x_i^2 = 4.899 \\ \dfrac{\left(\sum x_i\right)^2}{n} = \dfrac{(171)^2}{6} \end{array}\right\} \Rightarrow S_{XX} = 25{,}5.$$

$$\left.\begin{array}{l} \sum x_i\, y_i = 1.384{,}5 \\ \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{N} = \dfrac{171 \cdot 48{,}2}{6} \end{array}\right\} \Rightarrow S_{XY} = 10{,}8.$$

$$\left.\begin{array}{l} \hat{\beta} = \dfrac{S_{XY}}{S_{XX}} = 0{,}42 \\ \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = -4{,}04 \end{array}\right\} \Rightarrow Y = -4{,}04 + 0{,}42x.$$

Ejercicio 5.10. *Dada la siguiente tabla:*

X	5	10	15	20
Y	12,4	16,5	20,2	18,7
	15,2	19,1		21,3

Se pide:

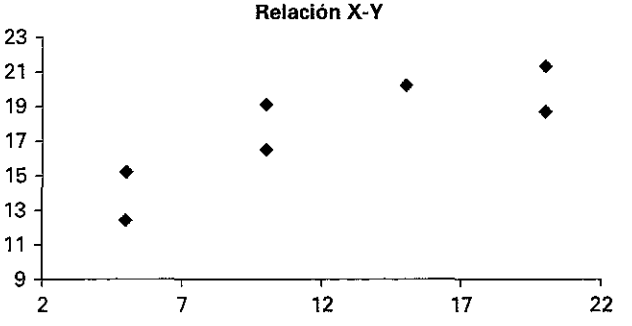
a) *Represente estos datos en un diagrama de dispersión y observe si es razonable suponer que existe una relación lineal entre* X *e* Y.

b) *Plantee el modelo lineal y estime los parámetros por el método de mínimos cuadrados.*

c) *Utilice la recta estimada para predecir el valor de* Y *sabiendo que* x = *12.*

Respuesta

a)



b)

$$\bar{x} = \frac{1}{n_x}\sum x_i = 12{,}143.$$

$$\bar{y} = \frac{1}{n_y}\sum y_i = 17{,}6285.$$

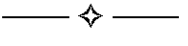
$$\left.\begin{array}{l} \sum n_i x_i^2 = 1.275 \\ \dfrac{\left(\sum n_i x_i\right)^2}{n} = \dfrac{(85)^2}{7} \end{array}\right\} \Rightarrow S_{XX} = 242{,}857.$$

$$\left.\begin{array}{l} \sum x_i y_i = 1.597 \\ \dfrac{\left(\sum n_i x_i\right)\left(\sum y_i\right)}{n} = \dfrac{85 \cdot 123{,}4}{7} \end{array}\right\} \Rightarrow S_{XY} = 98{,}571.$$

$$\left.\begin{array}{l} b = \dfrac{S_{XY}}{S_{XX}} = 0{,}4058 \\ a = \bar{y} - b\bar{x} = 12{,}7 \end{array}\right\} \Rightarrow Y = 12{,}7 + 0{,}406x.$$

c)

$$\hat{Y}_k = 12{,}7 + 0{,}406 \cdot 12 = 17{,}5694$$



Ejercicio 5.11. *Los datos siguientes corresponden a la frecuencia de entradas de personas en un espectáculo tomadas en intervalos de tiempo t de 5 minutos.*

t (minutos)	5	10	15	20	25
Y (personas)	72	65	70	62	64

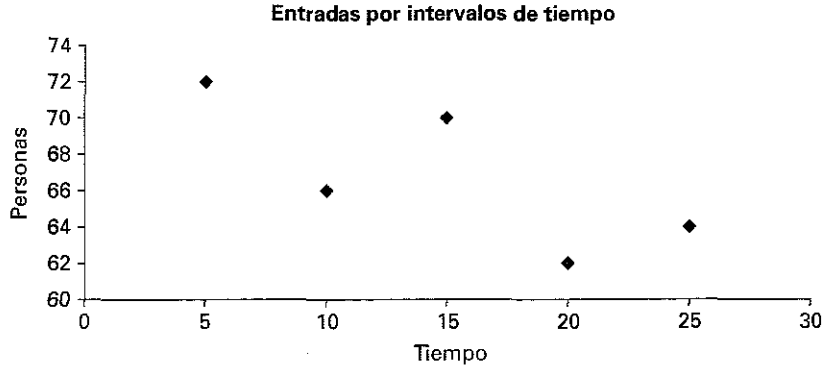
Se pide:

a) *El diagrama de dispersión.*

b) *El modelo lineal y los parámetros estimados por mínimos cuadrados.*

Respuesta

a)



b)

$$\bar{x} = \frac{1}{n}\sum x_i = 15$$

$$\bar{y} = \frac{1}{n}\sum y_i = 17{,}6285$$

$$\left.\begin{array}{l} \sum x_i^2 = 1275 \\ \dfrac{\left(\sum x_i\right)^2}{n} = \dfrac{(85)^2}{5} \end{array}\right\} \Rightarrow S_{XX} = 242{,}857$$

$$\left.\begin{array}{l} \sum x_i y_i = 4910 \\ \dfrac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = \dfrac{75 \times 334}{5} \end{array}\right\} \Rightarrow S_{XY} = -100$$

$$\left.\begin{array}{l} \hat{\beta} = \dfrac{S_{XY}}{S_{XX}} = -0{,}4 \\ \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} = 72{,}8 \end{array}\right\} \Rightarrow Y = 72{,}8 - 0{,}4\,x$$

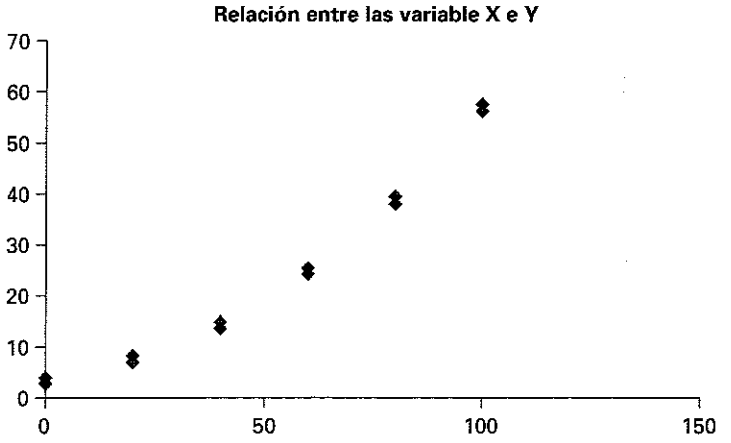
Ejercicio 5.12. *Consideremos la siguiente tabla:*

x	0	20	40	60	80	100
y	2,8	7,0	14,8	25,4	38,0	56,1
	3,9	8,2	13,6	24,2	39,5	57,5

Teniendo en cuenta el diagrama de dispersión, verifique que esta curva no sigue un modelo lineal. Busque el modelo que la ajuste y estime los parámetros.

Respuesta

Realicemos el diagrama de dispersión de los datos:



Observamos que la curva a ajustar podría ser una parábola de la forma:

$$y = \alpha + \beta\, x^2$$

Para aplicar las fórmulas correspondientes, debemos realizar el cambio de variables $u_i = x_i^2$ tal que:

$u_i = x_i^2$	0	400	1.600	3.600	6.400	10.000
y	2,8	7,0	14,8	25,4	38,0	56,1
	3,9	8,2	13,6	24,2	39,5	57,5

$$\overline{u} = \frac{1}{n_u}\sum u_i = 3.666,67.$$

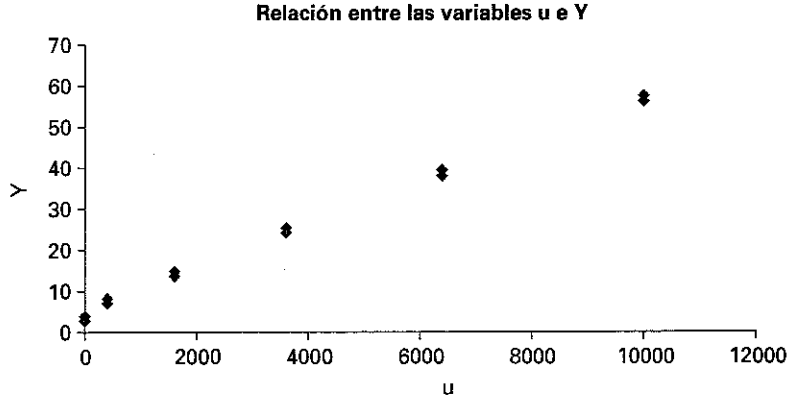
$$\overline{y} = \frac{1}{n_y}\sum y_i = 24,25.$$

$$\left.\begin{array}{l} \sum n_i u_i^2 = 313.280.000 \\ \dfrac{\left(\sum n_i u_i\right)^2}{n} = \dfrac{(44.000)^2}{12} \end{array}\right\} \Rightarrow S_{UU} = 151.946.666,7.$$

$$\left.\begin{array}{l} \sum u_i y_i = 1.862.080 \\ \dfrac{\left(\sum n_i u_i\right)\left(\sum y_i\right)}{n} = \dfrac{44.000 \cdot 291}{12} \end{array}\right\} \Rightarrow S_{UY} = 795.080.$$

$$\left.\begin{array}{l} \hat{\beta} = \dfrac{S_{UY}}{S_{UU}} = 0,005233 \\ \hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x} = 5,06 \end{array}\right\} \Rightarrow Y = 5,06 + 0,005u = 5,06 + 0,005x^2.$$

Veamos ahora el diagrama de dispersión para los datos modificados:



Ejercicio 5.13. *Dada la siguiente tabla se pide:*

X	6	18	30	42	54
Y	30,0	8,9	4,1	1,8	0,8
	28,6	8,0	4,6	2,6	0,6
	28,5	10,8	4,7	2,2	1,0

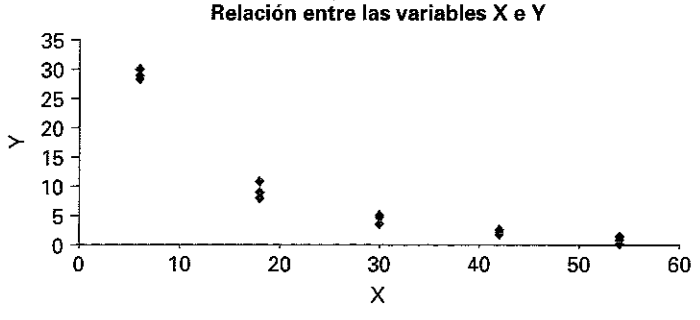
a) *Realice un diagrama de dispersión y compruebe que el modelo lineal no es correcto.*

b) *Ajuste al modelo ln* $Y = \alpha + \beta x$. *Realice un diagrama de dispersión con los datos modificados.*

c) *Utilizando la recta hallada en el punto anterior, calcule el valor de* Y *para* $x = 35$.

Respuesta

a)



b)

x	6	18	30	42	54
	3,4	2,19	1,41	0,59	–0,22
$\ln y = V$	3,35	2,08	1,53	0,96	–0,51
	3,35	2,38	1,55	0,79	0

$$\overline{x} = \frac{1}{n_x} \sum x_i = 30.$$

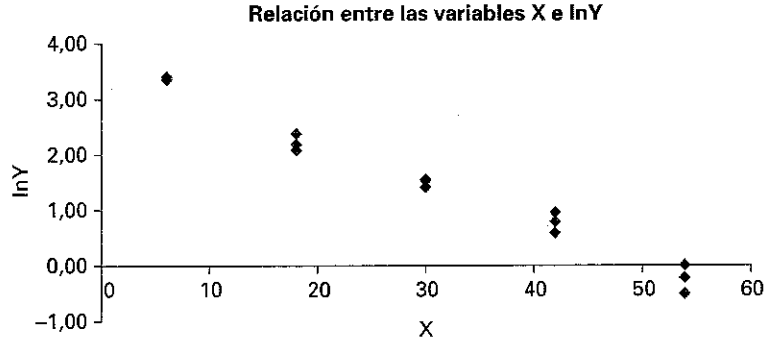
$$\bar{V} = \frac{1}{n_V}\sum V_i = 1{,}52.$$

$$\left.\begin{array}{l} \sum n_i x_i^2 = 17.820 \\ \dfrac{\left(\sum n_i x_i\right)^2}{n} = \dfrac{(450)^2}{15} \end{array}\right\} \Rightarrow S_{XX} = 4.320.$$

$$\left.\begin{array}{l} \sum x_i V_i = 1.862.080 \\ \dfrac{\left(\sum n_i x_i\right)\left(\sum V_i\right)}{n} = \dfrac{450 \cdot 22{,}85}{15} \end{array}\right\} \Rightarrow S_{XV} = 311{,}64.$$

$$\left.\begin{array}{l} \hat{\beta} = \dfrac{S_{XV}}{S_{XX}} = 0{,}072 \\ \hat{\alpha} = \bar{V} - \hat{\beta}\,\bar{x} = 3{,}68 \end{array}\right\} \Rightarrow V = 3{,}68 - 0{,}072x \quad \Rightarrow \quad \ln Y = 3{,}68 - 0{,}072x.$$

Veamos ahora el diagrama de dispersión para los datos modificados:



c) $\ln Y = 3{,}68 - 0{,}072 \cdot 35 = 1{,}16 \quad \Rightarrow \quad Y = e^{1{,}16} = 3{,}19\,.$

Capítulo 6

NÚMEROS ÍNDICES

6.1. INTRODUCCIÓN

Hasta ahora hemos trabajado con variables, que hemos tratado de forma estadística para conocer descriptivamente sus características, pero existen valores referidos a variables económicas (precios, cantidades producidas, costes, ingresos,...) que cambian con el tiempo.

Cuando para una de esta variables, tomamos una serie de valores correspondientes a diferentes momentos de tiempo, tenemos una «serie temporal»; en este capítulo y en el siguiente estudiaremos un conjunto de instrumentos estadísticos para tratar este tipo de valores.

El primer grupo de instrumentos, tratado en este capítulo, es el genéricamente denominado como «números índice», que permiten medir y comparar de forma sencilla los cambios sufridos por la variable a lo largo del tiempo.

El segundo, el tratamiento de «series temporales», lo abordamos en el capítulo siguiente.

Un ***Número Índice*** es una *medida estadística diseñada para poner de manifiesto los cambios en una variable (o grupo de variables relacionadas) con respecto a una determinada característica (generalmente, el tiempo).*

Se utilizan para efectuar comparaciones en diferentes momentos o periodos, respecto de una variable.

Cada índice es un número que indica una variación; el número índice 105 correspondiente a 2010, se interpreta como que «la variable en cuestión creció en un 5% respecto a un período anterior, el período que se ha tomado como referencia».

Los números índices se elaboran tanto con precios (p) como con cantidades (q). El año en que se inicia el cálculo de un número índice se denomina *año base* y se denotan por p_0 y q_0 los precios y las cantidades de dicho año. A los precios y las cantidades de los años sucesivos los anotamos como p_t y q_t. Si trabajamos con diferentes tipos de mercancías, utilizamos los subíndices (i) para referirnos a un tipo de mercancía, de modo que usamos los símbolos p_{it} o q_{it} o para señalar el precio o la cantidad de la mercancía i en el período t. Si hubiese N mercancías el valor total de la cesta de productos durante el periodo t se expresa:

$$\text{Valor total durante el período } t = \sum_{i=1}^{n} p_{it}\, q_{it}$$

Los números índices se clasifican en *ponderados* y *no ponderados*. Los números índices no ponderados son los más sencillos de calcular, pero deben de utilizarse con especial cuidado.

Los números índices ponderados requieren que definamos previamente a su construcción los criterios de ponderación o de peso. Una vez definida una ponderación debe de respetarse, en general, en los sucesivos períodos. En este apartado estudiaremos los índices ponderados que son de aplicación común.

Así, por ejemplo, un número índice ponderado es el Índice de Precios Hoteleros en España; este Índice es una medida estadística de la evolución mensual del conjunto de las principales tarifas de precios que los empresarios aplican a sus clientes.

Para su obtención se utiliza la Encuesta de Ocupación en Alojamientos Turísticos, en la que, mensualmente, se recoge información sobre la ocupación hotelera (viajeros entrados, pernoctaciones, grado de ocupación etc.), su estructura (plazas, personal, etc.) y demás variables de interés, con una amplia desagregación geográfica y por categorías de los establecimientos. En el cuestionario, se les pide, entre otras variables, el ADR (Average Daily Rate) o tarifas promedio diarias aplicadas a distintos tipos de clientes por una habitación doble con baño. Esos precios se desglosan según el tipo de cliente al que se le ha aplicado: Tour operador tradicional, Agencia de viajes tradicional (incluyendo bonos y talones de hoteles), Empresas, Particulares, Grupos, Contratación directa en la Web del hotel y/o de la cadena hotelera, Tour operador on-line, Agencia de viajes on-line y Otros.

Cada tarifa pondera de forma diferente cada mes, en función del uso real que los hoteles han dado a la misma; integradas, conforman el índice general, que tiene una presentación del siguiente tipo[10]:

Índice general nacional y desglose por tarifas

Mes de enero de 2010	Índice	Tasa de variación Interanual
Total categorías	91,1	–6,0
Cinco estrellas de oro	86,0	–7,9
Cuatro estrellas de oro	89,8	–6,7
Tres estrellas de oro	93,3	–5,6
Dos estrellas de oro	97,3	–1,7
Una estrella de oro	97,1	–3,1
Tres y dos estrellas de plata	95,4	–2,2
Una estrella de plata	98,9	–1,9

[10] La metodología de esta Encuesta puede verse con detalle en la Web: http://www.ine.es/daco/daco42/prechote/metoiphp1.htm; la información puede extraerse desde la Web: http://www.ine.es/inebmenu/mnu_hosteleria.htm.

La interpretación es la siguiente: en el mes de enero de 2010, y en relación con el año base, 2008, los precios (todas las categorías) se habían reducido un 8,9 %, los de los hoteles de 5 estrellas de oro un 14 %, etc.

Del mismo modo, en la última columna, se indica el crecimiento o tasa de variación anual con el mismo mes del año anterior (Enero de 2009), en el que se pone de manifiesto, según el INE, una bajada de todas las tarifas, en porcentajes que oscilan entre el 1,7% y el 7,9%.

6.2. PROPIEDADES DE LOS NÚMEROS ÍNDICES

Los números Índice tienen las siguientes propiedades:

a) ***Existencia:*** Todo número índice ha de tener un valor finito distinto de cero.

b) ***Identidad:*** Si se hacen coincidir el período base y el período actual el valor del índice tiene que ser igual a la unidad (o a 100 si se elabora en porcentajes); en la notación habitual: : $I_n^n = 100$.

Un ejemplo de número índice que cumple esta propiedad es el índice simple:

$$I_n^n = \frac{x_n}{x_n} 100 \quad \Rightarrow \quad I_n^n = 100$$

c) ***Inversión:*** El valor del índice ha de ser invertible al intercambiar los períodos entre sí. Es decir:

$$I_t^0 = \frac{1}{I_0^t}, \quad \text{o bien} \quad \frac{I_t^n}{100} = \frac{1}{\frac{I_n^t}{100}}, \quad \text{o bien} \quad \frac{I_t^n}{100} \cdot \frac{I_n^t}{100} = 1$$

Es decir, que el índice del año *o* calculado con la base del año *t*, ha de ser igual al inverso del índice del año *t* calculado en base del año *0*.

d) ***Proporcionalidad:*** Si en el período actual todas las magnitudes experimentan una variación proporcional, el número índice tiene que experimentar también dicha variación.

e) ***Homogeneidad:*** Un número índice no puede estar afectado por los cambios que se realicen en las unidades de medida.

f) ***Propiedad cíclica o circular:*** Determina una relación de igualdad entre un número índice construido con periodos base sucesivos y otro elaborado a partir del periodo inicial como base y el último periodo como periodo de estudio. Si se toma la relación que surge de tres periodos.

$$\frac{I_a^b}{100} \cdot \frac{I_b^c}{100} \cdot \frac{I_c^a}{100} = 1$$

g) ***Propiedad cíclica o circular modificada:*** Se desprende de las propiedades cíclica y de inversión temporal. Si se toma la relación que surge de tres periodos, se obtiene:

$$\frac{I_a^b}{100}\cdot\frac{I_b^c}{100}=\frac{I_a^c}{100}$$

6.3. NÚMEROS ÍNDICES SIMPLES Y COMPLEJOS

Los números índices simples son los más elementales; se elaboran a partir de la razón de precios (precios relativos) o cantidades (cantidades relativas) respecto a su valor en el período base.

$$I_{it}=\frac{X_{it}}{X_{i0}}\cdot 100$$

En el siguiente período el índice simple sería:

$$I_{i(t+1)}=\frac{X_{i(t+1)}}{X_{i0}}\cdot 100$$

Al comparar los números índice I_{it} e $I_{i(t+1)}$ se obtiene el incremento del precio de dicho producto en cuestión.

Ejemplo 6.1. Con los siguientes datos, elaborar un Índice de los Precios Medios de Alojamiento aplicados por un determinado hotel a las habitaciones dobles.

Año	Precio Medio del Alojamiento en habitación doble
2009	50 €
2010	51 €

Tomando como base la media de precios aplicados en el año 2009 (50 euros por habitación doble) tendríamos, expresado en porcentajes:

$$I_t=\frac{P_t}{P_0}\cdot 100=\frac{P_{2010}}{P_{2009}}\cdot 100=\frac{51}{50}\cdot 100=102$$

Indicativo de que los precios habrían aumentado un 2%, o lo que es lo mismo: número índice = 102%

Obsérvese que la operación realizada equivale a una simple regla de tres; sí 50 € equivalen a un índice 100, 51 € equivaldrán a x, dónde $x = \frac{51 \cdot 100}{50} = 102$.

Ejemplo 6.2. Con los siguientes datos:

Años	Precio
2005	150
2006	158
2007	168
2008	175
2009	183
2010	196

Elaborar un Índice de los precios con base año 2005 = 100 y obtener el incremento porcentual anual.

Años	Precio	Índice 2005 = 100	Incremento anual
2005	150	100,00%	
2006	158	105,33%	5,33%
2007	168	112,00%	12,00%
2008	175	116,67%	16,67%
2009	183	122,00%	22,00%
2010	196	130,67%	30,67%

Cuando lo que se desea es estudiar la variación de una serie de variables, sintetizándola en un sólo índice, se habrán de usar índices complejos. Los dos tipos principales: son los índices complejos con ponderación y sin ponderación.

6.3.1. Números Índices complejos de precios sin ponderación

Los índices simples pueden agregarse de diferentes formas; a dichas agregaciones se les conoce como índices complejos. Sí en la agregación no se utilizan elementos especiales de ponderación, tendremos Números Índices complejos sin ponderación.

Ejemplo 6.3. Disponemos de las notas medias de los alumnos aprobados en diversos exámenes de Estadística en los centros asociados de la UNED.

	Madrid	Córdoba	Palencia	Sevilla	Mérida
Junio 2009	5,3	5,3	6,5	6	7
Septiembre 2009	5,6	5,2	5,2	6,3	6,8
Junio 2010	7,3	5,3	5,3	6,3	5,2
Septiembre 2010	6,1	5,5	5,5	6,4	4,1
Junio 2011	6,3	5,7	5,7	5,7	3,6
Septiembre 2011	5,8	5,2	5,2	6,9	8

Podríamos obtener para cada centro los índices de evolución del período, con base Junio de *2009* igual a *100*, obteniendo:

	Madrid	Córdoba	Palencia	Sevilla	Mérida
Junio 2009	100,0%	100,0%	100,0%	100,0%	100,0%
Septiembre 2009	105,7%	98,1%	80,0%	105,0%	97,1%
Junio 2010	137,7%	100,0%	81,5%	105,0%	74,3%
Septiembre 2010	115,1%	103,8%	84,6%	106,7%	58,6%
Junio 2011	118,9%	107,5%	87,7%	95,0%	51,4%
Septiembre 2011	109,4%	98,1%	80,0%	115,0%	114,3%

Sin embargo, estos índices tan sólo nos pueden indicar la variación de la nota media de cada uno de lo centros por separado; sí deseáramos sintetizar en un sólo índice esta información tendríamos que obtener un índice complejo, en el que consideráse-mos, por ejemplo, la media de evolución de los índices obtenidos en los diversos centros; tendríamos así:

	Nota promedio	Índice medio de evolución
Junio 2009	6,02	100,0%
Septiembre 2009	5,82	96,7%
Junio 2010	5,88	97,7%
Septiembre 2010	5,52	91,7%
Junio 2011	5,4	89,7%
Septiembre 2011	6,22	103,3%

Si suponemos que tenemos N diferentes productos, podemos obtener los siguientes índices complejos:

— *Índice media aritmética de índices simples*:

$$I = \frac{I_1 + I_2 + I_3 + \ldots + I_N}{N} = \frac{\sum_{i=1}^{N} I_i}{N}$$

Este Índice también se denomina Índice de Sauerbeck.

— *Índice media geométrica de índices simples* :

$$I = \sqrt[N]{I_1 \cdot I_2 \cdot I_3 \cdot \ldots \cdot I_N} = \sqrt[N]{\prod_{i=1}^{N} I_i}$$

— *Índice media armónica de índices simples*:

$$I = \frac{N}{\frac{1}{I_1} + \frac{1}{I_2} + \ldots + \frac{1}{I_n}} = \frac{N}{\sum_{i=1}^{n} \frac{1}{I_i}}$$

— *Índice media agregativa de índices simples* :

$$I = \frac{x_{1t} + x_{2t} + x_{3t} + \ldots + x_{Nt}}{x_{10} + x_{20} + x_{30} + \ldots + x_{N0}} \cdot 100 = \frac{\sum_{i=1}^{N} x_{it}}{\sum_{i=1}^{N} x_{i0}} \cdot 100$$

Este Índice se denomina Índice de Bradstreet-Dutot.

6.3.2. Números índices de precios complejos ponderados

Con el objeto de reducir las desventajas que representa usar índices que otorgan igual importancia a cada variable, se generan índices ponderados. De esta manera, se asigna un peso o ponderación w_i que indica la importancia relativa de la variable en el índice total.

Ejemplo 6.4. Supongamos, que durante el año 2005 un hotel aplicó un precio medio de 52 euros teniendo un 97% de ocupación y durante 2004 el precio medio fue de 50 euros y la ocupación del 99%; podemos construir un índice, precio-cantidad, del tipo:

$$I_{it} = \frac{p_{i2005} \cdot q_{i2005}}{p_{i2004} \cdot q_{i2004}} \cdot 100 = \frac{52 \cdot 97}{50 \cdot 99} \cdot 100 = 101{,}9$$

Indicativo de que la evolución de los ingresos ha tenido un incremento del 1,9% en el período; estamos hablando en este caso de un *índice complejo* ponderado por cantidades.

La ponderación de los índices complejos se hace imprescindible cuando se dispone de diversos índices simples con diferente importancia cada uno; por ejemplo, para establecer un índice general de precios o de ingresos de una cadena hotelera, parece apropiado considerar por separado las diferentes grupos tarifarios, lo que nos permitiría hacer un seguimiento de la evolución de cada uno de los precios aplicados a cada grupo; sin embargo, para integrar el índice general, hay que tener en cuenta que cada grupo tiene una importancia diferente en la estructura de la ocupación y de los ingresos de la cadena hotelera y en consecuencia a cada índice de grupo hay que darle una importancia diferente en la elaboración de un índice general; esta importancia viene determinada por un ponderador y se denomina ponderación.

Para elaborar un ponderador hay que considerar una magnitud que sea comparable entre los diferentes tipos de productos utilizados. En el ejemplo del hotel se podría considerar para ponderar el número de habitaciones ocupadas por cada grupo tarifario en un período concreto o los ingresos totales obtenidos en la contratación de cada uno de los grupos. Para evitar los problemas de homogeneización que tienen las magnitudes físicas se suele acudir a las valoraciones monetarias para obtener los ponderadores.

Ejemplo 6.5. Elaborar un ponderador para los grupos tarifarios aplicados por una cadena hotelera teniendo en cuenta que los ingresos que le aporta cada grupo tarifario, en miles de euros/año, son los siguientes:

Grupos tarifarios	Ingresos
Tour-operadores	300
Clientes individuales	125
Empresas concertadas	45
Total	470

Haciendo 470 = 100, tendríamos los siguientes ponderadores:

Grupos tarifarios	Ponderador
Tour-operadores	63,8%
Clientes individuales	26,6%
Empresas concertadas	9,6%
Total	100%

Una vez obtenidos los ponderadores se calculan los índices complejos ponderados utilizando las siguientes fórmulas alternativas:

1. *Índice media aritmética ponderada de índices simples*:

$$I = \frac{I_1 w_1 + I_2 w_2 + I_3 w_3 + \ldots + I_n w_N}{w_1 + w_2 + w_3 + \ldots + w_n} = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i}$$

2. *Índice media geométrica ponderada de índices simples*:

$$I = \sqrt[n]{I_1^{w_1} \cdot I_2^{w_2} \cdot I_3^{w_3} \cdot \ldots \cdot I_N^{w_N}} = \sqrt[n]{\prod_{i=1}^{N} I_N^{wN}}$$

3. *Índice media armónica ponderada de índices simples*:

$$I = \frac{w_1 + w_2 + \ldots + w_n}{\frac{1}{I_1} w_1 + \frac{1}{I_2} w_2 + \ldots + \frac{1}{I_n} w_n} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{1}{I_i} w_i}$$

4. *Índice media agregativa ponderada de índices simples*:

$$I = \frac{x_{1t} w_1 + x_{2t} w_2 + x_{3t} w_3 + \ldots + x_{Nt} w_N}{x_{10} w_1 + x_{20} w_2 + x_{30} w_3 + \ldots + x_{N0} w_N} \cdot 100 = \frac{\sum_{i=1}^{N} x_{it} w_i}{\sum_{i=1}^{N} x_{i0} w_i} \cdot 100$$

Al definir el ponderador en tanto por uno, el procedimiento de cálculo de los índices ponderados se simplifica notablemente.

Ejemplo 6.6. Con los siguientes datos correspondientes a la nota media de una asignatura en distintos exámenes y centros de estudio:

	Madrid	Tarrasa	Cádiz	Córdoba	Las Palmas
Junio 2009	5,1	5,3	6,5	6	7
Septiembre 2009	5,2	5,2	6,3	5,3	6,3
Junio 2010	5,3	5,3	5,5	7,3	6,3
Septiembre 2010	5,4	5,5	5,4	5,3	6,4
Junio 2011	5,8	5,7	5,8	6,3	5,7
Septiembre 2011	5,9	5,2	6,3	5,3	6,9

Obtener, mediante una media *aritmética ponderada de índices simples, la evolución del índice de notas de la asignatura,* teniendo en cuenta que el número de alumnos de cada centro se distribuye en la siguiente forma: Madrid: 50%, Tarrasa 20%, Cádiz, Córdoba y Las Palmas 10%.

Tenemos:

	Madrid	Tarrasa	Cádiz	Córdoba	Las Palmas
Junio 2009	100,00	100,00	100,00	100,00	100,00
Septiembre 2009	101,96	98,11	96,92	88,33	90,00
Junio 2010	103,92	100,00	84,62	121,67	90,00
Septiembre 2010	105,88	103,77	83,08	88,33	91,43
Junio 2011	113,73	107,55	89,23	105,00	81,43
Septiembre 2011	115,69	98,11	96,92	88,33	98,57

Aplicando la formulación

$$I = \frac{I_1 w_1 + I_2 w_2 + I_3 w_3 + \ldots + I_n w_n}{w_1 + w_2 + w_3 + \ldots + w_n} = \frac{\sum_{i=1}^{n} I_i w_i}{\sum_{i=1}^{n} w_i}$$

en el segundo periodo, septiembre de 2009, se obtiene:

$$I = \frac{101{,}96 \cdot 0{,}5 + 98{,}11 \cdot 0{,}2 + 96{,}92 \cdot 0{,}1 + 88{,}33 \cdot 0{,}1}{0{,}5 + 0{,}2 + 0{,}1 + 0{,}1} = 98{,}1$$

Aplicando la misma expresión para el resto de los períodos, se obtendrá la segunda columna del siguiente cuadro:

	Índice medio ponderado	Índice medio sin ponderar
Junio 2009	100,0	100,00
Septiembre 2009	98,1	95,07
Junio 2010	101,6	100,04
Septiembre 2010	100,0	94,50
Junio 2011	105,9	99,39
Septiembre 2011	105,8	99,53

La mejor evolución de Madrid, que pesa o pondera un 50% en el colectivo de alumnos, hace que el índice ponderado tenga una evolución más positiva que el índice sin ponderar. Digamos, finalmente, que la decisión de otorgar un 50% de peso a Madrid, aunque podría ser arbitraria, debe estar motivada por algún criterio objetivo de medición de la importancia relativa de cada subíndice; en este caso, por ejemplo, por el número total de alumnos que tiene el centro de Madrid.

6.4. ÍNDICES DE PRECIOS COMPUESTOS PONDERADOS

Los Índices de precios más comunes se elaboran utilizando índices complejos ponderados; el más empleado es el denominado Índice de Laspeyres, pero existen otros, que también veremos a continuación:

1. Índice de Laspeyres

Se trata de un tipo particular de Índice compuesto o agregativo ponderado. Para obtener un índice de precios de Laspeyres, se debe valorar el consumo del año base a precios del año en estudio (en el numerador) y a precios del año base (en el denominador).

El Índice de Laspeyres es, pues, una media aritmética ponderada de índices simples, cuyo criterio de ponderación es: $w_i = p_{i0} \cdot q_{i0}$.

La fórmula que define el Índice de Laspeyres, expresada en porcentaje, es la siguiente:

$$L_p = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} p_{it} q_{i0}}{\sum_{i=1}^{N} p_{i0} q_{i0}} \cdot 100 \quad \text{siendo} \quad I_i = \frac{p_{it}}{p_{i0}}$$

Otra notación alternativa es:

$$IPL_0^n = \frac{\sum p_n \cdot q_0}{\sum p_0 \cdot q_0} \cdot 100$$

Se suele utilizar este Índice a la hora de elaborar los índices de precios por cuestiones prácticas ya que únicamente requiere investigar en el año base el valor de los ponderadores, que es la parte mas costosa de la elaboración del índice (téngase en cuenta que en el IPC se realiza una encuesta de presupuestos familiares en los años base que requiere una muestra de aproximadamente 20.000 hogares). Una vez determinados los ponderadores el índice de Laspeyres únicamente requiere que se investigue en los sucesivos períodos la evolución de los precios.

2. Índice de Paasche

También es una media aritmética ponderada de los Índices simples, pero utilizando como coeficiente ponderador $w_i = p_{i0} \cdot q_{it}$; por tanto su definición, expresada en porcentaje, queda como:

$$P_p = \frac{\sum_{i=1}^{N} I_i w_i}{\sum_{i=1}^{N} w_i} = \frac{\sum_{i=1}^{N} p_{it} q_{it}}{\sum_{i=1}^{N} p_{i0} q_{it}} \cdot 100$$

Como notación alternativa otros autores utilizan la siguiente:

$$IPP_0^n = \frac{\sum p_n \cdot q_n}{\sum p_0 \cdot q_n} \cdot 100$$

La diferencia entre el Índice Paasche y el Índice Laspeyres es que exige calcular las ponderaciones para cada periodo corriente t, haciendo su cálculo estadístico más laborioso. Además presenta el inconveniente de que sólo permite comparar la evolución del precio de cada año con el del año base, dado que las ponderaciones varían de período en período. Ambas razones han determinado que este índice sea más inusual que el anterior.

3. Índice de Fisher

El Índice de Fisher es la media geométrica de los Índices de Laspeyres y Paasche, es decir:

$$F_p = \sqrt{L_p \cdot P_p}$$

Finalmente, aunque menos utilizados en la práctica, cabe destacar los siguientes Índices de precios y cantidades:

4. **Índice de Drovisch-Bowley**

$$I_{DB} = \frac{I_p + P_p}{2}$$

5. **Índice de Edgeworth-Marshall**

$$I^n_{EM0} = \frac{\sum p_n \cdot (q_0 + q_t)}{\sum p_0 \cdot (q_0 + q_t)} \cdot 100$$

6. **Índice de Walch**

$$IPW^n_0 = \frac{\sum p_n \cdot (q_0 \cdot q_t)}{\sum p_0 \cdot (q_0 \cdot q_t)} \cdot 100$$

Ejemplo 6.7. Elaborar los Índices de Laspeyres, Paasche y Fisher para la siguiente información sobre precios y ocupación de una cadena hotelera.

	2008		2009		2010	
	Precio	**Ocupación**	**Precio**	**Ocupación**	**Precio**	**Ocupación**
Tarifa A	39	3.000	37	3.000	40	2.400
Tarifa B	40	2.500	45	2.500	47	2.500
Tarifa C	45	2.000	50	1.700	58	2.000
Tarifa D	50	1.500	55	1.200	60	1.750

1. *Índice de precios de Laspeyres*

a) Año 2008.

$p_n \cdot q_0$	$p_0 \cdot q_0$
117.000	117.000
100.000	100.000
90.000	90.000
75.000	75.000
382.000	382.000

En este caso $n = 2008$; multiplicamos en la primera columna y en la segunda los precios de 2008 por las ocupaciones de 2008:

$$IPL_0^n = \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot 100 = \frac{382.000}{382.000} \cdot 100 = 100$$

b) Año 2009 con base en 2008.

$p_n \cdot q_0$	$p_0 \cdot q_0$
111.000	117.000
112.500	100.000
100.000	90.000
82.500	75.000
406.000	382.000

En este caso $n = 2009$; multiplicamos en la primera columna los precios del año 2009 por las ocupaciones 2008 y en la segunda los precios de 2008 por las ocupaciones de 2008:

$$IPL_0^n = \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot 100 = \frac{406.000}{382.000} \cdot 100 = 100$$

c) Año 2010 con base en 2008.

$p_n \cdot q_0$	$p_0 \cdot q_0$
120.000	117.000
117.500	100.000
116.000	90.000
90.000	75.000
443.500	382.000

En este caso $n = 2010$; multiplicamos en la primera columna los precios del año 2010 por las ocupaciones 2008 y en la segunda los precios de 2008 por las ocupaciones de 2008:

$$IPL_0^n = \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot 100 = \frac{443.500}{382.000} \cdot 100 = 100$$

2. *Índice de precios de Paasche*

a) Año 2008.

$p_n \cdot q_n$	$p_0 \cdot q_n$
117.000	117.000
100.000	100.000
90.000	90.000
75.000	75.000
382.000	382.000

En este caso 0 = 2008 y n = 2008; multiplicamos en la primera y en la segunda columna los precios de 2008 por las ocupaciones de 2008:

$$IPP_0^n = \frac{\sum p_n q_n}{\sum p_0 q_n} \cdot 100 = \frac{382.000}{382.000} \cdot 100 = 100$$

b) Año 2009 con base en 2008.

$p_n \cdot q_n$	$p_0 \cdot q_n$
111.000	117.000
112.500	100.000
85.000	76.500
66.000	60.000
374.500	353.500

En este caso 0 = 2008 y n = 2009; multiplicamos en la primera columna los precios y las ocupaciones del año 2009 y en la segunda los precios de 2008 por las ocupaciones de 2009:

$$IPP_0^n = \frac{\sum p_n q_n}{\sum p_0 q_n} \cdot 100 = \frac{374.500}{353.500} \cdot 100 = 105{,}94$$

c) Año 2010 con base en 2008.

$p_n \cdot q_n$	$p_0 \cdot q_n$
96.000	93.600
117.500	100.000
116.000	90.000
105.000	87.500
434.500	371.100

En este caso 0 = 2008 y n = 2010; multiplicamos en la primera columna los precios del año 2010 por las ocupaciones de 2010 y en la segunda los precios de 2008 por las ocupaciones de 2010:

$$IPP_0^n = \frac{\sum p_n q_n}{\sum p_0 q_n} \cdot 100 = \frac{434.500}{371.100} \cdot 100 = 117{,}08$$

3. *Índice de Fisher con base en 2008.*

Año	$F_p = \sqrt{L_p \cdot P_p}$	Valor del Índice
2008	$F_p = \sqrt{100 \cdot 100}$	100,0
2009	$F_p = \sqrt{106,22 \cdot 105,94}$	106,1
2010	$F_p = \sqrt{116,1 \cdot 117,08}$	116,6

6.5. ENLACE Y CAMBIO DE PERÍODO BASE EN LOS NÚMEROS ÍNDICES

El Índice de Laspeyres, que es el más utilizado, y en general todos los índices ponderados que utilizan un año base inicial, tienen como principal problema su pérdida de representatividad a medida que los datos se alejan del periodo base.

Periódicamente, es necesario, en consecuencia, cambiar el año base a fin de hacerlo más representativo; así, por ejemplo, el INE cambia periódicamente el año base del Índice de Precios al Consumo; al cambiar el año base, suele también cambiarse la metodología, de forma que los dos índices son esencialmente distintos, y por lo tanto no se pueden comparar a priori entre sí. El procedimiento a través del cual hacemos comparables números índices obtenidos con bases distintas es lo que se denomina *Enlace*.

Si consideramos que $P_1, P_2, P_3, \ldots$, representan precios correspondientes a intervalos sucesivos de tiempo, llamamos relaciones de enlace a $I_1^2, I_2^3, I_3^4, \ldots,$, el denominado coeficiente legal de enlace, el cual viene dado por la expresión:

$$I_h^0 = \frac{I_0^0}{I_0^h}$$

Dicho coeficiente se basa en la propiedad de la inversión de los números índices, tal que:

$$I_h^t = I_0^t \cdot \frac{I_0^0}{I_0^h}$$

Esta propiedad garantiza que podamos intercambiar índices calculados con distintas bases.

Así, por ejemplo, el INE publica las tablas de enlaces para el cambio de base del Índice de Precios al Consumo de base 2001 a base 2006; en este caso, por grupos de IPC, son las siguientes:

General. Base 2001	0,740268
Alimentos y Bebidas No Alcohólicas	0,780515
Bebidas Alcohólicas y Tabaco	0,546851
Vestido y Calzado	0,843242
Vivienda	0,701667
Menaje	0,78033
Medicina	0,766029
Transporte	0,712176
Comunicaciones	0,825968
Ocio y Cultura	0,753008
Enseñanza	0,575517
Hoteles, Cafés y Restaurantes	0,681374
Otros Bienes y Servicios	0,70226

Fuente: Instituto Nacional de Estadística.

Con este procedimiento, en el momento de transición de un sistema a otro (diciembre de 2005 a enero de 2006) se adopta la variación calculada con el nuevo Sistema. Las series enlazadas se calculan multiplicando cada uno de los índices en base 2001 por estos coeficientes de enlace.

A veces no disponemos del coeficiente de enlace pero sí disponemos de los dos valores del índice en el año de cambio de base; en este caso operamos con una simple regla de tres.

Ejemplo 6.8. Completar la siguiente serie de índices:

Año	Índice base 1988	Índice base 1993	Índice base 1997
1988	100		
1999	150		
1990	145		
1991	154		
1992	132		
1993	165	100	
1994		112	
1995		106	
1996		122	
1997		140	100
1998			112
1999			114
2000			116
2001			118
2002			144
2003			155
2004			160

La serie completada sería la siguiente:

Año	Índice base 1988	Índice base 1993	Índice base 1997
1988	100	60,61	43,29
1989	150	90,91	64,94
1990	145	87,88	62,77
1991	154	93,33	66,67
1992	132	80,00	57,14
1993	165	100	71,43
1994	184,80	112	80,00
1995	174,90	106	75,71
1996	201,30	122	87,14
1997	231,00	140	100
1998	258,72	156,80	112
1999	263,34	159,60	114
2000	267,96	162,40	116
2001	272,58	165,20	118
2002	332,64	201,60	144
2003	358,05	217,00	155
2004	369,60	224,00	160

En la que el valor del índice para 1994, con base 1988, se obtiene a partir del hecho de conocer, que para 1993 tenemos dos datos equivalentes (165 en base 1988 equivale a 100 en base 1993); por ello:

$$I_{88}^{94} = \frac{112 \cdot 165}{100} = 185; \qquad I_{88}^{95} = \frac{106 \cdot 165}{100} = 175; \qquad \text{etc.}$$

Del mismo modo sabemos que para el año 1997 la base 100 equivale a 140 en base 1993, etc.

6.6. DEFLACTACIÓN DE SERIES

La utilidad más importante que tienen los índices de precios, a parte de describir el comportamiento de los precios durante un período concreto, es la de deflactar series cronológicas o temporales valoradas en unidades monetarias.

Deflactar es eliminar el componente de subida de precios que es inherente a toda serie temporal que viene referida a un valor monetario (ventas de una empresa, los depósitos y créditos bancarios, etc...).

Las ventas de una empresa, por ejemplo, se incrementan de un año a otro (ó de un mes a otro), bien por haber aumentado el número de pedidos que realizan los clientes

o bien por que la empresa o el mercado haya decidido una subida en los precios de los artículos pedidos. Si nosotros valoramos el número de pedidos del año actual utilizando los precios vigentes el ejercicio pasado, dispondríamos de un elemento comparativo con respecto al ejercicio anterior que nos señalaría de manera inequívoca si nuestro volumen de negocio se ha incrementado con independencia de lo ocurrido con los precios, y este análisis se alcanza deflactando la serie de ventas de dicha empresa por un índice de precios adecuado.

En consecuencia, cuando utilizamos una serie temporal con referencia para su valoración el precio que rige en un período determinado (un año base), utilizamos una «valoración a precios constantes», en tanto que cuando dicha serie está valorada a los precios vigentes en cada período tendría su valor a «precios corrientes».

En la práctica, para pasar de una serie en *moneda corriente* a otra en *moneda constante* se procede dividiendo la primera por un índice de precios adecuado. Este procedimiento recibe el nombre de deflactación y al índice de precios elegido se le denomina deflactor.

No obstante, hay que señalar que, cuando utilizamos como deflactor un Índice de Laspeyres:

$$\frac{v_t}{L_p} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\frac{\sum_{i=1}^{n} p_{it} \cdot q_{i0}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}}} = \sum_{i=1}^{n} p_{it} \cdot q_{i0} \cdot \frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}} = v_0 \cdot Q_p$$

No pasamos exactamente valores corrientes a valores constantes, cosa que sí ocurre con el Índice de Paasche. Cuando es utilizado como deflactor la expresión queda:

$$\frac{v_t}{P_p} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\sum_{i=1}^{n} p_{i0} \cdot q_{it}}} = \sum_{i=1}^{n} p_{i0} \cdot q_{it}$$

Ejemplo 6.9. Suponiendo un inflación anual constante del 3% durante el período 2006-2010, deflactar la siguiente serie de ventas de una empresa:

Año	Ventas
2006	30
2007	34
2008	38
2009	40
2010	46

El resultado buscado sería el siguiente:

Año	Ventas	Deflactor	Ventas deflactadas (ventas/deflactor)
2006	30	1,0000	30,00
2007	34	1,0300	33,01
2008	38	1,0609	35,82
2009	40	1,0927	36,61
2010	46	1,1255	40,87

Para obtener el deflactor de cada año tenemos que operar de la siguiente forma:

2006	1,0000
2007	$1{,}0000 \cdot (1 + 0{,}03) = 1{,}03$
2008	$1{,}03 \cdot (1 + 0{,}03) = 1{,}0609$
2009	$1{,}0609 \cdot (1 + 0{,}03) = 1{,}0927$
2010	$1{,}0927 \cdot (1 + 0{,}03) = 1{,}1255$

6.7. EJERCICIOS

Sobre números índices

Ejercicio 6.1. *A continuación se detallan las ventas, en millones de unidades, de una determinada empresa:*

Año	1997	1998	1999	2000	2001	2002
Ventas	23	16	22	24	15	13

a) *Calcule un índice simple para cada periodo, tomando como año base 1999.*

b) *Calcule la variación porcentual de ventas entre los años 2001 y 1999.*

Respuesta

a) Para calcular el índice simple, debemos realizar el cociente entre las ventas de cada año y la cantidad correspondiente al año base (1999).

$$I_{99}^{97} = \frac{23}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{97} = 104{,}55.$$

$$I_{99}^{98} = \frac{16}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{98} = 72{,}73.$$

$$I_{99}^{99} = \frac{22}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{99} = 100.$$

$$I_{99}^{00} = \frac{24}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{00} = 109{,}09.$$

$$I_{99}^{01} = \frac{15}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{01} = 68{,}18.$$

$$I_{99}^{02} = \frac{13}{22} \cdot 100 \quad \Rightarrow \quad I_{99}^{02} = 59{,}09.$$

Año	Precio en €	Índice base 1999
1997	23	104,55
1998	16	72,73
1999	22	100,00
2000	24	109,09
2001	15	68,18
2002	13	59,09

b) Para calcular la variación porcentual de las ventas solicitada, al índice correspondiente al año 2001 con base 1999, se le debe restar el valor de 100.

$$VP_{99}^{01} = 68{,}18 - 100 \quad \Rightarrow \quad VP_{99}^{01} = -31{,}8182\%$$

Se produjo una caída en las ventas del 31,82%.

Ejercicio 6.2. *A continuación se detallan los precios mínimos, en euros, por reservas de pasajes efectuadas a un destino promocional entre los meses de Noviembre de 2009 y Aabril de 2010.*

Mes	Precio mínimo
Noviembre	130
Diciembre	160
Enero	200
Febrero	180
Marzo	175
Abril	165

a) *Calcule el incremento de los precios de Marzo respecto a Enero.*

b) *Calcule el incremento de los precios de Marzo respecto a Febrero.*

c) *Obtenga un índice de precios simple cuya base sea Enero.*

d) *Obtenga un índice de precios simple cuya base sea Febrero.*

Respuesta

a) Se debe aplicar la fórmula de incremento o variación porcentual.

$$VP_{\text{enero}}^{\text{marzo}} = \frac{175}{200} \cdot 100 - 100 \quad \Rightarrow \quad VP_{\text{enero}}^{\text{marzo}} = -12,5\%$$

b)

$$VP_{\text{febrero}}^{\text{marzo}} = \frac{175}{180} \cdot 100 - 100 \quad \Rightarrow \quad VP_{\text{febrero}}^{\text{marzo}} = -2,78\%$$

c) Los números índice de cada mes, con base Enero, se obtienen a partir de los siguientes cálculos:

$$I_{\text{enero}}^{\text{noviembre}} = \frac{130}{200} \cdot 100 = 65$$

$$I_{\text{enero}}^{\text{diciembre}} = \frac{160}{200} \cdot 100 = 80$$

$$I_{\text{enero}}^{\text{enero}} = \frac{200}{200} \cdot 100 = 100$$

$$I_{\text{enero}}^{\text{febrero}} = \frac{180}{200} \cdot 100 = 90$$

$$I_{\text{enero}}^{\text{marzo}} = \frac{175}{200} \cdot 100 = 87,5$$

$$I_{\text{enero}}^{\text{abril}} = \frac{165}{200} \cdot 100 = 82,5$$

d) Los números índice de cada mes, con base Febrero, se obtienen a partir de los siguientes cálculos:

$$I_{\text{febrero}}^{\text{noviembre}} = \frac{130}{180} \cdot 100 = 72,22$$

$$I_{\text{febrero}}^{\text{diciembre}} = \frac{160}{180} \cdot 100 = 88,89$$

$$I_{\text{febrero}}^{\text{enero}} = \frac{200}{180} \cdot 100 = 111,11$$

$$I_{\text{febrero}}^{\text{febrero}} = \frac{180}{180} \cdot 100 = 100$$

$$I_{\text{febrero}}^{\text{marzo}} = \frac{175}{180} \cdot 100 \qquad = 97{,}22$$

$$I_{\text{febrero}}^{\text{abril}} = \frac{165}{180} \cdot 100 \qquad = 91{,}67$$

Ejercicio 6.3. *Una empresa vende sus productos en 4 regiones. El volumen de ventas para los meses de Diciembre, Enero, Febrero y Marzo del 2010 resultaron ser los siguientes:*

Mes	Diciembre	Enero	Febrero	Marzo
Región 1	1.500	1.300	1.200	1.000
Región 2	1.450	1.550	1.260	1.050
Región 3	1.600	1.460	1.620	990
Región 4	1.300	1.390	1.400	1.160

a) *Obtenga un índice agregativo simple, sobre la base de Diciembre para el mes de Febrero.*

b) *Calcule el incremento que se ha producido entre Diciembre y Marzo.*

Respuesta

a) Para obtener el índice agregativo simple pedido, se suman los valores de volúmenes de venta correspondientes a Febrero (periodo en estudio) y los correspondientes a Diciembre (periodo base). Luego se efectúa el cociente entre la suma del año en estudio y la suma del año base. Finalmente, el resultado obtenido se multiplica por 100.

Mes	Diciembre	Enero	Febrero
Destino 1	1.500	1.300	1.200
Destino 2	1.450	1.550	1.260
Destino 3	1.600	1.460	1.620
Destino 4	1.300	1.390	1.400
Sumas	5.850	5.700	5.480

$$I_{\text{diciembre}}^{\text{febrero}} = \frac{5.480}{5.850} \cdot 100 = 93{,}6752$$

El índice agregativo simple para el mes de febrero, con base en diciembre del 2002 es 93,6752.

b) $$VP_{\text{diciembre}}^{\text{marzo}} = 93{,}6752 - 100 = -6{,}325\%$$

La variación experimentada entre los meses de Diciembre y Marzo fue una caída del 6,325%.

Ejercicio 6.4. *Una compañía de transporte ofrece un paquete turístico hacia tres destinos distintos. Se detalla en el siguiente cuadro la cantidad de reservas realizadas a cada destino en el periodo comprendido entre 1999 y 2002:*

Año	Londres	Berlín	París
1999	8.500	9.000	10.000
2000	10.800	9.500	11.300
2001	9.000	10.000	9.800
2002	9.100	11.000	12.000

En base a los datos anteriores, calcule los siguientes índices ponderados:

a) *Media Aritmética.*

b) *Media Geométrica.*

c) *Media Armónica.*

d) *Media Agregativa.*

Tomando como año base 1999 y como coeficiente de ponderación para cada destino:

Londres: 1
Berlín: 2
Paris: 3.

Respuesta

Los índices simples para Londres son:

$$IL_{99}^{99} = \frac{8.500}{8.500} \cdot 100 \quad \Rightarrow \quad IL_{99}^{99} = 100$$

$$IL_{99}^{00} = \frac{10.800}{8.500} \cdot 100 \quad \Rightarrow \quad IL_{99}^{00} = 127,06$$

$$IL_{99}^{01} = \frac{9.000}{8.500} \cdot 100 \quad \Rightarrow \quad IL_{99}^{01} = 105,88$$

$$IL_{99}^{02} = \frac{9.100}{8.500} \cdot 100 \quad \Rightarrow \quad IL_{99}^{02} = 107,06$$

Se multiplica por la correspondiente ponderación y se obtienen los valores resumidos en el siguiente cuadro junto con los índices de Berlín y París.

Base: 1999	8.500	9.000	10.000	Ponderación		
				3	2	1
Año	***IL* · 100**	***IB* · 100**	***IP* · 100**	***IL* · pond**	***IB* · pond**	***IP* · pond**
1999	100,00	100,00	100,00	300,00	200,00	100,00
2000	127,06	105,56	113,00	381,18	211,11	113,00
2001	105,88	111,11	98,00	317,65	222,22	98,00
2002	107,06	122,22	120,00	321,18	244,44	120,00

Con estos datos podemos calcular los índices pedidos.

a) La media aritmética ponderada para cada año se obtiene mediante los siguientes cálculos:

$$I\text{ma}_{99}^{99} = \frac{100 \cdot 3 + 100 \cdot 2 + 100 \cdot 1}{100 \cdot 3 + 100 \cdot 2 + 100 \cdot 1} \cdot 100 \quad \Rightarrow \quad I\text{ma}_{99}^{99} = 100$$

$$I\text{ma}_{99}^{00} = \frac{127,06 \cdot 3 + 105,56 \cdot 2 + 113 \cdot 1}{100 \cdot 3 + 100 \cdot 2 + 100 \cdot 1} \cdot 100 \quad \Rightarrow \quad I\text{ma}_{99}^{00} = 117,55$$

$$I\text{ma}_{99}^{01} = \frac{105{,}883 \cdot 3 + 111{,}11 \cdot 2 + 98 \cdot 1}{100 \cdot 3 + 100 \cdot 2 + 100 \cdot 1} \cdot 100 \quad \Rightarrow \quad I\text{ma}_{99}^{01} = 106{,}31$$

$$I\text{ma}_{99}^{02} = \frac{107{,}06 \cdot 3 + 122{,}22 \cdot 2 + 120 \cdot 1}{100 \cdot 3 + 100 \cdot 2 + 100 \cdot 1} \cdot 100 \quad \Rightarrow \quad I\text{ma}_{99}^{02} = 114{,}27$$

b) La media geométrica de cada año será:

$$I\text{mg}_{99}^{99} = \sqrt[6]{100^3 \cdot 100^2 \cdot 100^1} \quad \Rightarrow \quad I\text{mg}_{99}^{99} = 100$$

$$I\text{mg}_{99}^{00} = \sqrt[6]{127{,}06^3 \cdot 105{,}56^2 \cdot 113^1} \quad \Rightarrow \quad I\text{mg}_{99}^{00} = 117{,}13$$

$$I\text{mg}_{99}^{01} = \sqrt[6]{105{,}88^3 \cdot 111{,}11^2 \cdot 98^1} \quad \Rightarrow \quad I\text{mg}_{99}^{01} = 106{,}22$$

$$I\text{mg}_{99}^{02} = \sqrt[6]{107{,}06^3 \cdot 122{,}22^2 \cdot 120^1} \quad \Rightarrow \quad I\text{mg}_{99}^{02} = 114{,}04$$

c) La media armónica de cada año se calcula como sigue:

$$I\text{mar}_{99}^{99} = \frac{6}{\frac{3}{100} + \frac{2}{100} + \frac{1}{100}} \quad \Rightarrow \quad I\text{mar}_{99}^{99} = 100$$

$$I\text{mar}_{99}^{00} = \frac{6}{\frac{3}{127{,}06} + \frac{2}{105{,}56} + \frac{1}{113}} \quad \Rightarrow \quad I\text{mar}_{99}^{00} = 116{,}71$$

$$I\text{mar}_{99}^{01} = \frac{6}{\frac{3}{105{,}88} + \frac{2}{111{,}11} + \frac{1}{98}} \quad \Rightarrow \quad I\text{mar}_{99}^{01} = 106{,}12$$

$$I\text{mar}_{99}^{02} = \frac{6}{\frac{3}{107{,}06} + \frac{2}{122{,}22} + \frac{1}{120}} \quad \Rightarrow \quad I\text{mar}_{99}^{02} = 113{,}81$$

d) Por último, obtenemos la media agregativa:

$$\text{Im}agr_{99}^{99} = \frac{3 \cdot 8.500 + 2 \cdot 9.000 + 1 \cdot 10.000}{3 \cdot 8.500 + 2 \cdot 9.000 + 1 \cdot 10.000} \cdot 100 = 100$$

$$\text{Im}agr_{99}^{00} = \frac{3 \cdot 10.800 + 2 \cdot 9.500 + 1 \cdot 11.300}{3 \cdot 8.500 + 2 \cdot 9.000 + 1 \cdot 10.000} \cdot 100 = 117{,}20$$

$$\text{Im}agr_{99}^{01} = \frac{3 \cdot 9.000 + 2 \cdot 10.000 + 1 \cdot 9.800}{3 \cdot 8.500 + 2 \cdot 9.000 + 1 \cdot 10.000} \cdot 100 = 106{,}17$$

$$\text{Im}agr_{99}^{99} = \frac{3 \cdot 9.100 + 2 \cdot 11.000 + 1 \cdot 12.000}{3 \cdot 8.500 + 2 \cdot 9.000 + 1 \cdot 10.000} \cdot 100 = 114{,}58$$

Conclusión

Año	Aritmética	Geométrica	Armónica	Agregativa
1999	100,00	100,00	100,00	100,00
2000	117,55	117,13	116,71	117,20
2001	106,31	106,22	106,12	106,17
2002	114,27	114,04	113,81	114,58

Ejercicio 6.5. *Los precios promedios y niveles de demanda de tres empresas para un determinado servicio durante los años 2008 y 2010 fueron los siguientes:*

	2008		2010	
	Precio	Demanda en miles	Precio	Demanda en miles
E-I	600	80	650	83
E-II	750	75	775	95
E-III	620	92	630	110

a) *Obtenga los índices de Paasche, Laspeyres, Fisher, Drovisch-Bowley, Edgeworth-Marshall y Walch con base en el año 2008.*

b) *Calcule el incremento en cada caso.*

Respuesta

— Índice de precios de Laspeyres.

	$P(08)\cdot Q(10)$	$P(08)\cdot Q(10)$
	52.000	48.000
	58.125	56.250
	57.960	57.040
Suma	168.085	161.290

$$IPL_{08}^{10} = \frac{168,085}{161.290}\cdot 100 \quad \Rightarrow \quad IPL_{08}^{10} = 104,2129$$

$$Variación = 104,2129 - 100 \quad \Rightarrow \quad Variación = 4,2129\%$$

— Índice de precios de Paasche.

	$P(10)\cdot Q(10)$	$P(08)\cdot Q(08)$
	53.950	49.800
	73.625	71.250
	69.300	68.200
Suma	196.875	189.250

$$IPP_0^n = \frac{196.875}{189.250}\cdot 100 \quad \Rightarrow \quad IPP_0^n = 104,0291$$

$$Variación = 104,0291 - 100 \quad \Rightarrow \quad Variación = 4,0291\%$$

— Índice de precios de Fisher

$$IPF_{08}^{10} = \sqrt{104,2129\cdot 104,0291} \quad \Rightarrow \quad IPF_{08}^{10} = 104,12094$$

$$Variación = 4,012094\%$$

— Índice de Drovisch-Bowley

$$IPDB_{08}^{10} = \frac{104,2129 + 104,0291}{2} \quad \Rightarrow \quad IPDB_{08}^{10} = 104,12099$$

Variación = 4,12099%

— Índice de Edgeworth-Marshall

	$P(10) \cdot [Q(08) + Q(10)]$	$P(08) \cdot [Q(08) + Q(10)]$
	105.950	97.800
	131.750	127.500
	127.260	125.240
Suma	364.960	350.540

$$IPEM_{08}^{10} = \frac{364.960}{350.540} \cdot 100 \quad \Rightarrow \quad IPEM_{08}^{10} = 104,113645$$

Variación: 4,1136 %

— Índice de Walch

	$P(10) \cdot Q(08) \cdot Q(10)$	$P(08) \cdot Q(08) \cdot Q(10)$
	4.316.000	3.984.000
	5.521.875	5.343.750
	6.375.600	6.274.400
Suma	16.213.475	15.602.150

$$IPW_{08}^{10} = \frac{16.213.475}{15.602.150} \cdot 100 \quad \Rightarrow \quad IPW_{08}^{10} = 103,91821$$

Variación: 3,91821 %

Ejercicio 6.6. *En la siguiente tabla se agrupan los ingresos semanales promedio en euros de tres empresas:*

Empresas	2008		2009		2010	
	Ingreso promedio	Cantidad	Ingreso promedio	Cantidad	Ingreso promedio	Cantidad
E-I	105,6	120	110,25	140	190,12	150
E-II	140,8	200	135,26	230	220,36	225
E-III	75,2	60	68,69	70	120,20	100

En base a los datos anteriores se pide calcular:

a) *Índice de precios de Laspeyres y su variación.*

b) *Índice de precios de Paasche y su variación.*

c) *Índice de precios de Fisher y su variación.*

d) *Comprobar si se cumple la propiedad de inversión temporal para cada índice calculado.*

Tomando como base el año 2009 y como año de estudio el año 2010.

Respuesta

a) Índice de Laspeyres. Base 2009.

	$P(10)\cdot Q(09)$	$P(09)\cdot Q(09)$
	26.616,80	15.435,00
	50.682,80	31.109,80
	8.414,00	4.808,30
Suma	85.713,60	51.353,10

$$IPL_{09}^{10} = \frac{85.713,6}{51.353,1}\cdot 100 \quad \Rightarrow \quad IPL_{09}^{10} = 166,91$$

$$Variación = 166,91 - 100 \quad \Rightarrow \quad Variación = 66,91\%$$

b) Índice de Paasche. Base 2009.

	$P(10)\cdot Q(10)$	$P(09)\cdot Q(10)$
	28.518,00	16.537,50
	49.581,00	30.433,50
	12.020,00	6.869,00
Suma	90.119,00	53.840,00

$$IPP_{09}^{10} = \frac{90.119}{53.840}\cdot 100 \quad \Rightarrow \quad IPP_{09}^{10} = 167{,}383$$

$$Variación = 167{,}383 - 100 \quad \Rightarrow \quad Variación = 67{,}383\%$$

c) Índice de Fisher. Base 2009.

$$IPF_{09}^{10} = \sqrt{166{,}91\cdot 167{,}383} \quad \Rightarrow \quad IPF_{09}^{10} = 167{,}1465$$

$$Variación = 167{,}1465 - 100 \quad \Rightarrow \quad Variación = 67{,}1465\%$$

d) Se cumple la propiedad de inversión temporal en el caso que se verifique la siguiente relación:

$$\frac{I_{09}^{10}}{100}\cdot\frac{I_{10}^{09}}{100} = 1$$

Para hallar esta relación debemos obtener los correspondientes índices de Laspeyres, Paasche y Fisher con base 2010.

— Índice de Laspeyres. Base 2010.

	$P(09)\cdot Q(10)$	$P(10)\cdot Q(10)$
	16.537,50	28.518,00
	30.433,50	49.581,00
	6.869,00	12.020,00
Suma	53.840,00	90.119,00

$$IPL_{10}^{09} = \frac{53.840}{90.119} \cdot 100 \quad \Rightarrow \quad IPL_{10}^{09} = 59,7432$$

— Índice de Paasche. Base 2010

	$P(09) \cdot Q(09)$	$P(10) \cdot Q(09)$
	15.435,00	26.616,80
	31.109,80	50.682,80
	4.808,30	8.414,00
Suma	51.353,10	85.713,60

$$IPP_{10}^{09} = \frac{51.353,10}{85.713,6} \cdot 100 = 59,9124$$

— Índice de Fisher. Base 2010

$$IPF_{10}^{09} = \sqrt{59,7432 \cdot 59,9124} = 59,82777$$

Con estos resultados, ya estamos en condiciones de verificar la propiedad de inversión temporal para cada índice.

— Inversión temporal para índice de Laspeyres:

$$\frac{IPL_{09}^{10}}{100} \cdot \frac{IPL_{10}^{09}}{100} = \frac{166,9102742}{100} \cdot \frac{59,7432284}{100} = 0,997175863$$

Por lo tanto se concluye que no se verifica esta propiedad para el índice de Laspeyres.

— Inversión temporal para índice de Paasche:

$$\frac{IPP_{09}^{10}}{100} \cdot \frac{IPP_{10}^{09}}{100} = \frac{167,3829866}{100} \cdot \frac{59,9124293}{100} = 1,002832135$$

Por lo tanto se concluye que no se verifica esta propiedad para el índice de Paasche.

— Inversión temporal para índice de Fisher:

$$\frac{IPF_{09}^{10}}{100}\cdot\frac{IPF_{10}^{09}}{100}=\frac{167{,}1464633}{100}\cdot\frac{59{,}82776903}{100}=1$$

El índice de Fisher siempre satisface la inversión temporal. Este es un ejemplo donde se puede verificar esta propiedad.

Ejercicio 6.7. *Calcule los siguientes índices de precios para los artículos detallados en la tabla:*

a) *Agregativo Simple.*
b) *Laspeyres.*
c) *Paasche.*
d) *Fisher.*
e) *Drovisch-Bowley.*
f) *Edgeworth-Marshall.*
g) *Walch.*

Para ello, utilice como año base 1999.

Artículos	2002		1999	
	Euros	Cantidad	Euros	Cantidad
Leche	1,3	1.500	1,1	1.050
Yogurt	2	1.300	1,7	2.000
Manteca	3	300	3,5	300
Queso	4	500	5,2	400

Respuesta

a) Agregativo Simple

	$P(02)$	$P(99)$
	1,3	1,1
	2	1,7
	3	3,5
	4	5,2
Suma	10,30	11,50

$$I_{99}^{02} = \frac{10,3}{11,5} \cdot 100 = 89,5652$$

b) Índice de Laspeyres

	$P(02) \cdot Q(99)$	$P(99) \cdot Q(99)$
	1.365,00	1.155,00
	4.000,00	3.400,00
	900,00	1.050,00
	1.600,00	2.080,00
Suma	7.865,00	7.685,00

$$IPL_{99}^{02} = \frac{7.865}{7.685} \cdot 100 = 102,34$$

c) Índice de Paasche

	$P(02) \cdot Q(02)$	$P(99) \cdot Q(02)$
	1.950,00	1.650,00
	2.600,00	2.210,00
	900,00	1.050,00
	2.000,00	2.600,00
Suma	7.450,00	7.510,00

$$IPP_{99}^{02} = \frac{7.450}{7.510} \cdot 100 = 99,2010$$

d) Índice de Fisher

$$IPF_{99}^{02} = \sqrt{102,3422 \cdot 99,201} = 100,7594$$

e) Índice de Drovisch-Bowley

$$IPDB_{99}^{02} = \frac{102,3422 + 99,2010}{2} = 100,77165$$

f) Índice de Edgeworth-Marshall

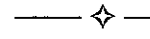
	$P(02)\cdot[Q(99)+Q(02)]$	$P(99)\cdot[Q(99)+Q(02)]$
	331.500	280.500
	660.000	561.000
	180.000	210.000
	360.000	468.000
Suma	1.531.500	1.519.500

$$IPEM_{99}^{02}=\frac{1.531.500}{1.519.500}\cdot 100=100,7897$$

g) Índice de Walch

	$P(02)\cdot Q(99)\cdot Q(02)$	$P(99)\cdot Q(99)\cdot Q(02)$
	2.047.500	1.732.500
	5.200.000	4.420.000
	270.000	315.000
	800.000	1.040.000
Suma	8.317.500	7.507.500

$$IPW_{99}^{02}=\frac{8.317.500}{7.507.500}\cdot 100 \quad \Rightarrow \quad IPW_{99}^{02}=110,7892$$



Ejercicio 6.8. *A continuación se detallan los costes de cultivo en los que ha incurrido una determinada empresa agrícola. Se detallan las hectáreas cultivadas y las diversas especies agrícolas:*

Especies	2007		2008		2009	
	Coste/Ha.	Has. (en miles)	Coste/Ha.	Has. (en miles)	Coste/Ha.	Has. (en miles)
Trigo	80	6	106	10	115	13
Cebada	90	10	100	9	106	10
Avena	105	13,5	116	12	125	15
Centeno	72	8,6	80	5	82	7
Tomate	115	15,9	112	16	120	19
Pimiento	95	18,2	91	9	98	11
Judías verdes	100	9,25	100	8,5	112	10,6
Lechugas	86	9,68	85,6	13	90,5	15
Alcachofa	104	7,2	116	7	119	10
Coliflor	89	5	91	6	105	8

Calcule el índice de coste (Laspeyres) para el año 2009 con base 2008 y el índice cantidad (Paasche) de hectáreas cultivadas para el año 2009 con base 2007.

Respuesta

Para resolver el ejercicio utilizaremos el esquema de precio-cantidad visto en los puntos anteriores. Para ello consideraremos al coste como el precio y al número de de hectáreas como la cantidad.

Sumas para el índice de precios de Laspeyres

	$P(09) \cdot Q(08)$	$P(08) \cdot Q(08)$
	1.150,00	1.060,00
	954,00	900,00
	1.500,00	1.392,00
	410,00	400,00
	1.920,00	1.792,00
	882,00	819,00
	952,00	850,00
	1.176,50	1.112,80
	833,00	812,00
	630,00	546,00
Suma	10.407,50	9.683,80

Índice de precios de Laspeyres:

$$IPL_{08}^{09} = \frac{10.407,50}{9.683,8} \cdot 100 = 107,47$$

Sumas para el índice de cantidades de Paasche

	$Q(09) \cdot P(09)$	$Q(07) \cdot P(09)$
	1.495,00	690,00
	1.060,00	1.060,00
	1.875,00	1.687,50
	574,00	705,20
	2.280,00	1.908,00
	1.078,00	1.783,60
	1.187,20	1.036,00
	1.357,50	876,04
	1.190,00	856,80
	840,00	525,00
Suma	12.936,70	11.128,14

Índice de cantidades de Paasche

$$IQP_{07}^{09} = \frac{\sum q_{02} \cdot p_{02}}{\sum q_{00} \cdot p_{02}} \cdot 100 \Rightarrow IQP_{07}^{09} = \frac{12.936,7}{11.128,14} \cdot 100 \Rightarrow IQP_{07}^{09} = 116,2521$$

Ejercicio 6.9. *Para un conjunto de servicios que brinda una empresa, se tienen dos series de números índices de precios de Laspeyres con distintas bases.*

Halle el coeficiente de enlace y realice el enlace correspondiente, tomando como base el año 1998.

Años	Base 1988	Base 1998
1995	113	
1996	120	
1997	114	
1998	109	100
1999		115
2000		118
2001		117
2002		120

Respuesta

El coeficiente de enlace es:

$$I_{1998}^{1988} = \frac{I_{1988}^{1988}}{I_{1988}^{1998}} = \frac{100}{109} = 0{,}9174$$

El enlace resulta de los siguientes cálculos:

$$I_{98}^{95} = I_{88}^{95} \cdot \frac{I_{88}^{88}}{I_{88}^{98}} = 113 \cdot 0{,}9174 = 103{,}6662$$

$$I_{98}^{96} = 120 \cdot 0{,}9174 = 110{,}09$$

$$I_{98}^{97} = 114 \cdot 0{,}9174 = 104{,}59$$

$$I_{98}^{98} = 109 \cdot 0{,}9174 = 100$$

El nuevo conjunto de índices actualizados a la nueva base es, por tanto:

Años	Base 1998
1995	103,67
1996	110,09
1997	104,59
1998	100
1999	115
2000	118
2001	117
2002	120

Ejercicio 6.10. *En la siguiente serie de números índice, efectúe un cambio de base desde 2006 a 2008.*

Años	Índices base 2006
2006	100
2007	117,6
2008	135,6
2009	154,3
2010	210,5

Respuesta

El coeficiente de transformación es:

$$I_{2008}^{2006} = \frac{100}{135,6} = 0,7375$$

El cambio de base se efectúa de la siguiente manera:

$$I_{08}^{06} = I_{06}^{06} \cdot 0,7375 = 100 \cdot 0,7375 = 73,75$$
$$I_{08}^{07} = 117,6 \cdot 0,7375 = 86,73$$
$$I_{08}^{08} = 135,6 \cdot 0,7375 = 100$$
$$I_{08}^{09} = 154,3 \cdot 0,7375 = 113,796$$
$$I_{08}^{10} = 210,5 \cdot 0,7375 = 155,24$$

El resultado del cambio de base es, por tanto:

Años	Índices base 2008
2006	73,75
2007	86,73
2008	100
2009	113,79
2010	155,24

Capítulo 7

SERIES TEMPORALES

7.1. INTRODUCCIÓN

Una serie temporal, también llamada serie cronológica o histórica, se define se define como una **sucesión *de observaciones de una variable en distintos momentos del tiempo;*** habitualmente estas observaciones se presentan en ***intervalos regulares de tiempo y ordenadas cronológicamente.***

La serie temporal puede estar generada con datos continuos o discretos, flujos o stocks[11], valorados en unidades monetarias o en magnitudes físicas, con periodicidad diaria, semanal, mensual, trimestral, anual, bianual, etc. pero lo que caracteriza a la serie temporal es la presencia de una referencia cronológica concreta y determinada.

El análisis estadístico de series temporales se utiliza actualmente en múltiples ramas de la ciencia (medicina, física, ingeniería, meteorología, etc.), y muy especialmente, en economía.

Este análisis tiene como objetivo principal la predicción de comportamientos futuros de la variable en función del comportamiento pasado de la misma (serie). Se trata, en consecuencia, de analizar la serie con el fin de extraer regularidades o patrones de comportamiento que se hayan producido en el pasado y que en consecuencia se pueda preverse que se van a reproducir en el futuro.

El estudio estadístico de las series temporales se ha llevado a cabo históricamente mediante tres enfoques diferentes pero relacionados entre sí:

— El denominado «tratamiento o modelado clásico de series temporales», que se basa en la descomposición de la serie en los diversos componentes que estudiaremos más tarde: la tendencia, las variaciones estacionales y los ciclos. Su enfoque es eminentemente descriptivo o de predicción a corto plazo.

---

[11] Son datos flujo datos generados en un período determinado de tiempo: un día, un mes, un año, etc.. y datos stock los referidos a una fecha determinada: 31 de diciembre de cada año.

Un ejemplo de datos flujos son las ventas de una empresa, ya que tendrán un valor si se toma al cabo de un día, una semana, un mes ó un año; sin embargo, el valor de las acciones en la bolsa solo puede ser registrado a una fecha determinada por ejemplo a 31 de diciembre. Nótese que con datos stock también se puede tomar una serie diaria, semanal, mensual o anual, lo que dependerá de la frecuencia con la que registremos el dato, si lo hacemos cuando cierra la jornada de la bolsa generaremos una serie diaria, si lo hacemos únicamente un día determinado de la semana estaremos generando una serie semanal, si fuera a determinada fecha de cada mes, una mensual o si lo hacemos al finalizar el año, una serie anual.

— Los modelos ARIMA, también conocidos como modelos Box-Jenkins, en honor a sus creadores[12]. Además de la descomposición de la serie, como modelos que son, se orientan a la predicción, siendo necesario en este caso un número mínimo de observaciones (entre 60 a 100 según los autores) y, en todo caso, dicho número también dependerá de la complejidad del modelo generado.
— El denominado análisis espectral. Se centra especialmente en el análisis de la componente estacional, utilizando para ello los denominados armónicos de frecuencia, que son funciones que dependen del tiempo expresadas en forma de senos y cosenos.

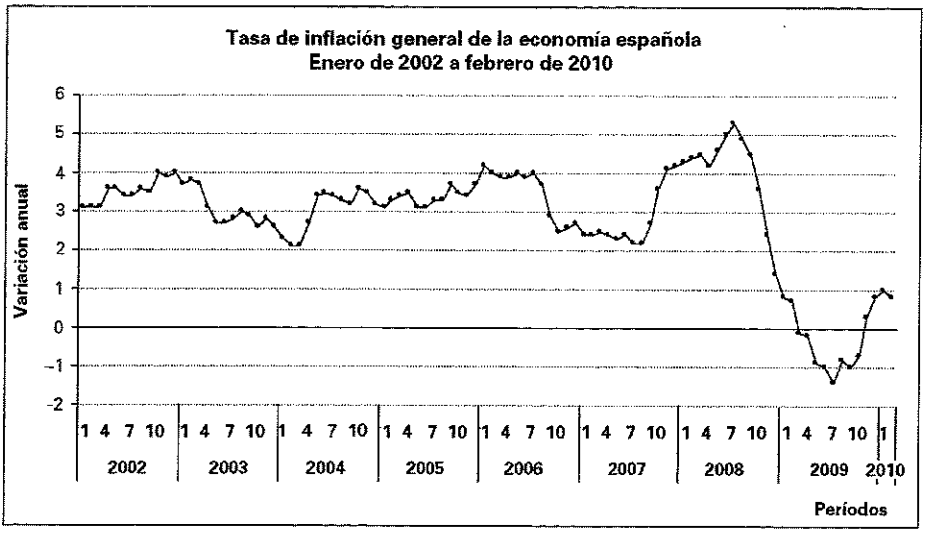
En este libro sólo abordaremos el primero de estos enfoques; el alumno interesado en profundizar sobre esta materia puede estudiar otros manuales especializados[13].

7.2. REPRESENTACIÓN GRÁFICA

La representación gráfica de las series es bastante simple; en la más habitual se indica en el eje cartesiano de abscisas los datos temporales (años, trimestres, meses, días, etc.) y en el de ordenadas los datos de la variable.

Un ejemplo tipo es el siguiente:

Gráfico 7.1.



[12] G.E.P. Box, profesor de Estadística de la Universidad de Wisconsin, y G.M. Jenkins, profesor de Ingeniería de Sistemas de la Universidad de Lancaster, publicaron en 1976 su trabajo *Time Series Analysis: Forecasting and Control*, que dio origen a este nuevo enfoque muy utilizado en los años 80 y 90 en los más diversos campos de la predicción temporal.

[13] Puede verse, por ejemplo, Uriel, E. (1995). *Análisis de Datos: Series Temporales y Análisis de la Varianza*. AC, Madrid o Gujarati, Damodar N. (1997). *Econometría*. Tercera Edición. McGrawHill.

El primer paso obligatorio para analizar una serie temporal es presentar un gráfico de la evolución de la variable a lo largo del tiempo para tratar de encontrar visualmente algún patrón o regularidad.

7.3. COMPONENTES DE UNA SERIE TEMPORAL

El análisis o modelado tradicional de una serie temporal se basa en considerar que la misma puede dividirse en cuatro componentes diferenciadas, llamadas tendencia (T), fluctuación cíclica (C), variación estacional (S) y movimientos irregulares (I).

La **tendencia** de una serie temporal es el componente que indica la dirección en la que se mueve la serie en el largo plazo.

La **estacionalidad** recoge las oscilaciones a corto plazo, entendiendo como tales aquellas cuya duración es igual o inferior a un año; la estacionalidad se produce por el efecto de algún periodo de tiempo que influye en la serie (estaciones del año, días de la semana, etc.); las razones de la estacionalidad son de tipo físico-natural (tiempo meteorológico, ciclos biológicos, etc.) o de tipo institucional (vacaciones escolares, fiestas, horarios comerciales, etc.).

El **factor cíclico** recoge las oscilaciones de carácter periódico, pero no regular, y a medio plazo; se considera que el periodo de cada ciclo siempre es superior al año; este componente es frecuente encontrarlo en las series económicas y se debe a los cambios en la actividad económica.

Las **fluctuaciones irregulares** se producen con motivo de un acontecimiento especial y ocasional concreto; una vez extraídos los anteriores componentes (tendencia, estacionalidad y ciclo) suele quedar este componente derivado de movimientos irregulares, ocasionales o aleatorios.

Si tomamos como ejemplo una serie mensual con la evolución del número de turistas en una determinada ciudad, tendremos probablemente variaciones estacionales (meses de verano, invierno, etc.), cíclicas (períodos de crisis) o irregulares (con motivo de un acontecimiento especial y ocasional concreto), aisladas todas ellas, la tendencia nos dirá sí a largo plazo el turismo de la ciudad está estabilizado, crece o disminuye a un determinado ritmo.

La asociación de las cuatro componentes en la serie temporal (Y) dan como resultado el dato observado; esta asociación de componentes puede ser:

$$\textbf{Aditiva: } y = T + C + S + I$$

En este caso el valor de la serie en cada instante es igual a la suma correspondiente a los cuatro componentes.

$$\textbf{Multiplicativa: } y = T \cdot C \cdot S \cdot I$$

En este caso el valor de la serie en cada instante es igual al producto correspondiente a los cuatro componentes.

La razón de utilizar una de estas dos hipótesis, en lugar de otras más complejas, radica fundamentalmente en la sencillez y en la operatividad de las mismas.

También puede suponerse un **esquema mixto**, como combinación de ambas, por ejemplo: $y = T \cdot C \cdot S + I$, $y = T \cdot C + S + I$, etc.

Por otra parte, es necesario señalar que en una serie concreta no tienen por qué darse los cuatro componentes. De hecho, una serie con periodicidad anual carece de estacionalidad.

7.4. CÁLCULO Y ANÁLISIS DE LA TENDENCIA

La tendencia es la componente de la serie temporal que representa la evolución a largo plazo de la misma. La tendencia se asocia al movimiento uniforme o regular observado en la serie durante un período de tiempo extenso.

La tendencia es la información más relevante de la serie temporal ya que es *el componente fundamental para realizar predicciones* sobre el comportamiento futuro de la serie.

Los medios más utilizados para detectar y eliminar la tendencia de una serie se basan en la aplicación de filtros a los datos. *Un filtro no es más que una función matemática que aplicada a los valores de la serie produce una nueva serie con unas características determinadas.* Entre esos filtros encontramos las medias móviles.

Los métodos clásicos de análisis de la tendencia son tres: los semipromedios, los ajustes de una función por mínimos cuadrados y el método de los promedios (o medias) móviles.

El método de los semipromedios sólo es válido para los ajustes de tipo lineal. Respecto a los ajustes por mínimos cuadrados, las funciones de tiempo han de ser continuas y diferenciables; las funciones de tendencia más utilizadas son:

— Lineal.
— Polinómica.
— Exponencial.
— Modelo autorregresivo.
— Función logística.
— Curva de Gompertz.
— Modelo logarítmico recíproco.

7.4.1. Cálculo de la tendencia por el método de los semipromedios

El método de los semipromedios es la forma más rápida de estimar una línea de tendencia recta. El método requiere dividir la serie de datos en dos mitades y calcular el promedio de cada mitad que se centra en el punto medio. La recta que una ambas medias (o semipromedios) sería la línea de tendencia estimada.

Ejemplo 7.1. Utilizando la serie cronológica del coste laboral por trabajador en España, realizar un ajuste de una tendencia basada en el método de semipromedios:

Evolución del coste total por trabajador (euros) en España, periodo 2005-2008

Año	Trimestre	Euros
2005	1	2.063,93
2005	2	2.141,24
2005	3	2.055,75
2005	4	2.251,93
2006	1	2.154,32
2006	2	2.254,38
2006	3	2.152,88
2006	4	2.358,85
2007	1	2.239,53
2007	2	2.339,64
2007	3	2.242,03
2007	4	2.459,71
2008	1	2.342,28
2008	2	2.451,40
2008	3	2.350,17
2008	4	2.583,82

Fuente: Encuesta Trimestral de Coste Laboral. INE

Dividimos la serie en dos mitades, cada una de 8 trimestres y calculamos los promedios de cada mitad.

$$\text{Promedio}_{2005\text{-}2006} = \frac{2.063{,}93 + 2.141{,}24 + 2.055{,}75 + 2.251{,}93 + 2.154{,}32 + 2.254{,}38 + 2.152{,}88 + 2.358{,}85}{8} = 2.179{,}16$$

$$\text{Promedio}_{2007\text{-}2008} = \frac{2.239{,}53 + 2.339{,}64 + 2.242{,}03 + 2.459{,}71 + 2.342{,}28 + 2.451{,}40 + 2.350{,}17 + 2.583{,}82}{8} = 2.376{,}07$$

El primer semipromedio se centra entre el cuarto trimestre de 2005 y el primero de 2006 y el segundo entre el cuarto de 2007 y el primero de 2008.

La tendencia se obtiene calculando una línea recta:

$$y_t = a + bx_t$$

Donde los valores de x_t se elaboran a partir de una sucesión de puntuaciones consecutivas, que van desde un mínimo 1 en el primer trimestre de 2005 hasta un máximo de 16 en el cuarto de 2008; el coeficiente de la pendiente de la recta b representaría por tanto el incremento trimestral de la tendencia, disponiendo para estimar los parámetros a y b de los puntos correspondientes a los semipromedios, es decir:

$$x_t = 4{,}5 \quad y_t = 2.179{,}16$$
$$x_t = 12{,}5 \quad y_t = 2.376{,}07$$

El coeficiente b se calcula a partir de los dos semipromedios del siguiente modo:

$$b = \frac{2.376{,}07 - 2.179{,}16}{8} = 24{,}61$$

Estimando, por tanto, un incremento medio del coste laboral de 24,61 euros al trimestre. El valor de a se puede obtener considerando cualquiera de ambos puntos; considerando el primer semipromedio:

$$a = 2.179{,}16 - 4{,}5 \cdot 24{,}61 = 2.068{,}40$$

Si consideramos el segundo:

$$a = 2.376{,}07 - 12{,}5 \cdot 24{,}61 = 2.068{,}40$$

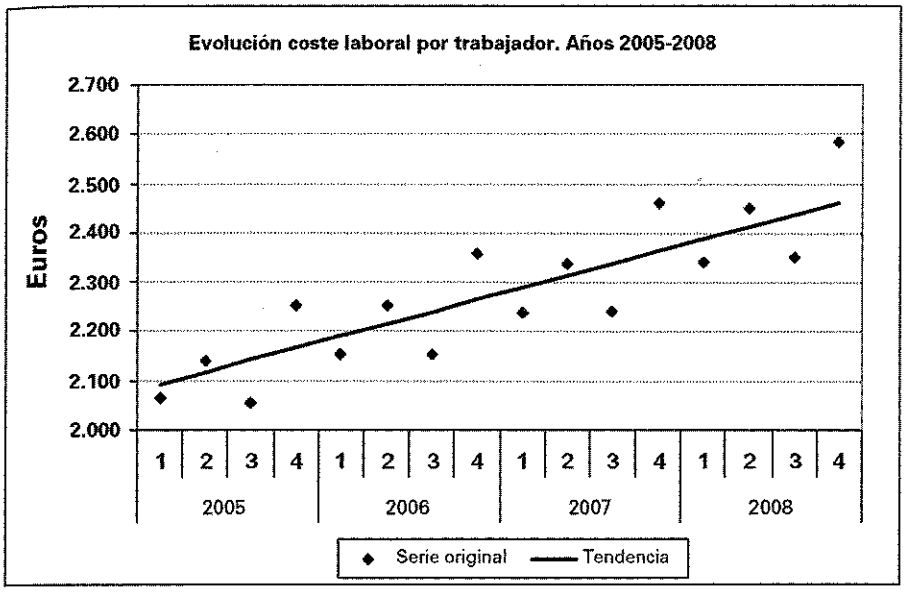
Siendo por tanto la ecuación de tendencia:

$$y_t = 2.068{,}40 + 24{,}61x_t \qquad \text{con } x_t = 1, 2, \ldots, 16$$

Se muestra en la siguiente tabla la línea de tendencia calculada.

Año	Trimestre	Serie original Euros	x_t	y_t Tendencia Euros
2005	1	2.063,93	1	2.093,01
2005	2	2.141,24	2	2.117,62
2005	3	2.055,75	3	2.142,24
2005	4	2.251,93	4	2.166,85
2006	1	2.154,32	5	2.191,47
2006	2	2.254,38	6	2.216,08
2006	3	2.152,88	7	2.240,70
2006	4	2.358,85	8	2.265,31
2007	1	2.239,53	9	2.289,92
2007	2	2.339,64	10	2.314,54
2007	3	2.242,03	11	2.339,15
2007	4	2.459,71	12	2.363,77
2008	1	2.342,28	13	2.388,38
2008	2	2.451,40	14	2.412,99
2008	3	2.350,17	15	2.437,61
2008	4	2.583,82	16	2.462,22

Representamos a continuación gráficamente la serie original y la línea de tendencia obtenida:



7.4.2. Cálculo de la tendencia por el método de los mínimos cuadrados

El método de mínimos cuadrados es el que más se utiliza para ajustar tendencias. Este método ya fue explicado en los epígrafes 5.8 a 5.11, por lo que no nos extenderemos nuevamente sobre el mismo.

7.4.3. Cálculo de la tendencia por el método de las Medias Móviles

Si queremos calcular tendencias sin tener que ajustarnos a una función previa, debemos utilizar la ***técnica de las medias móviles***. Una media móvil es un promedio de un número determinado de datos, u **orden de la media móvil**, que se imputa siempre a la fecha central si se elige un número impar de datos; sí el número es par, lo más correcto es imputarlo a la primera referencia de las dos fechas centrales. *La ventaja que tiene la media móvil es la flexibilidad y la facilidad de cálculo.*

Frente a estas ventajas tiene, sin embargo, dos inconvenientes a tener en cuenta:

— **La pérdida de información**; uno de los principales inconvenientes de la media móvil es que se pierde información de la tendencia en los ejercicios inicial y final (en el caso de media móvil con periodicidad de tres términos se han perdido dos datos, el primero y el último, pero en el caso de medias móviles con mayor periodicidad perderíamos más información).

— **La decisión, que es relativamente arbitraria, del número de periodos utilizados para calcularla y la variabilidad que ello conlleva**, ya que los da-

tos obtenidos con una media móvil de 3 períodos son bastante diferentes de la media para la misma serie pero con 5 períodos de cálculo.

Ejemplo 7.2. Ajustar una tendencia por el método de las medias móviles a la serie correspondiente al número de visados de dirección de obra, obra nueva, de viviendas por meses en España durante el periodo 2006 a 2009.

Ajustaremos dos líneas de tendencia, una de orden 3 (3 meses) y otra de orden 9 (9 meses). Los resultados son los siguientes:

Año	Mes	Dato	Media móvil orden 3	Media móvil orden 9
2006	Enero	17.056		
2006	Febrero	19.097	18.772	
2006	Marzo	20.163	18.062	
2006	Abril	14.925	18.081	
2006	Mayo	19.155	17.810	18.870
2006	Junio	19.351	19.387	18.615
2006	Julio	19.654	17.831	17.980
2006	Agosto	14.487	20.029	17.048
2006	Septiembre	25.945	18.398	16.765
2006	Octubre	14.761	18.028	16.089
2006	Noviembre	13.378	13.306	16.071
2006	Diciembre	11.780	12.510	15.051
2007	Enero	12.373	12.409	14.820
2007	Febrero	13.075	14.879	13.244
2007	Marzo	19.189	14.245	12.997
2007	Abril	10.470	14.023	12.358
2007	Mayo	12.409	11.548	12.090
2007	Junio	11.764	12.235	11.833
2007	Julio	12.532	10.642	11.403
2007	Agosto	7.630	9.842	10.114
2007	Septiembre	9.365	9.018	9.779
2007	Octubre	10.060	9.543	9.204
2007	Noviembre	9.205	8.951	8.625
2007	Diciembre	7.587	8.083	8.163
2008	Enero	7.456	7.428	7.994
2008	Febrero	7.241	7.083	7.588
2008	Marzo	6.551	7.387	7.099
2008	Abril	8.369	7.009	6.388
2008	Mayo	6.108	6.729	6.022
2008	Junio	5.711	5.828	5.702
2008	Julio	5.665	4.727	5.329
2008	Agosto	2.805	4.255	4.960
2008	Septiembre	4.294	3.892	4.338
2008	Octubre	4.576	4.250	3.979
2008	Noviembre	3.880	3.897	3.692
2008	Diciembre	3.236	3.294	3.397

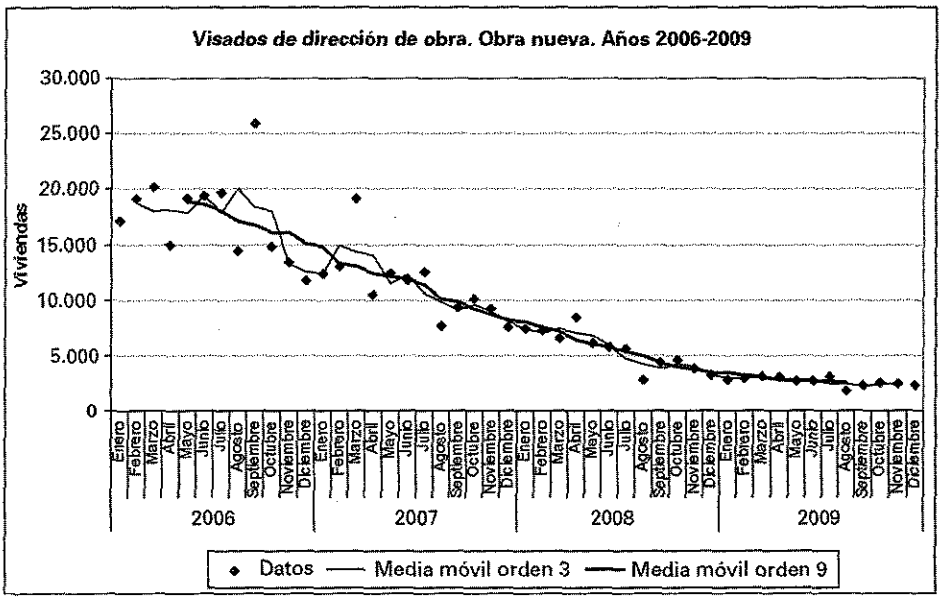
(Continúa)

(Continúa)

Año	Mes	Dato	Media móvil orden 3	Media móvil orden 9
2009	Enero	2.765	2.960	3.387
2009	Febrero	2.880	2.924	3.205
2009	Marzo	3.127	3.005	3.038
2009	Abril	3.008	2.951	2.799
2009	Mayo	2.719	2.792	2.694
2009	Junio	2.650	2.817	2.676
2009	Julio	3.081	2.485	2.625
2009	Agosto	1.725	2.367	2.534
2009	Septiembre	2.294	2.206	
2009	Octubre	2.599	2.439	
2009	Noviembre	2.425	2.442	
2009	Diciembre	2.303		

FUENTE: Ministerio de Fomento.

La representación gráfica de los datos originales y las medias móviles calculadas sería la siguiente.



Como puede observarse en el gráfico, a medida que aumentamos el orden de la media móvil, iremos obteniendo una tendencia más clara (o más suavizada), pero en contraposición aumentamos la pérdida de información; dos meses en el caso de la media móvil de orden tres, uno al principio y otro al final, y 8 meses en la de orden nueve, cuatro al principio y cuatro al final.

7.5. ANÁLISIS DE LAS VARIACIONES ESTACIONALES

Entendemos por variaciones estacionales los ciclos regulares de duración inferior al año. Las variaciones o ciclos estacionales son muy frecuentes en las series temporales sea cual sea su naturaleza, pueden presentar un esquema horario, diario, semanal, mensual, trimestral o incluso semestral, no siendo necesario que tengan alguna relación con las estaciones del año. Lo verdaderamente importante de los ciclos estacionales es su temporalidad o repetición regular. Precisamente la práctica totalidad de las series que tienen como referencia al sector turístico se ven fuertemente afectadas por los ciclos estacionales (afluencia de turistas, gasto turístico, clientes de una agencia de viajes, etc.).

El procedimiento de descontar los efectos que provoca la existencia de un ciclo estacional se llama ***desestacionalización***. Existen diferentes procedimientos para realizar un ajuste estacional en las series temporales cuya solución requiere de un cálculo matemático relativamente complejo, aquí únicamente estudiaremos los dos procedimientos de desestacionalización más sencillos: el método del porcentaje promedio y el método del porcentaje promedio móvil.

Con carácter previo conviene, no obstante, aplicar un test para determinar si una serie temporal presenta variaciones estacionales de relevancia; para ello se utilizan algunas herramientas cuyo detalle no veremos en este texto (análisis de la varianza del componente estacional de la serie y el estadístico denominado *F de Snedecor*, cuyo valor, comparado con el denominado valor crítico, nos permite determinar si tiene significación el factor temporal para explicar la varianza de la serie; si el test nos resulta positivo indicaría que los movimientos estacionales de la serie son lo suficientemente determinantes como para proceder a su desestacionalización posterior, mientras que en caso contrario debemos despreciar el componente estacional.

Los cálculos manuales para llevar a cabo este tipo de análisis suelen ser complejos y exigen la disponibilidad de una tabla estadística para la citada *F de Snedecor*, debiendo acudirse normalmente a programas informáticos como la Hoja de Cálculo EXCEL, el SPSS, etc.

7.5.1. Cálculo de la variación estacional por el método del porcentaje promedio

El método del porcentaje promedio es un procedimiento rápido y simple para elaborar un índice estacional, que permite valorar y visualizar mejor el grado de estacionalidad de dicha serie. El índice estacional se supone de carácter multiplicativo.

Se procede de la siguiente forma:

— Se obtienen los promedios anuales.
— Se obtienen los porcentajes de las cifras mensuales en relación al promedio anual.
— Se elabora un índice estacional para cada mes, con el promedio de las cantidades obtenidas en el paso anterior.

Ejemplo 7.3. Elaborar un índice de estacionalidad por el método del porcentaje promedio para la siguiente serie temporal, correspondiente al volumen de exportaciones españolas durante el periodo 2002 a 2009 y desestacionalizar la serie.

Volumen de exportaciones españolas (millones de euros). Años 2002-2009

	2002	2003	2004	2005	2006	2007	2008	2009
Enero	10.139,2	10.319,9	10.698,8	10.904,6	12.753,0	13.968,5	14.928,3	11.092,4
Febrero	10.802,2	11.211,9	11.528,1	12.141,3	13.992,1	14.859,7	16.621,4	12.400,6
Marzo	11.061,1	12.249,8	13.079,6	12.884,6	15.449,7	16.301,8	15.881,6	13.714,2
Abril	11.409,6	12.273,5	12.361,0	13.405,2	13.161,2	14.399,1	17.963,8	13.192,1
Mayo	11.151,1	12.356,6	12.702,3	13.307,2	15.471,5	16.018,8	16.621,3	12.893,4
Junio	10.700,4	11.679,9	13.084,4	13.581,1	15.192,1	16.109,7	15.464,0	13.895,9
Julio	11.159,6	11.430,9	12.880,6	12.799,6	13.596,9	15.321,8	17.188,9	14.474,7
Agosto	8.777,3	8.268,2	8.891,0	9.920,4	11.264,4	12.125,7	12.121,8	10.072,3
Septiembre	10.520,8	11.707,8	12.103,0	13.516,0	14.092,5	14.903,1	17.290,4	13.871,1
Octubre	12.610,3	13.068,7	13.002,1	13.215,8	15.263,5	16.706,6	16.671,5	14.918,2
Noviembre	11.863,9	11.674,8	13.779,2	14.592,6	15.096,8	16.568,1	14.288,9	14.067,7
Diciembre	10.618,5	11.573,1	12.350,2	13.290,6	14.538,1	14.195,6	13.142,3	13.661,5

FUENTE: Boletín Mensual de Estadística INE.

Los totales y promedios anuales serían:

	2002	2003	2004	2005	2006	2007	2008	2009
Total	130.814,1	137.815,3	146.460,4	153.559,0	169.871,9	181.478,5	188.184,3	158.254,3
Media	10.901,2	11.484,6	12.205,0	12.796,6	14.156,0	15.123,2	15.682,0	13.187,9

Se muestra en la siguiente tabla los porcentajes de cada mes con respecto al promedio anual. Así, para el mes 1 (enero de 2002), operamos de la siguiente forma:

$$Promedio_{ene\,2002} = \frac{10.139,2}{10.901,2} = 93,01$$

y así sucesivamente,...,

$$Promedio_{dic\,2009} = \frac{13.661,5}{13.187,9} = 103,59$$

Obteniendo el índice estacional como el promedio de los porcentajes de cada mes.

	2002	2003	2004	2005	2006	2007	2008	2009	Índice estacional
Enero	93,01	89,86	87,66	85,21	90,09	92,36	95,19	84,11	89,69
Febrero	99,09	97,63	94,45	94,88	98,84	98,26	105,99	94,03	97,90
Marzo	101,47	106,66	107,17	100,69	109,14	107,79	101,27	103,99	104,77
Abril	104,66	106,87	101,28	104,76	92,97	95,21	114,55	100,03	102,54
Mayo	102,29	107,59	104,07	103,99	109,29	105,92	105,99	97,77	104,62
Junio	98,16	101,70	107,20	106,13	107,32	106,52	98,61	105,37	103,88
Julio	102,37	99,53	105,53	100,02	96,05	101,31	109,61	109,76	103,02
Agosto	80,52	71,99	72,85	77,52	79,57	80,18	77,30	76,38	77,04
Septiembre	96,51	101,94	99,16	105,62	99,55	98,54	110,26	105,18	102,10
Octubre	115,68	113,79	106,53	103,28	107,82	110,47	106,31	113,12	109,63
Noviembre	108,83	101,66	112,90	114,04	106,65	109,55	91,12	106,67	106,43
Diciembre	97,41	100,77	101,19	103,86	102,70	93,87	83,80	103,59	98,40

Para obtener la serie de las exportaciones ajustada estacionalmente, esto es descontando el efecto que provoca el ciclo estacional, se dividiría el valor de cada mes por el correspondiente índice estacional y se multiplicaría por 100, obteniendo los siguientes resultados.

Volumen de exportaciones españolas (millones de euros). Años 2002-2009
Series desestacionalizada

	2002	2003	2004	2005	2006	2007	2008	2009
Enero	11.305,1	11.506,5	11.928,9	12.158,4	14.219,4	15.574,7	16.644,7	12.367,8
Febrero	11.034,4	11.452,8	11.775,8	12.402,2	14.292,8	15.179,0	16.978,6	12.667,0
Marzo	10.557,2	11.691,8	12.483,8	12.297,7	14.746,0	15.559,3	15.158,2	13.089,5
Abril	11.126,8	11.969,3	12.054,6	13.072,9	12.834,9	14.042,1	17.518,5	12.865,1
Mayo	10.659,1	11.811,5	12.142,0	12.720,1	14.789,0	15.312,1	15.888,0	12.324,6
Junio	10.301,0	11.244,0	12.596,0	13.074,2	14.625,1	15.508,5	14.886,9	13.377,3
Julio	10.832,1	11.095,4	12.502,5	12.423,9	13.197,8	14.872,0	16.684,4	14.049,9
Agosto	11.393,5	10.732,5	11.541,0	12.877,1	14.621,8	15.739,8	15.734,8	13.074,4
Septiembre	10.304,7	11.467,3	11.854,5	13.238,4	13.803,1	14.597,1	16.935,4	13.586,3
Octubre	11.503,1	11.921,3	11.860,5	12.055,4	13.923,3	15.239,8	15.207,8	13.608,3
Noviembre	11.147,5	10.969,9	12.947,2	13.711,5	14.185,2	15.567,7	13.426,1	13.218,3
Diciembre	10.791,3	11.761,5	12.551,2	13.506,9	14.774,7	14.426,6	13.356,1	13.883,8

7.5.2. Cálculo de la variación estacional por el método del porcentaje promedio móvil

El método del porcentaje del promedio móvil es uno de los métodos más usados para la medición de la variación estacional. Su cálculo es también bastante sencillo:

1. Se obtiene un promedio móvil de doce meses de la serie de datos originales (o de cuatro trimestres si se utilizan los datos trimestrales).
2. Se calcula (con $n = 2$) el promedio móvil de los datos calculados en el paso anterior, con el fin de centrar convenientemente dichos datos, al que se le denomina promedio móvil centrado de doce meses (o de cuatro trimestres).
3. Finalmente se obtiene el índice dividiendo los datos originales por el promedio móvil centrado de doce meses.

Dado el procedimiento señalado, los promedios móviles centrados constituyen en este caso la serie desestacionalizada.

Ejemplo 7.4. Elaborar un índice de estacionalidad por el método del porcentaje promedio móvil para los dos últimos años de la serie del ejemplo anterior.

Los resultados finalmente obtenidos son los siguientes:

Año	Mes	Datos	Media móvil 12 meses	Media móvil centrada	Porcentaje sobre media móvil
2008	Enero	14.928,3			
	Febrero	16.621,4			
	Marzo	15.881,6			
	Abril	17.963,8			
	Mayo	16.621,3			
	Junio	15.464,0			
			15.682,0		
	Julio	17.188,9		15.522,2	110,74
			15.362,4		
	Agosto	12.121,8		15.186,5	79,82
			15.010,6		
	Septiembre	17.290,4		14.920,3	115,89
			14.830,0		
	Octubre	16.671,5		14.631,2	113,95
			14.432,4		
	Noviembre	14.288,9		14.277,1	100,08
			14.121,7		
	Diciembre	13.142,3		14.056,4	93,50
			13.991,1		

(Continúa)

(Continuación)

Año	Mes	Datos	Media móvil 12 meses	Media móvil centrada	Porcentaje sobre media móvil
2009	Enero	11.092,4		13.878,0	79,93
			13.764,9		
	Febrero	12.400,6		13.679,5	90,65
			13.594,1		
	Marzo	13.714,2		13.451,6	101,95
			13.309,1		
	Abril	13.192,1		13.236,1	99,67
			13.163,0		
	Mayo	12.893,4		13.153,8	98,02
			13.144,6		
	Junio	13.895,9		13.166,2	105,54
			13.187,9		
	Julio	14.474,7			
	Agosto	10.072,3			
	Septiembre	13.871,1			
	Octubre	14.918,2			
	Noviembre	14.067,7			
	Diciembre	13.661,5			

Como puede observarse, al utilizarse medias móviles, se produce una pérdida de información correspondiente al número de periodos por año considerados, en nuestro caso, seis en el inicio de la serie y seis en el final (en el caso trimestral dos en el inicio y dos en el final).

7.6. ANÁLISIS DE LAS VARIACIONES CÍCLICAS E IRREGULARES

Aunque los dos componentes anteriores constituyen los fundamentales en el análisis de las series de tiempo, debemos hacer alguna referencia al componente cíclico, fundamental en el análisis de determinadas series temporales de carácter económico, y al componente irregular.

Se entiende por componente o variación cíclica las variaciones regulares que se producen en las series temporales con periodo superior a un año. De hecho una serie temporal puede estar originada por diversos ciclos: un ciclo de medio plazo, otro ciclo de largo plazo, etc.

Un ciclo tiene dos componentes básicos: la amplitud o la distancia que media entre el cero y el máximo valor que alcanza el ciclo, y el periodo o el tiempo que tarda en ocurrir un ciclo completo.

En teoría, cabe entender una serie temporal como una suma de un número indeterminado de ciclos de amplitud y período diferentes, y puede demostrarse que la varianza que muestra en el tiempo una serie temporal se obtiene a partir de la suma de las amplitudes de los diferentes ciclos en que se descompone la serie temporal. El tratamiento de los ciclos es bastante complejo y no vamos a considerarlo en este libro; en los ejercicios que siguen a continuación hemos efectuado, no obstante (Ejercicio 7.11) un ejemplo de algunos de los tratamientos más elementales que suelen realizarse en el análisis de ciclos[14].

Las variaciones irregulares de una serie temporal son movimientos ocasionales que tienen generalmente un carácter impredecible o aleatorio. No obstante, a veces, hay circunstancias comprensibles que determinan dicha irregularidad y que permiten corregir previamente los datos iniciales.

Este tipo de eventos o circunstancias para las que se puede arbitrar una solución son los denominados efectos calendario (festividades, vacaciones, etc...), las huelgas, los paros, las regulaciones de empleo, etc... .La forma más sencilla de compensar estas variaciones es multiplicar la serie original de datos por la siguiente razón: días (laborales y/o turísticos) de un mes en un mes determinado en un promedio de años / días efectivos en el mes dado.

Por lo general, para eliminar las variaciones se suele utilizar una media móvil de pocos periodos (3 ó 5 periodos).

7.7. LA SUAVIZACIÓN EXPONENCIAL

Las técnicas utilizadas hasta ahora pertenecen al ámbito eminentemente descriptivo; tratamos en este apartado un nuevo método denominado *Suavización Exponencial* o *Alisado Exponencial*.

El propósito del *Alisado Exponencial* es eliminar la fluctuación aleatoria. Esto permite captar cualquier «patrón» de conducta que sea evidente en la serie temporal observada, y usar ese patrón para predecir los nuevos valores. Estudiamos aquí las técnicas más importantes de este método.

7.7.1. Suavizado exponencial simple

Cuando la serie presenta un comportamiento estacionario, es decir, no tiene tendencia y puede ser modelizada como:

$$X_t = a + u_t \quad con\ t = 1, 2, \ldots, T$$

[14] Los alumnos interesados en el tratamiento de los ciclos pueden estudiar los siguientes textos:
— ÁLVAREZ VÁZQUEZ, N. (1999): *Introducción a la Econometría*. Ediciones Académicas, Madrid, 2006; temas 3 y 4.
— ÁLVAREZ VÁZQUEZ, , N. J. (2001): *Econometría II*. Ediciones Académicas, Madrid; capítulo 7.

Donde u_t es un término de perturbación aleatorio (fluctuación irregular), con valor esperado cero y varianza constante para todo t, e independiente de X_t para todo t, el método de predicción adecuado es el **Alisado Exponencial Simple** (AES). Este método estima para cada período t el parámetro a como suma ponderada de todas las observaciones anteriores, dando mayor importancia a las observaciones más recientes que a las más antiguas. La ponderación decrece exponencialmente, de ahí el nombre de éste método.

La expresión de cálculo de la estimación del parámetro a en el periodo t es:

$$\hat{a}_t = S_t = aX_t + (1 - \alpha)\, S_{t-1}$$

Donde S_{t-1} es la estimación de a obtenida en el período t–1 y α es la constante de alisado que toma valores entre 0 y 1.

La predicción en el periodo t se obtiene:

$$\hat{X}_t = S_{t-1}$$

Como se observa, el AES actualiza período a período las estimaciones de a incorporando la nueva información.

La elección de la constante de alisado determina las características operativas del AES, ya que la rapidez con que se adaptan las predicciones a los posibles cambios experimentados por el valor de a depende de α. Si α es grande (próximo a 1) el AES se adapta rápidamente a los cambios experimentados en el valor de a y, en consecuencia, deberá escogerse un valor grande de α cuando a es poco estable.

Por el contrario, si la serie es muy estable, el valor de α deberá ser pequeño para conseguir eliminar al máximo las fluctuaciones aleatorias debidas al término de perturbación y conseguir un mejor alisado.

Alternativamente, se puede seleccionar aquel valor de α para el que se obtenga una Raíz del Error Cuadrático Medio $\sqrt{\sum (X_t - S_{t-1})^2 / T}$ menor en la predicción del período muestral (esto es factible hacerlo con el SPSS), si bien en general, parece que un valor de α igual a 0,2 es apropiado en la mayor parte de los casos.

Respecto a la asignación del primer valor S_0, se suelen hacer estos supuestos: cuando la serie tiene muchas oscilaciones se toma $S_0 = X_1$; por el contrario, cuando la serie tiene una cierta estabilidad se hace $S_0 = \bar{X}$.

La predicción del valor de X_t para los períodos T+1, T+2,... en base a los T periodos de la serie observada, será:

$$\hat{X}_{T+1/T} = \hat{X}_{T+2/T\ldots} = S_T$$

Ejemplo 7.5. Se muestra en la siguiente tabla las ventas de una empresa en los últimos 20 años. Obtener la predicción de las ventas para el año 2010, utilizando el AES.

Año	Ventas (miles de euros)	Año	Ventas (miles de euros)
1990	646	2000	641
1991	513	2001	535
1992	656	2002	688
1993	659	2003	525
1994	674	2004	622
1995	693	2005	501
1996	526	2006	659
1997	622	2007	593
1998	652	2008	690
1999	591	2009	632

Como constante de alisado vamos a considerar $\alpha = 0{,}1$ y como inicio de la serie suavizada $S_0 = \overline{X}$, cuyo valor es igual a:

$$\overline{X} = \frac{\sum_{t=1}^{T} X_t}{T} = \frac{12.318}{20} = 615{,}90$$

De esta manera iremos generando la serie suavizada como:

$$S_0 = \overline{X} = 615{,}90$$
$$S_1 = 0{,}2X_1 + 0{,}8S_0 = 0{,}2 \cdot 646 + 0{,}8 \cdot 615{,}90 = 621{,}92$$
$$S_2 = 0{,}2X_2 + 0{,}8S_1 = 0{,}2 \cdot 513 + 0{,}8 \cdot 621{,}92 = 600{,}14$$
$$S_{20} = 0{,}2X_{20} + 0{,}8S_{19} = 0{,}2 \cdot 632 + 0{,}8 \cdot 617{,}09 = 620{,}07$$

Se muestra en la siguiente tabla la serie original, la alisada y el error cometido en la estimación, es decir, $e_t = X_t - S_{t-1}$, en valor absoluto y porcentaje.

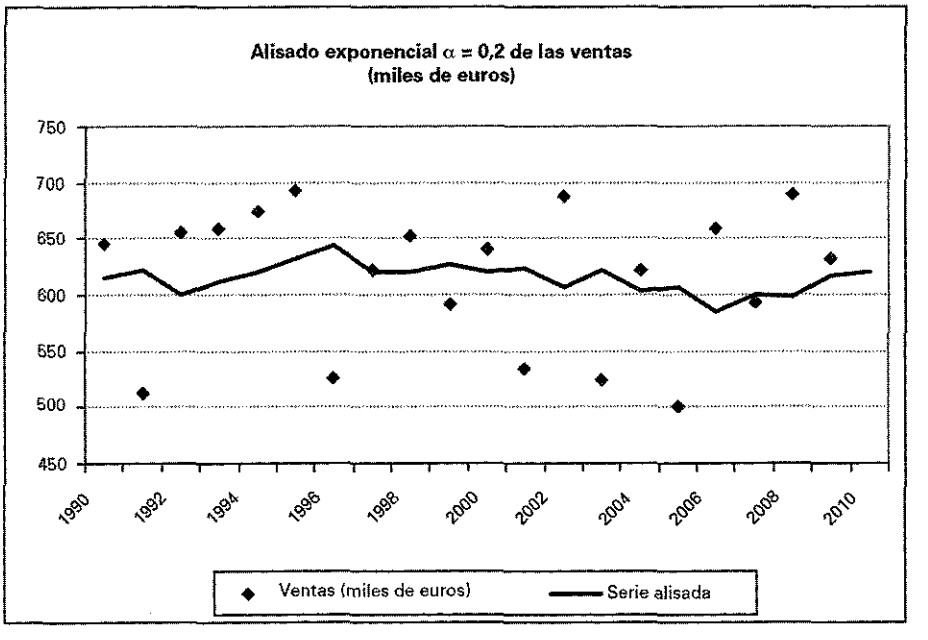
Año		Ventas (miles de euros)		Serie alisada (miles de euros)	Error (miles de euros)	%
1990	X_1	646	S_0	615,90	30,10	4,66
1991	X_2	513	S_1	621,92	–108,92	–21,23
1992	X_3	656	S_2	600,14	55,86	8,52
1993	X_4	659	S_3	611,31	47,69	7,24
1994	X_5	674	S_4	620,85	53,15	7,89
1995	X_6	693	S_5	631,48	61,52	8,88

(Continúa)

(Continuación)

Año		Ventas (miles de euros)		Serie alisada (miles de euros)	Error (miles de euros)	%
1996	X_7	526	S_6	643,78	–117,78	–22,39
1997	X_8	622	S_7	620,23	1,77	0,29
1998	X_9	652	S_8	620,58	31,42	4,82
1999	X_{10}	591	S_9	626,86	–35,86	–6,07
2000	X_{11}	641	S_{10}	619,69	21,31	3,32
2001	X_{12}	535	S_{11}	623,95	–88,95	–16,63
2002	X_{13}	688	S_{12}	606,16	81,84	11,89
2003	X_{14}	525	S_{13}	622,53	–97,53	–18,58
2004	X_{15}	622	S_{14}	603,02	18,98	3,05
2005	X_{16}	501	S_{15}	606,82	–105,82	–21,12
2006	X_{17}	659	S_{16}	585,66	73,34	11,13
2007	X_{18}	593	S_{17}	600,32	–7,32	–1,24
2008	X_{19}	690	S_{18}	598,86	91,14	13,21
2009	X_{20}	632	S_{19}	617,09	14,91	2,36
2010			S_{20}	620,07		

Las ventas predichas para 2010 (y posteriores) son de 620,07 miles de euros. Se representa en el siguiente gráfico la serie original y el alisado realizado.



7.7.2. Suavizado exponencial de Holt

Cuando la serie presenta tendencia lineal, creciente o decreciente, y puede ser modelizada como

$$X_t = a + bt + u_t \quad con\ t = 1, 2, \ldots, T$$

Donde u_t sería la fluctuación irregular, un método de predicción adecuado es el propuesto por Holt (***Suavizado Exponencial de Holt***); este método ha dado muy buenos resultados en la previsión de distintas áreas de la economía empresarial: gestión de stocks, financiación, ventas, etc. El procedimiento se basa en dos ecuaciones de alisado:

$$\hat{a}_t = S_t = \alpha X_t + (1-\alpha)\left(S_{t-1} + \hat{b}_{t-1}\right)$$
$$\hat{b}_t = \beta\left(S_t - S_{t-1}\right) + (1-\beta)\hat{b}_{t-1}$$

La primera de las ecuaciones proporciona una estimación del nivel de la serie en el período t y la segunda permite obtener una estimación de la pendiente de la recta de tendencia para el período t.

Las constantes de alisado α y β toman valores comprendidos entre 0 y 1. Cuanto menores sean estas constantes más alisada será la serie de predicciones. Al igual que en el caso anterior y en el siguiente, el programa informático SPSS permite hacer una búsqueda en rejilla y seleccionar aquellas que presenten la menor Raíz del Error Cuadrático Medio de predicción en los periodos observados de la serie.

Finalmente, la predicción para el periodo t se obtiene a partir de:

$$\hat{X}_t = S_{t-1} + \hat{b}_{t-1}$$

Respecto a los valores iniciales, valores apropiados serían[15]:

$$\hat{b}_0 = \frac{X_T - X_1}{T-1} \qquad S_0 = X_1 - \frac{1}{2}\hat{b}_0$$

Otra opción alternativa sería considerar $S_0 = X_1$ y b_0 la pendiente obtenida del ajuste por mínimos cuadrados de la serie observada, tomando como valores explicativos la variable formada por los subíndices de tiempo, es decir, 1, 2, ..., T.

La predicción para los períodos futuros $T+1$, ...$T+k$, condicionada a los T periodos observados es:

$$\hat{X}_{T+k/T} = S_T + k \cdot \hat{b}_T$$

[15] Estos valores iniciales son los que considera el programa SPSS.

Ejemplo 7.6. Utilizando los datos del ejercicio 7.1, realizar un alisado exponencial de Holt para predecir la evolución del coste laboral en 2009, suponiendo $\alpha = \beta = 0{,}1$.

Para calcular los valores iniciales de las ecuaciones de alisado, tomamos la opción considerada en el SPSS, es decir:

$$b_0 = \frac{2.583{,}82 - 2.063{,}93}{15} = 34{,}66 \qquad S_0 = 2.063{,}93 - \frac{34{,}66}{2} = 2.046{,}60$$

Siendo por tanto la predicción de la serie suavizada para el primer periodo:

$$\hat{X}_1 = S_0 + b_0 = 2.046{,}60 + 34{,}66 = 2.081{,}26$$

Para el segundo periodo se opera de la siguiente manera:

$$S_1 = \alpha X_1 + (1-\alpha)(S_0 + b_0) = \alpha X_1 + (1-\alpha)\hat{X}_1 = 0{,}1 \cdot 2.063{,}93 + 0{,}9 \cdot 2.081{,}26 = 2.079{,}53$$

$$b_1 = \beta(S_1 - S_0) + (1-\beta)b_0 = 0{,}1 \cdot (2.079{,}53 - 2.046{,}60) + 0{,}9 \cdot 34{,}66 = 34{,}49$$

$$\hat{X}_2 = S_1 + b_1 = 2.079{,}53 + 34{,}49 = 2.114{,}01$$

Y así sucesivamente. Los resultados obtenidos para todos los periodos son los siguientes.

Evolución del coste total por trabajador (euros) en España, periodo 2005-2008

Año	Trimestre	Euros		Nivel		Pendiente		Serie alisada	Error	%
2005	1	2.063,93	S_0	2.046,60	$\hat{b}_0$	34,66	$S_0+\hat{b}_0$	2.081,26	–17,33	–0,84
2005	2	2.141,24	S_1	2.079,53	$\hat{b}_1$	34,49	$S_1+\hat{b}_1$	2.114,01	27,23	1,27
2005	3	2.055,75	S_2	2.116,74	$\hat{b}_2$	34,76	$S_2+\hat{b}_2$	2.151,49	–95,74	–4,66
2005	4	2.251,93	S_3	2.141,92	$\hat{b}_3$	33,80	$S_3+\hat{b}_3$	2.175,72	76,21	3,38
2006	1	2.154,32	S_4	2.183,34	$\hat{b}_4$	34,56	$S_4+\hat{b}_4$	2.217,90	–63,58	–2,95
2006	2	2.254,38	S_5	2.211,55	$\hat{b}_5$	33,93	$S_5+\hat{b}_5$	2.245,47	8,91	0,40
2006	3	2.152,88	S_6	2.246,36	$\hat{b}_6$	34,02	$S_6+\hat{b}_6$	2.280,38	–127,50	–5,92
2006	4	2.358,85	S_7	2.267,63	$\hat{b}_7$	32,74	$S_7+\hat{b}_7$	2.300,37	58,48	2,48
2007	1	2.239,53	S_8	2.306,22	$\hat{b}_8$	33,33	S_8+b_8	2.339,54	–100,01	–4,47
2007	2	2.339,64	S_9	2.329,54	$\hat{b}_9$	32,33	$S_9+\hat{b}_9$	2.361,87	–22,23	–0,95

(Continúa)

(Continúa)

Año	Trimestre	Euros		Nivel		Pendiente		Serie alisada	Error	%
2007	3	2.242,03	S_{10}	2.359,65	$\hat{b}_{10}$	32,10	$S_{10}+\hat{b}_{10}$	2.391,75	–149,72	–6,68
2007	4	2.459,71	S_{11}	2.376,78	$\hat{b}_{11}$	30,61	$S_{11}+\hat{b}_{11}$	2.407,38	52,33	2,13
2008	1	2.342,28	S_{12}	2.412,62	$\hat{b}_{12}$	31,13	$S_{12}+\hat{b}_{12}$	2.443,75	–101,47	–4,33
2008	2	2.451,40	S_{13}	2.433,60	$\hat{b}_{13}$	30,11	$S_{13}+\hat{b}_{13}$	2.463,71	–12,31	–0,50
2008	3	2.350,17	S_{14}	2.462,48	$\hat{b}_{14}$	29,99	$S_0+\hat{b}_{14}$	2.492,48	–142,31	–6,06
2008	4	2.583,82	S_{15}	2.478,24	$\hat{b}_{15}$	28,57	$S_0+\hat{b}_{15}$	2.506,81	77,01	2,98
			S_{16}	2.514,51	$\hat{b}_{16}$	29,34	$S_0+\hat{b}_{16}$	2.543,85		

Las predicciones para el año 2009 se calculan del siguiente modo:

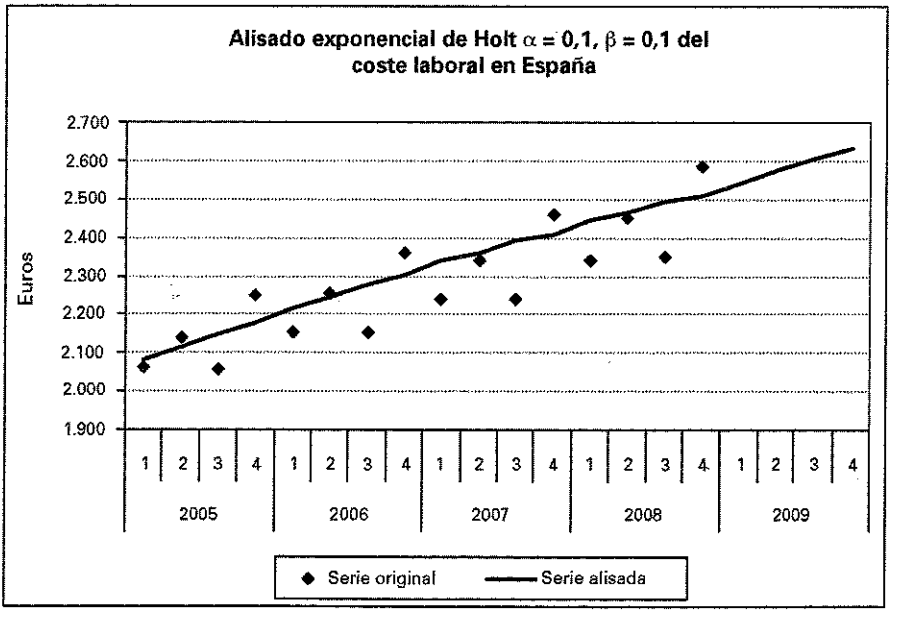
Primer trimestre de 2009: $\hat{X}_{T+1/T} = S_T + \hat{b}_T$ == 2.514,51 + 29,34 = 2.543,85

Segundo trimestre de 2009: $\hat{X}_{T+2/T} = S_T + 2\hat{b}_T = 2.514,51 + 2 \cdot 29,34 = 2.573,19$

Tercer trimestre de 2009: $\hat{X}_{T+3/T} = S_T + 3\hat{b}_T = 2.514,51 + 3 \cdot 29,34 = 2.602,53$

Cuarto trimestre de 2009: $\hat{X}_{T+4/T} = S_T + 4\hat{b}_T = 2.514,51 + 4 \cdot 29,34 = 2.631,87$

Por último, el gráfico con la serie original y la alisada sería la siguiente.



Tal como se observa en el gráfico, la serie presenta un claro patrón estacional, con valores inferiores en el primer y cuarto trimestre, debido seguramente a la generación de un empleo de carácter estacional menos remunerado en los periodos correspondientes a las fiestas de Navidad y a las vacaciones de verano, generando en consecuencia errores de predicción más elevados, en especial en el tercer trimestre.

En el siguiente apartado, estudiaremos como tener en cuenta las variaciones estacionales en las ecuaciones del alisado exponencial y mejorar así el modelo obteniendo predicciones más adecuadas.

7.7.3. Suavizado exponencial de Winters

Una serie con tendencia lineal y patrón estacional multiplicativo puede modelizarse como:

$$X_t = (a + b_t)\, c_t + u_t \quad con\, t = 1, 2, \ldots, T$$

Donde c_t es el índice estacional correspondiente al período t. Las estimaciones de a_t, b_t y c_t vienen dadas por:

$$\hat{a}_t = S_t = \alpha \frac{X_t}{\hat{c}_{t-S}} + (1-\alpha)\left(S_{t-1} + \hat{b}_{t-1}\right)$$

$$\hat{b}_t = \beta(S_t - S_{t-1}) + (1-\beta)\hat{b}_{t-1}$$

$$\hat{c}_t = \gamma \frac{X_t}{S_t} + (1-\gamma)\hat{c}_{t-S}$$

Donde S es la periodicidad de la serie.

Las constantes de alisado α, β y γ deben satisfacer únicamente la condición de tomar valores comprendidos entre 0 y 1.

La predicción para el periodo t se obtiene a partir de:

$$\hat{X}_t = \left(S_{t-1} + \hat{b}_{t-1}\right)\hat{c}_{t-S}$$

Respecto a los valores iniciales, señalamos como valores apropiados los que toma como referencia el SPSS, los cuales son:

$$\hat{b}_0 = \frac{\overline{X}_n - \overline{X}_1}{(n-1)S} \qquad S_0 = \overline{X}_1 - \frac{S}{2}\hat{b}_0$$

Siendo n el número de años de la serie.

Los factores estacionales iniciales pueden ser obtenidos a partir de cualquiera de los dos métodos estudiados en el apartado correspondiente al estudio de las variaciones cíclicas o de cualquier otro.

La predicción para los períodos futuros $T+1$, ..., $T+k$ obtenida en el período T es:

$$\hat{X}_{T+k/T} = \left(S_T + k \cdot \hat{b}_T\right)\hat{c}_{T+k-S}$$

Ejemplo 7.7. Utilizando los datos del ejercicio 7.1, realizar un alisado exponencial de Winters para predecir la evolución del coste laboral en 2009, suponiendo $\alpha = \beta = \gamma = 0{,}1$ y utilizando el método del porcentaje promedio para la obtención de los factores estacionales iniciales.

Calculamos en primer lugar el coste laboral medio anual para cada uno de los años considerados. Así:

$$Promedio_{2005} = \overline{X}_1 = \frac{\sum_{t=1}^{4} X_t}{4} = \frac{8.512{,}85}{4} = 2.128{,}21$$

$$Promedio_{2006} = \overline{X}_2 = \frac{\sum_{t=5}^{8} X_t}{4} = \frac{8.920{,}43}{4} = 2.230{,}11$$

$$Promedio_{2007} = \overline{X}_3 = \frac{\sum_{t=9}^{12} X_t}{4} = \frac{9.280{,}91}{4} = 2.320{,}23$$

$$Promedio_{2008} = \overline{X}_4 = \frac{\sum_{t=13}^{16} X_t}{4} = \frac{9.727{,}67}{4} = 2.431{,}92$$

Se muestran en la siguiente tabla los porcentajes de cada trimestre respecto a su media anual, y el factor estacional inicial considerado, calculado como media de todos los porcentajes de cada trimestre en los distintos años considerados.

	2005	**2006**	**2007**	**2008**	**Índice estacional**
Trimestre 1	96,980	96,602	96,522	96,314	96,604
Trimestre 2	100,612	101,088	100,837	100,801	100,835
Trimestre 3	96,595	96,537	96,630	96,639	96,600
Trimestre 4	105,813	105,773	106,012	106,246	105,961

Luego consideramos:

$$\hat{c}_{-3} = \frac{96{,}604}{100} \qquad \hat{c}_{-2} = \frac{100{,}835}{100} \qquad \hat{c}_{-1} = \frac{96{,}600}{100} \qquad \hat{c}_{0} = \frac{105{,}961}{100}$$

Considerando como estimaciones iniciales las del SPSS, obtenemos para el nivel y la pendiente:

$$\hat{b}_0 = \frac{\overline{X}_n - \overline{X}_1}{(n-1)S} = \frac{\overline{X}_4 - \overline{X}_1}{(4-1)\cdot 4} = \frac{2.431{,}92 - 2.128{,}21}{12} = 25{,}31$$

$$S_0 = \overline{X}_1 - \frac{S}{2}\hat{b}_0 = 2.128{,}21 - 2 \cdot 25{,}31 = 2.077{,}60$$

Una vez calculados los valores iniciales ya estamos en condiciones de iniciar el algoritmo. Así, la primera predicción de la serie sería:

$$\hat{X}_1 = \left(S_0 + \hat{b}_0\right)\hat{c}_{-3} = (2.077{,}60 + 25{,}31) \cdot \frac{96{,}604}{100} = 2.031{,}50$$

Para el segundo periodo:

$$\hat{a}_1 = S_1 = \alpha \frac{X_1}{\hat{c}_{-3}} + (1-\alpha)\left(S_0 + \hat{b}_0\right) = 0{,}1 \cdot \frac{2.063{,}93}{0{,}96604} + 0{,}9 \cdot (2.077{,}60 + 25{,}31) = 2.106{,}26$$

$$\hat{b}_1 = \beta\left(S_1 - S_0\right) + (1-\beta)\hat{b}_0 = 0{,}1 \cdot (2.106{,}26 - 2.077{,}60) + 0{,}9 \cdot 25{,}31 = 25{,}64$$

$$\hat{c}_1 = \gamma \frac{X_1}{S_1} + (1-\gamma)\hat{c}_{-3} = 0{,}1 \cdot \frac{2.063{,}93}{2.106{,}26} + 0{,}9 \cdot 0{,}96604 = 0{,}96743$$

Obteniendo por tanto la siguiente predicción.

$$\hat{X}_2 = \left(S_1 + \hat{b}_1\right)\hat{c}_{-2} = (2.106{,}26 + 25{,}64) \cdot \frac{100{,}835}{100} = 2.149{,}70$$

Para el tercer periodo:

$$\hat{\alpha}_2 = s_2 = \alpha \frac{x_2}{\hat{c}_2} + (1-\alpha)\left(S_1 + \hat{b}_1\right) = 0{,}1 \cdot \frac{2.141{,}24}{1{,}00835} + 0{,}9 \cdot (2.106 + 25{,}64) = 2.131{,}07$$

$$\hat{b}_2 = \beta\left(S_2 - S_1\right) + (1-\beta)\hat{b}_1 = 0{,}1 \cdot (2.131{,}07 - 2.106{,}26) + 0{,}9 \cdot 25{,}64 = 25{,}56$$

$$\hat{c}_2 = \gamma \frac{X_2}{S_2} + (1-\gamma)\hat{c}_{-2} = 0{,}1 \cdot \frac{2.141{,}24}{2.131{,}07} + 0{,}9 \cdot 1{,}00835 = 1{,}00799$$

Y la predicción toma el valor.

$$\hat{X}_3 = \left(S_2 + \hat{b}_2\right)\hat{c}_{-1} = (2.131{,}07 + 25{,}56)\cdot 0{,}96600 = 2.083{,}30$$

Operando hasta el final obtenemos la siguiente tabla:

Periodo	Euros			Nivel		Pendiente		Factor estac. × 100	Serie alisada	Error	%
							$\hat{c}_{-3}$	96,60			
							$\hat{c}_{-2}$	100,83			
							$\hat{c}_{-1}$	96,60			
2005-1	2.063,93		S_0	2077,60	$\hat{b}_0$	25,31	$\hat{c}_0$	105,96	2031,50	32,43	1,57
2005-2	2.141,24		S_1	2106,26	$\hat{b}_1$	25,64	$\hat{c}_1$	96,74	2149,70	-8,46	-0,40
2005-3	2.055,75		S_2	2131,07	$\hat{b}_2$	25,56	$\hat{c}_2$	100,80	2083,30	-27,55	-1,34
2005-4	2.251,93		S_3	2153,78	$\hat{b}_3$	25,28	$\hat{c}_3$	96,48	2308,94	-57,01	-2,53
2006-1	2.154,32		S_4	2173,67	$\hat{b}_4$	24,74	$\hat{c}_4$	105,72	2126,80	27,52	1,28
2006-2	2.254,38		S_5	2201,25	$\hat{b}_5$	25,02	$\hat{c}_5$	96,86	2244,06	10,32	0,46
2006-3	2.152,88		S_6	2227,30	$\hat{b}_6$	25,12	$\hat{c}_6$	100,84	2173,25	-20,37	-0,95
2006-4	2.358,85		S_7	2250,31	$\hat{b}_7$	24,91	$\hat{c}_7$	96,40	2405,48	-46,63	-1,98
2007-1	2.239,53		S_8	2270,81	$\hat{b}_8$	24,47	$\hat{c}_8$	105,54	2223,11	16,42	0,73
2007-2	2.339,64		S_9	2296,98	$\hat{b}_9$	24,64	$\hat{c}_9$	96,92	2341,14	-1,50	-0,06
2007-3	2.242,03		S_{10}	2321,47	$\hat{b}_{10}$	24,63	$\hat{c}_{10}$	100,83	2261,72	-19,69	-0,88
2007-4	2.459,71		S_{11}	2344,06	$\hat{b}_{11}$	24,42	$\hat{c}_{11}$	96,33	2499,70	-39,99	-1,63
2008-1	2.342,28		S_{12}	2364,69	$\hat{b}_{12}$	24,04	$\hat{c}_{12}$	105,39	2315,16	27,12	1,16
2008-2	2.451,40		S_{13}	2391,53	$\hat{b}_{13}$	24,32	$\hat{c}_{13}$	97,02	2436,02	15,38	0,63
2008-3	2.350,17		S_{14}	2417,38	$\hat{b}_{14}$	24,48	$\hat{c}_{14}$	100,89	2352,19	-2,02	-0,09
2008-4	2.583,82		S_{15}	2441,65	$\hat{b}_{15}$	24,45	$\hat{c}_{15}$	96,32	2598,97	-15,15	-0,59
			S_{16}	2464,66	$\hat{b}_{16}$	24,31	$\hat{c}_{16}$	105,33	2414,85		

Las predicciones para el año 2009 se calculan del siguiente modo:

Primer trimestre de 2009:

$$\hat{X}_{T+1/T} = \left(S_T + \hat{b}_T\right)\hat{c}_{T-3} = (2.464{,}66 + 2\cdot 24{,}31)\cdot \frac{97{,}02}{100} \approx 2.414{,}85$$

Segundo trimestre de 2009:

$$\hat{X}_{T+2/T} = \left(S_T + 2\hat{b}_T\right)\hat{c}_{T-2} = (2.464{,}66 + 2\cdot 24{,}31)\cdot \frac{100{,}89}{100} = 2.535{,}71$$

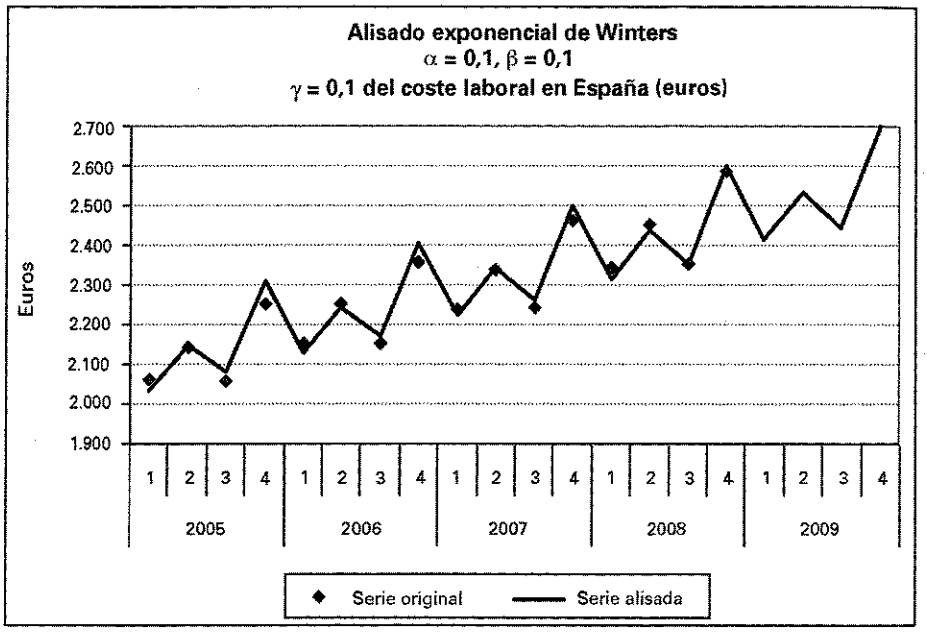
Tercer trimestre de 2009:

$$\hat{X}_{T+3/T} = (S_T + 3b_{1T})\hat{c}_{T-1} = (2.464{,}66 + 3 \cdot 24{,}31) \cdot \frac{96{,}32}{100} = 2.444{,}23$$

Cuarto trimestre de 2009:

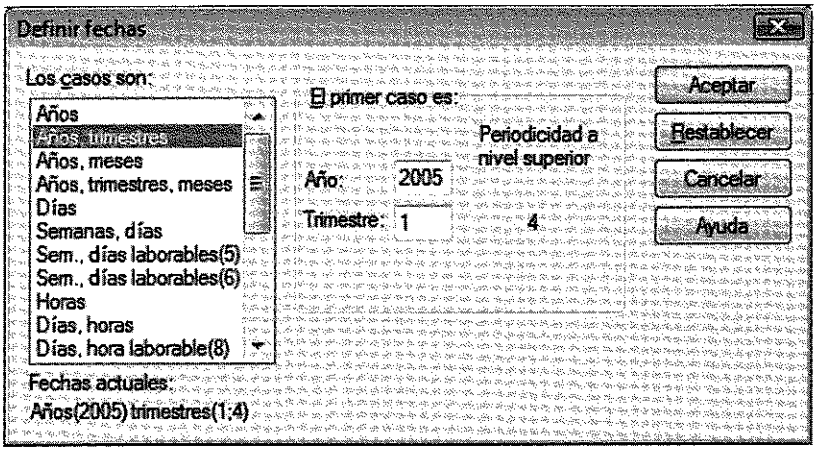
$$\hat{X}_{T+4/T} = (S_T + 4b_{1T})\hat{c}_T = (2.464{,}66 + 4 \cdot 24{,}31) \cdot \frac{105{,}33}{100} = 2.698.53$$

Por último, se representa en el gráfico siguiente la serie original y la alisada. Como vemos, el alisado reproduce perfectamente la serie observada, con lo cual podemos considerar como buenas las predicciones obtenidas para 2009.



7.8. SUAVIZACIÓN EXPONENCIAL DE SERIES TEMPORALES CON SPSS

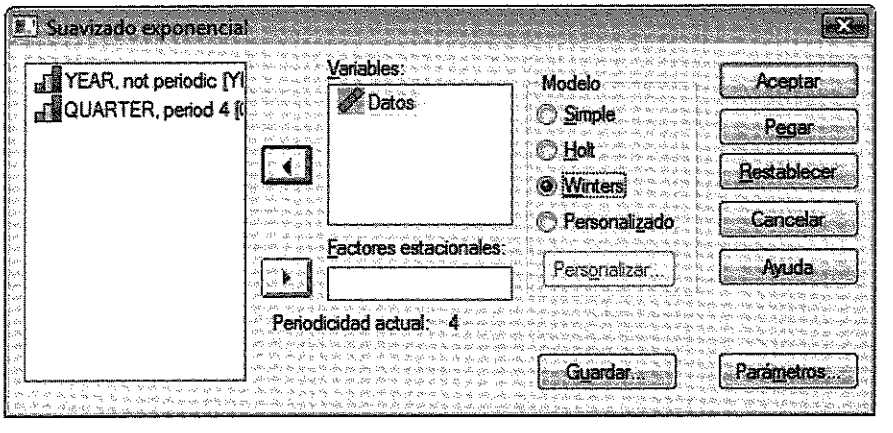
El análisis de series temporales en SPSS requiere la definición previa de variables fechas, lo cual está accesible en el menú Datos -> Definir fechas. La pantalla asociada a este menú es la siguiente:



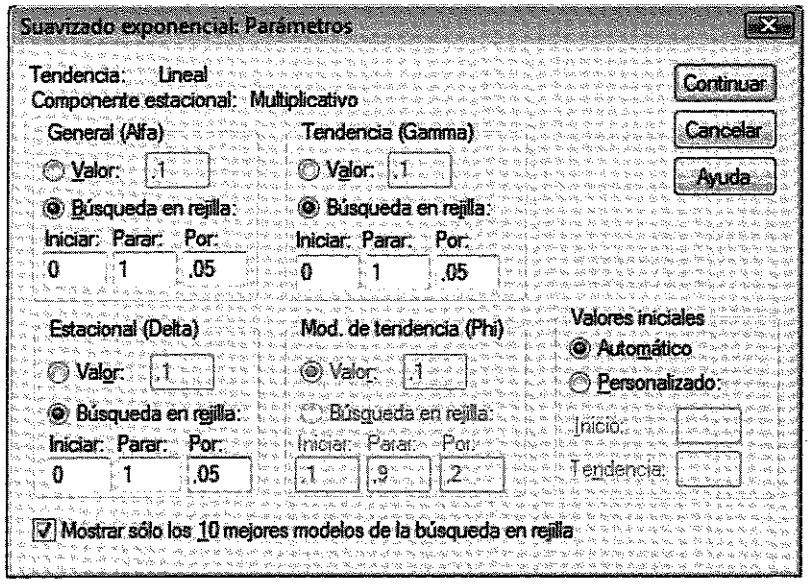
En ella especificamos el tipo de periodicidad de los datos, así como el año y el mes de inicio de la serie, en nuestro caso consideramos la serie ajustada en el ejercicio anterior.

Una vez hecho esto, para obtener predicciones mediante el AES la secuencia en SPSS es:

Analizar → Series Temporales → Suavizado exponencial



En el cuadro de diálogo *Suavizado exponencial* está activado por defecto el método AES que corresponde al Modelo *Simple*. Señalamos la opción correspondiente al modelo de Winters que, como se ha visto, es el más indicado para esta serie temporal. Se indica la serie que se quiere predecir en la ventana Variables. Con el botón Parámetros se abre el siguiente cuadro de diálogo en el que se puede modificar la elección del valor de las constantes de alisado:



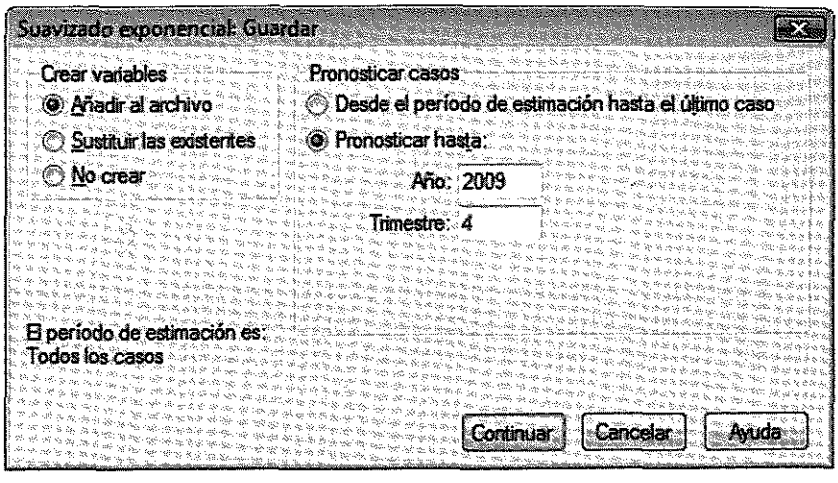
Por defecto está activada para *Alfa* la opción *Valor*: 0,1 el la primera y 0,2 en resto. Con la opción Búsqueda en rejilla se puede especificar un intervalo de valores para cada una de las constantes de alisado, Alfa, Gama y Delta, entre los cuales el sistema determinará aquellos que optimicen la predicción, seleccionando aquellos que proporcionen menor suma de cuadrados de los errores de estimación (SSE). La búsqueda puede afinarse tanto como se quiera modificando en la casilla Por el valor de incremento. Si se mantiene seleccionada la opción *Mostrar sólo los 10 mejores modelos de la búsqueda en rejilla* se muestran sólo en el visor de resultados los valores de los parámetros y las SSE para los 10 valores de α con menor error de estimación. En caso contrario, aparecen todos los modelos probados.

El valor inicial del alisado, por defecto, lo calcula el sistema automáticamente (tal como se ha comentado anteriormente) a partir de la serie observada. Con la opción *Valores iniciales*: *Personalizado* se puede fijar el valor de inicio que se desee.

El botón *Guardar,* del cuadro de diálogo *Suavizado Exponencial*, permite modificar las opciones relacionadas con la creación de nuevas variables y con la predicción:

Con respecto a la creación de nuevas variables con los resultados del alisado, la opción activada por defecto es *Añadir al archivo* con la que se añaden a la base de datos activa la serie alisada y los errores de predicción. La opción *Sustituir las existentes* guarda sólo las variables del último procedimiento sustituyéndolas si en el archivo activo ya existían. Por último, la opción *No crear* indica que no se desea guardar ninguna variable de resultados.

Con respecto a las predicciones, en *Pronosticar casos*, por defecto, está activada la opción Desde el período de estimación hasta el último caso que proporciona los valores alisados (S_t) únicamente para los períodos observados. La opción *Pronosticar hasta* permite generar predicciones para valores futuros de la variable hasta el período indicado en *Observación*.



El resultado final obtenido sería el siguiente:

Estado de suavizado inicial

		Datos
Índices estacionales	1	98,14035
	2	101,55584
	3	96,09396
	4	104,20984
Nivel		2077,595
Tendencia		25,30875

Sumas menores de los errores cuadráticos

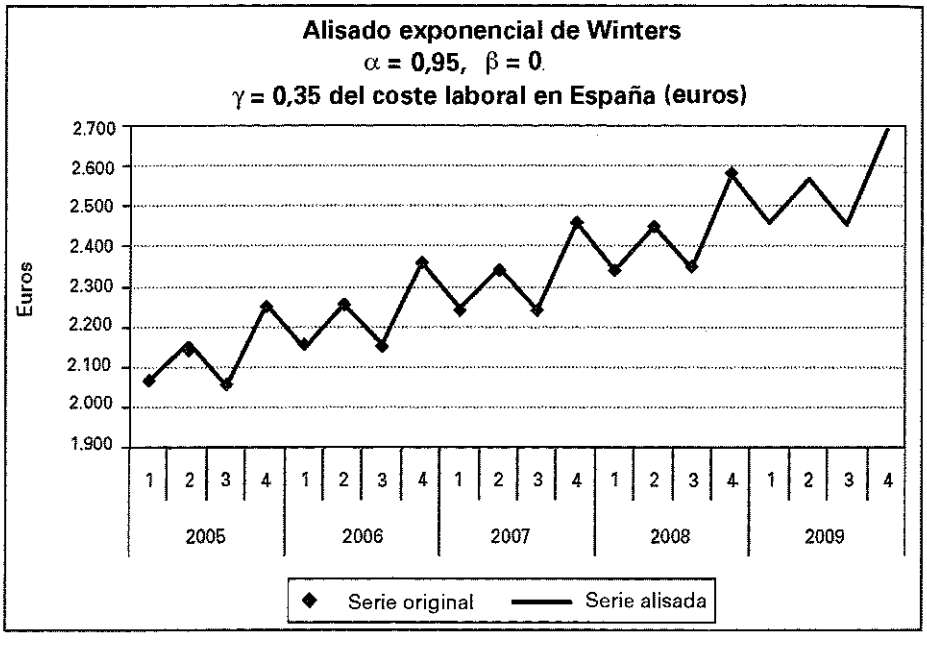
Serie	Rango del modelo	Alpha (nivel)	Gamma (tendencia)	Delta (estación)	Sumas de los errores cuadráticos
Datos	1	,95000	,0000	,35000	749,71594
	2	,95000	,0000	,40000	749,72489
	3	,95000	,0000	.30000	749,75722
	4	,95000	,0000	.45000	749,78300
	5	,95000	,0000	.25000	749,84982
	6	,95000	,0000	,50000	749,88923
	7	,95000	,0000	,20000	749,99485
	8	,95000	,0000	,55000	750,04256
	9	,95000	,0000	,15000	750,19346
	10	,95000	,0000	,60000	750,24199

Parámetros del suavizado

Serie	Alpha (nivel)	Gamma (tendencia)	Delta (estación)	Sumas de los errores cuadráticos	gl error
Datos	,95000	,0000	,35000	749,71594	11

A continuación, se muestran los parámetros con las sumas menores de errores cuadráticos. Estos parámetros se utilizan para pronosticar.

Se muestra en el siguiente gráfico la serie temporal y la suavizada obtenida.



7.9. EJERCICIOS

Sobre series temporales

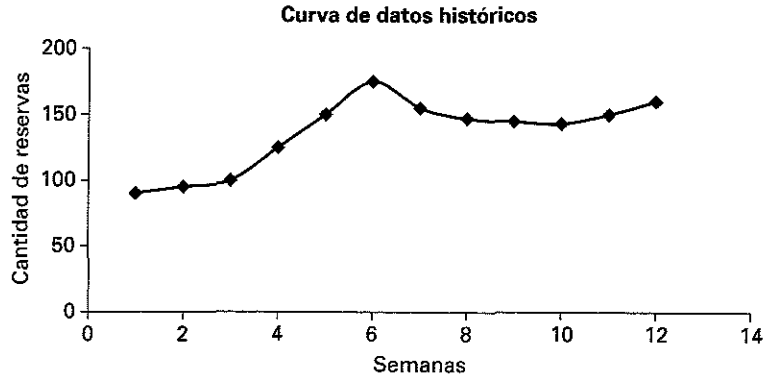
Ejercicio 7.1. *La siguiente serie muestra la cantidad de reservas efectuadas por una agencia de viajes a un destino particular, durante el verano del año 2001.*

Meses	Semanas	Períodos	Reservas
Julio	1	1	90
	2	2	95
	3	3	100
	4	4	125
Agosto	1	5	150
	2	6	175
	3	7	155
	4	8	147
Septiembre	1	9	145
	2	10	143
	3	11	150
	4	12	160

Utilice los diferentes métodos estudiados durante el curso para realizar los gráficos a partir de una misma serie de datos históricos de un determinado periodo.

Respuesta

NOTA: De la simple observación de los datos, se desprende cuál es el método que representa mejor el comportamiento de esta serie, un ajuste a partir de un polinomio de grado 4. Más adelante se verán problemas que analizan varios de esos métodos para efectuar pronósticos con el mayor grado de aproximación a la realidad posible, identificándose cuál de ellos resulta más apropiado en cada caso.


Curva de datos históricos
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas

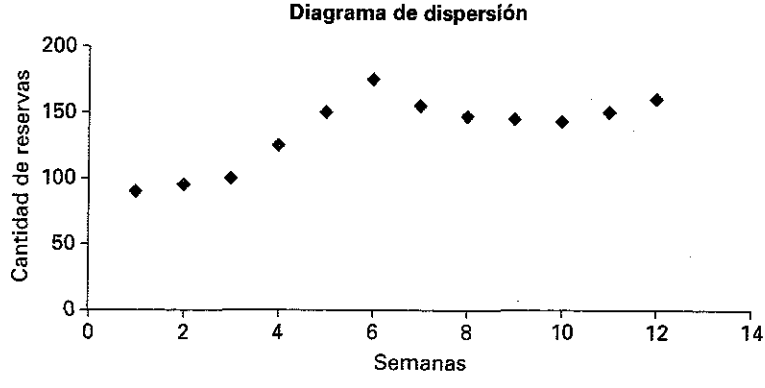
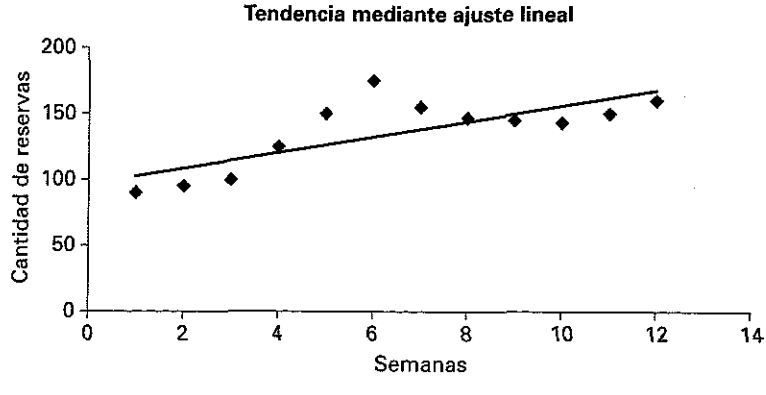
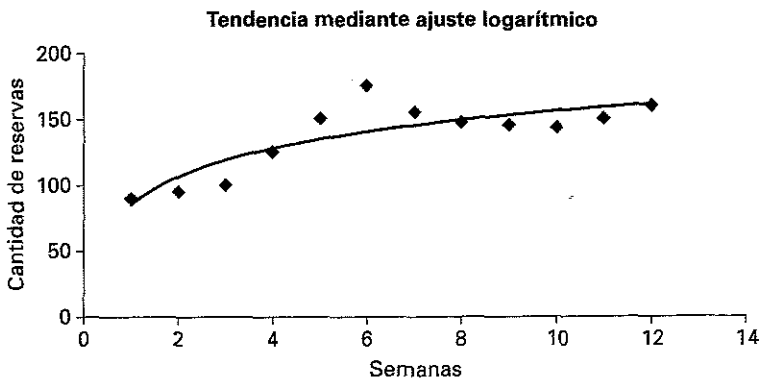
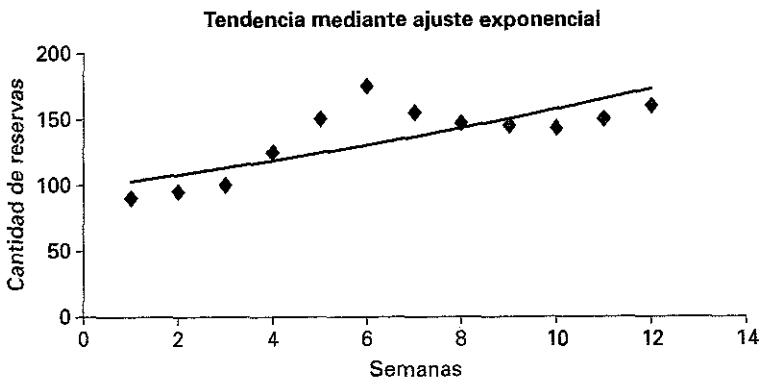
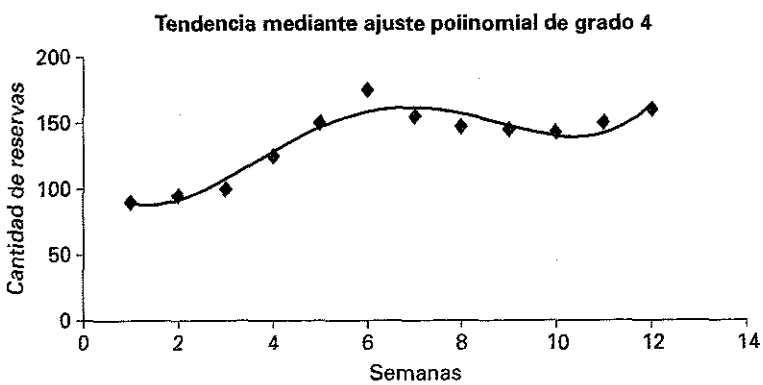


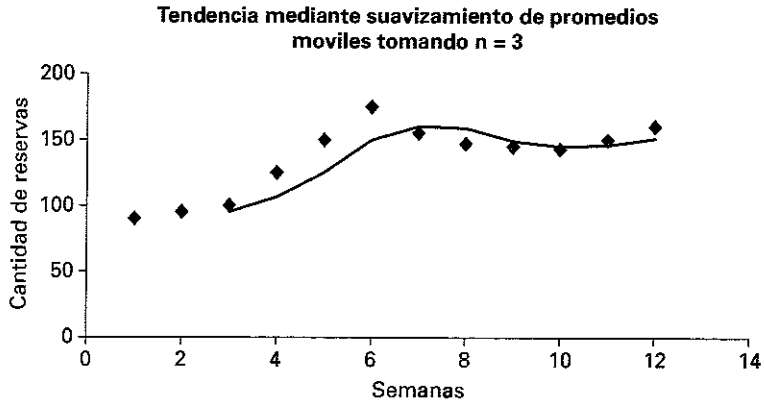
Diagrama de dispersión
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas


Tendencia mediante ajuste lineal
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas


Tendencia mediante ajuste logarítmico
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas


Tendencia mediante ajuste exponencial
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas


Tendencia mediante ajuste polinomial de grado 4
Cantidad de reservas
200
150
100
50
0
0
2
4
6
8
10
12
14
Semanas



Ejercicio 7.2. *Use el método de los semipromedios para determinar la tendencia lineal de los datos de devoluciones de pedidos de una determinada empresa. Tome como promedio la media aritmética. Represente gráficamente el diagrama de dispersión y la recta de tendencia.*

Los datos de devoluciones anuales son los siguientes:

Año	Devoluciones
1993	94
1994	92
1995	86
1996	82
1997	78
1998	81
1999	75
2000	72
2001	70
2002	68
2003	70

Respuesta

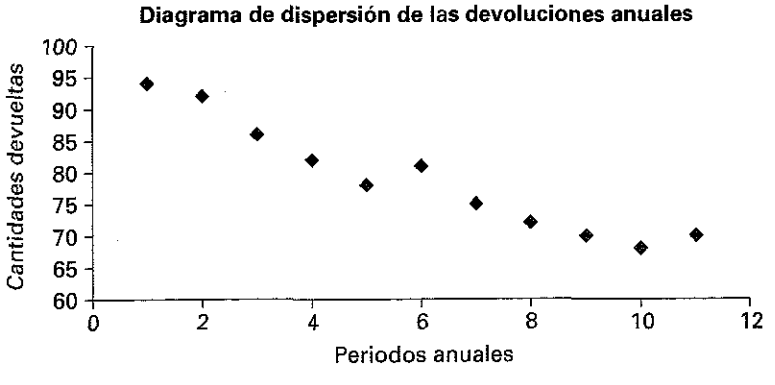
Debe dividirse la serie en dos partes iguales y se calcula la media de cada parte por separado. Para simplificar los cálculos, se omite el año central (1998). Con

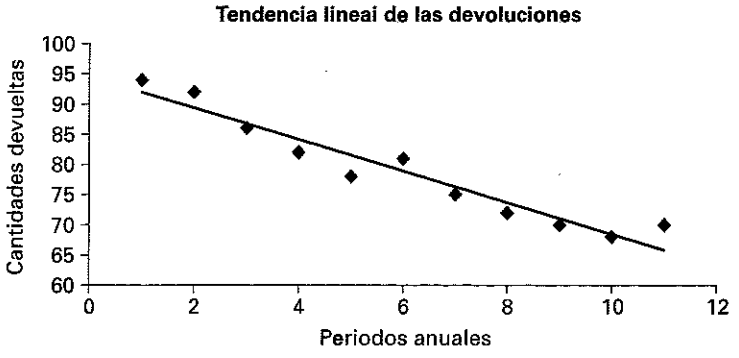
estos datos puede calcularse la pendiente y ordenada al origen de una recta tendencial. Por último, se utiliza la ecuación hallada para determinar los valores de tendencia.

Año	Periodos	Datos	Semipromedios	Valores Ajustados
1993	1	94		**91,53**
1994	2	92		**88,97**
1995	3	86	86,40	**86,40**
1996	4	82		**83,83**
1997	5	78		**81,27**
1998	6	81		**78,70**
1999	7	75		**76,13**
2000	8	72		**73,57**
2001	9	70	71,00	**71,00**
2002	10	68		**68,43**
2003	11	70		**65,87**

De donde:

$$b = \frac{71 - 86{,}4}{9 - 3} = -2{,}567\,; \qquad a = 71 - (-2{,}567)\cdot 9 = 94{,}1$$





Ejercicio 7.3. *El siguiente cuadro muestra la evolución del coste de los insumos de una empresa durante las últimas 10 temporadas. Utilice el método de los semipromedios para determinar la tendencia lineal de dichos costes. Tome como promedio la mediana. Represente gráficamente el diagrama de dispersión y la recta de tendencia.*

Los datos de costes históricos anuales (en miles de euros) son los siguientes:

Periodos	Datos
1	152
2	155
3	160
4	170
5	165
6	163
7	173
8	190
9	202
10	200
11	205

Respuesta

Se procede de manera análoga al ejercicio anterior. Debemos tener en cuenta que para calcular la mediana, previamente es necesario ordenar los datos.

La columna de Datos ordenados es:

Datos 152 155 160 163 165 170 173 190 200 202 205

Luego las medianas a considerar son la tercera y la novena observación, es decir, 160 y 200.

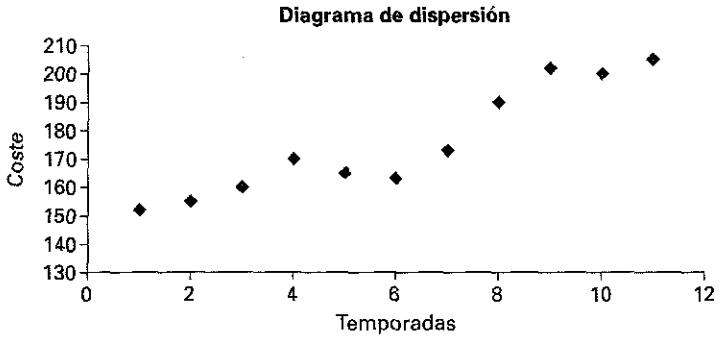
Año	Periodos	Datos	Datos Ordenados	Semipromedios	Valores ajustados
1993	1	152	152		**146,67**
1994	2	155	155		**153,33**
1995	3	160	160	160,00	**160,00**
1996	4	170	165		**166,67**
1997	5	165	170		**173,33**
1998	6	163	163		**180,00**
1999	7	173	173		**186,67**
2000	8	190	190		**193,33**
2001	9	202	200	200,00	**200,00**
2002	10	200	202		**206,67**
2003	11	205	205		**213,33**

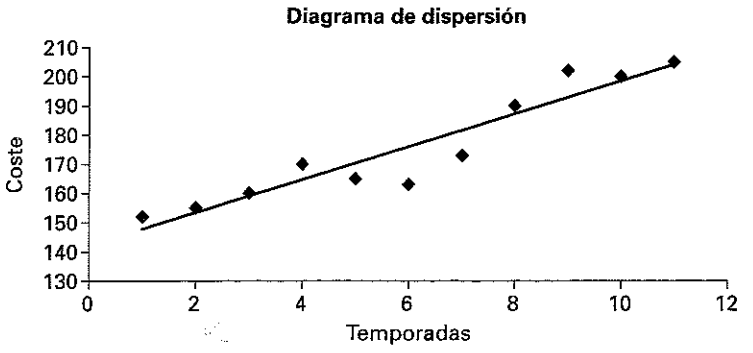
De donde:

$$b = \frac{200-160}{9-3} = 6{,}667\,; \qquad a = 200 - 6{,}667 \cdot 9 = 140$$

Y la ecuación de la recta de tendencia será

$$\hat{y} = 140 + 6{,}6\hat{7}x$$





Ejercicio 7.4. *Se detalla a continuación la evolución anual de los niveles de ocupación (en miles de personas) correspondientes al periodo 1998 a 2002 para la quincena de menor afluencia turística de la temporada.*

Años	Ocupación
1988	4,1
1989	4,5
1990	4,3
1991	4,5
1992	4,7
1993	4,9
1994	5,2
1995	5
1996	5,1
1997	5,16
1998	5,2
1999	5,25
2000	5,12
2001	4,95
2002	4

A partir de dichos datos:

a) *Estime, mediante ajuste lineal por mínimos cuadrados, el nivel de ocupación para la quincena en estudio correspondiente al año 2003.*

b) *Determine el porcentaje de la variación en los niveles de ocupación que se encuentra explicado por el modelo resultante. Interprete el resultado obtenido.*

c) *Represente gráficamente el resultado obtenido.*

Respuesta

a) Podemos redefinir el tiempo para simplificar los cálculos, expresando los mismos como periodos sucesivos y no como años.

Periodos (t)	Ocupación (y)	$(y-\bar{y})$	$(t-\bar{t})$	$(y-\bar{y})\cdot(t-\bar{t})$	$(y-\bar{y})^2$	$(t-\bar{t})^2$
1	4,1	–0,699	–7	–0,699	0,4881	49
2	4,5	–0,299	–6	–0,598	0,0892	36
3	4,3	–0,499	–5	–1,497	0,2487	25
4	4,5	–0,299	–4	–1,196	0,0892	16
5	4,7	–0,099	–3	–0,495	0,0097	9
6	4,9	0,101	–2	0,606	0,0103	4
7	5,2	0,401	–1	2,807	0,1611	1
8	5	0,201	0	1,608	0,0405	0
9	5,1	0,301	1	2,709	0,0908	1
10	5,16	0,361	2	3,61	0,1306	4
11	5,2	0,401	3	4,411	0,1611	9
12	5,25	0,451	4	5,412	0,2037	16
13	5,12	0,321	5	4,173	0,1033	25
14	4,95	0,151	6	2,114	0,0229	36
15	4	–0,799	7	–11,985	0,6379	49
Sumas: 120	71,98			10,98	2,4870	280
$\bar{t}=8$	$\bar{y}=4,79867$					

A partir de los productos y sumas anteriores tenemos que:

$$\text{Cov}(y,t)=\frac{\sum_{i=1}^{n}(y-\bar{y})(t-\bar{t})}{n}=\frac{10,98}{15}=0,732$$

$$\text{Var}(t)=\frac{\sum_{i=1}^{n}(t-\bar{t})^2}{n}=\frac{280}{15}=18,67$$

Con todos estos resultados ya podemos obtener los parámetros para el ajuste mínimo-cuadrático:

$$b = \frac{\text{Cov}(y,t)}{\text{Var}(t)} = \frac{0,732}{18,67} = 0,039207$$

$$a = \bar{y} - b\bar{t} = 4,79867 - (0,039207 \cdot 8) = 4,48501$$

Por tanto, el ajuste lineal estimado por mínimos cuadrados resulta ser:

$$Y = 4,48501 + 0,039207t$$

Para la 16.ª quincena tendremos que:

$$\hat{Y}(16) = 4,48501 + 0,039207 \cdot 16 = 5,11$$

Por tanto, el nivel de ocupación pronosticado para la quincena en estudio del año 2003 es de aproximadamente 5.000 personas.

b) Para resolver este apartado necesitamos calcular el coeficiente de determinación. Para ello debemos obtener inicialmente la varianza residual tal que:

y	$y^* = a + bt$	$e = y - y^*$	e^2
4,1	4,5242	–0,4242	0,1800
4,5	4,5634	–0,0634	0,0040
4,3	4,6026	–0,3026	0,0916
4,5	4,6418	–0,1418	0,0201
4,7	4,6810	0,0190	0,0004
4,9	4,7203	0,1797	0,0323
5,2	4,7595	0,4405	0,1941
5	4,7987	0,2013	0,0405
5,1	4,8379	0,2621	0,0687
5,16	4,8771	0,2829	0,0800
5,2	4,9163	0,2837	0,0805
5,25	4,9555	0,2945	0,0867
5,12	4,9947	0,1253	0,0157
4,95	5,0339	–0,0839	0,0070
4	5,0731	–1,0731	1,1516
$\bar{y} = 4,79867$			2,0533

Por tanto, la varianzas de los residuos y de la variable y son:

$$S_e^2 = \frac{\sum e_i^2}{n} = \frac{2,0533}{15} = 0,136887$$

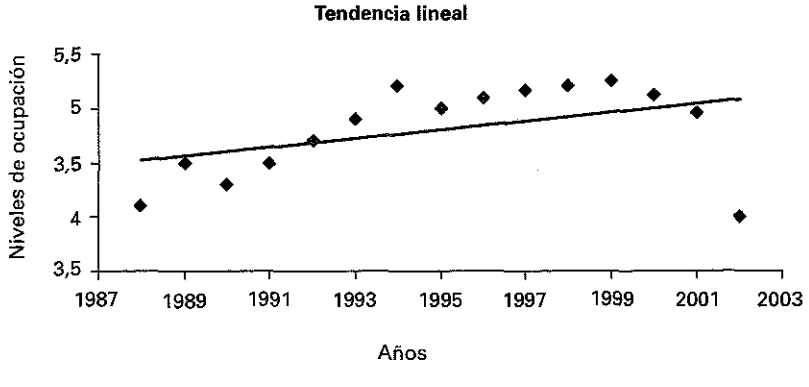
$$S_y^2 = \frac{\sum_{i=1}^{n} (y - \overline{y})^2}{n} = \frac{2,487}{15} = 0,1658$$

Por lo que el coeficiente de determinación resulta ser:

$$R^2 = 1 - \frac{0,136887}{0,1658} = 0,1744$$

La variación explicada por el modelo de regresión lineal es aproximadamente del 17,44% de la variación total de los datos; por lo tanto debemos interpretar que el ajuste realizado es deficiente, es decir, la recta no constituye una curva que explique el comportamiento de los datos de forma adecuada.

c)



Ejercicio 7.5. *Una empresa desea proyectar el volumen de ventas que se efectuará la próxima temporada. Para ello se cuenta con los datos de los últimos ocho años.*

Año	Ventas (millones €)
1996	21
1997	23
1998	26
1999	30
2000	50
2001	80
2002	115
2003	118

A partir de dichos datos:

a) *Obtenga el volumen de ventas estimado para el año 2004 mediante un ajuste a través de una curva exponencial. Centralice la variable explicada para simplificar los cálculos.*

b) *Calcule el error cometido por el ajuste para el año 2003.*

c) *Represente gráficamente la tendencia exponencial.*

Respuesta

a) Del mismo modo que antes, redefinimos el tiempo para simplificar los cálculos, expresando los mismos como periodos sucesivos y no como años. Seguidamente tomamos logaritmos neperianos en la variable y realizamos la siguiente tabla

	y	$\ln(y)$	t	$t \cdot \ln(y)$	t^2
	21	3,0445	1	3,0445	1
	23	3,1355	2	6,2710	4
	26	3,2581	3	9,7743	9
	30	3,4012	4	13,6048	16
	50	3,9120	5	19,5601	25
	80	4,3820	6	26,2922	36
	115	4,7449	7	33,2145	49
	118	4,7707	8	38,1655	64
Total	463	30,6490	36	149,9269	204

A partir del sistema de ecuaciones normales:

$$\sum_{i=1}^{8} \ln y_i = Na + b\sum_{i=1}^{8} t_i$$

$$\sum_{i=1}^{8} t_i \ln y_i = a\sum_{i=1}^{8} t_i + b\sum_{i=1}^{8} t_i^2$$

Podemos obtener los parámetros a y b tal que:

$$30{,}6490 = 8a + 36b$$

$$149{,}9269 = 36a + 204b$$

De donde se obtiene $a = 2{,}5389$ y $b = 0{,}2869$.
Deshaciendo la transformación logarítmica tenemos que:

$$\text{Ln}(a) = 2{,}5389 \quad \Rightarrow \quad a = e^{2{,}5389} \cong 12{,}6657$$

$$\text{Ln}(b) = 0{,}2869 \quad \Rightarrow \quad b = e^{0{,}2869} \cong 1{,}3323$$

Por tanto, el ajuste exponencial resultante es:

$$Y = 12{,}6657 \cdot 1{,}3323^t$$

Y el volumen de ventas esperado para la temporada 2004 será:

$$\hat{Y}(9) = 12{,}6657 \cdot 1{,}3323^9 = 167{,}51$$

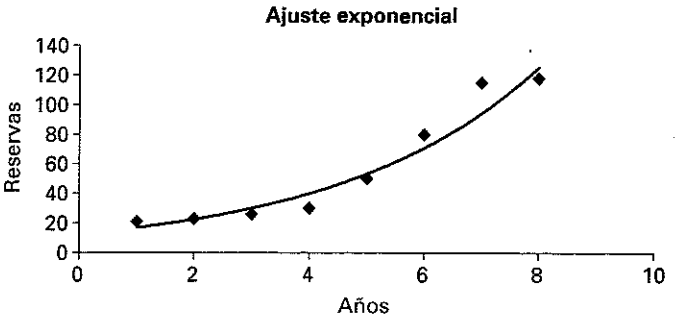
Se esperan unas ventas de 168 millones de euros para el próximo año.

b) El error de ajuste, medido como la diferencia entre el número de reservas observado y pronosticado será:

$$EP = 118 - \hat{Y}(8) = 118 - 125{,}732 = 7{,}73$$

La diferencia entre el dato histórico (real) y el estimado para el año 2003 fue de aproximadamente –8 millones de euros.

c)



Ejercicio 7.6. *Para los datos del ejercicio anterior, se pide:*

a) *Realice el mismo pronóstico a través de una curva potencial.*

b) *Calcule el error del pronóstico para el año 2003. Compare con el ajuste anterior.*

c) *Calcule una medida absoluta de la bondad del ajuste realizado. Compárelo con el ajuste anterior.*

d) *Represente gráficamente la curva de tendencia. Compárela con el ajuste anterior.*

Respuesta

a) Del mismo modo que antes, redefinimos el tiempo para simplificar los cálculos, expresando los mismos como periodos sucesivos y no como años. Seguidamente tomamos logaritmos neperianos en las variables y construimos la siguiente tabla:

	Periodos	Y	ln (*t*)	ln (*y*)	ln (*t*) · ln (*y*)	(ln (*t*))^2
	1	21	0	3,0445	0	0
	2	23	0,6931	3,1355	2,1734	0,48039
	3	26	1,0986	3,2581	3,5794	1,20692
	4	30	1,3863	3,4012	4,7151	1,92183
	5	50	1,6094	3,9120	6,2962	2,59017
	6	80	1,7918	4,3820	7,8515	3,21055
	7	115	1,9459	4,7449	9,2332	3,78653
	8	118	2,0794	4,7707	9,9204	4,32390
Total	36	463	10,6046	30,6490	43,7691	17,5203

A partir del sistema de ecuaciones normales:

$$\sum_{i=1}^{8} \ln y_i = Na + b \sum_{i=1}^{8} \ln t_i$$

$$\sum_{i=1}^{8} \ln t_i \cdot \ln y_i = a \sum_{i=1}^{8} \ln t_i + b \sum_{i=1}^{8} (\ln t_i)^2$$

Podemos obtener los parámetros a y b tal que:

$$30{,}6490 = 8a + 10{,}6046b$$

$$43{,}7691 = 10{,}6046a + 17{,}5203b$$

De donde se obtiene $a = 2{,}6287$ y $b = 0{,}9071$.
Deshaciendo la transformación logarítmica tenemos que:

$$\text{Ln}\,(a) = 2{,}6287 \quad \Rightarrow \quad a = e^{2{,}6287} = 13{,}856$$

Por tanto, el ajuste potencial obtenido es:

$$Y = 13{,}856 \cdot t^{0{,}907082}$$

Y el número volumen de ventas esperado para el año 2004 es:

$$\hat{Y}(9) = 13{,}856 \cdot 9^{0{,}907082} \cong 101{,}675$$

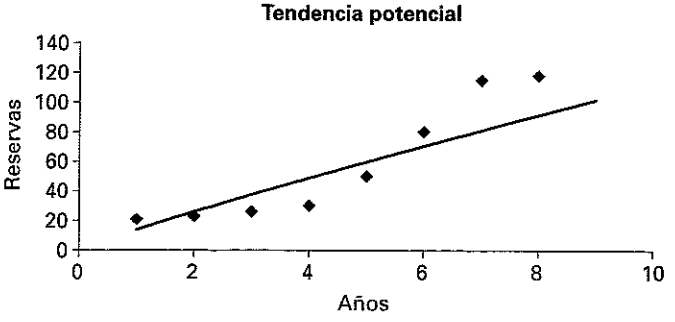
Se esperan aproximadamente 102 millones de euros de volumen de negocio para el próximo año.

b) Error del pronóstico para el año 2003:

$$EP = 118 - \hat{Y}(8) = 118 - 91{,}3722 = 26{,}627$$

Dicho error, como puede observarse, es mayor que en el caso anterior.

c) Como puede apreciarse, el grafico del ajuste exponencial describe mejor la evolución de los datos que el gráfico del ajuste potencial.



Ejercicio 7.7. *La siguiente tabla muestra la evolución del precio medio de las publicaciones de una compañía editorial en euros en los últimos 15 años.*

a) *Construya una nueva serie suavizada a partir de:*
   1. *Un promedio móvil de 2 años.*
   2. *Un promedio móvil de 5 años.*

b) *Calcule el error cuadrático medio en cada caso.*

c) *Analice los gráficos de tendencia correspondiente.*

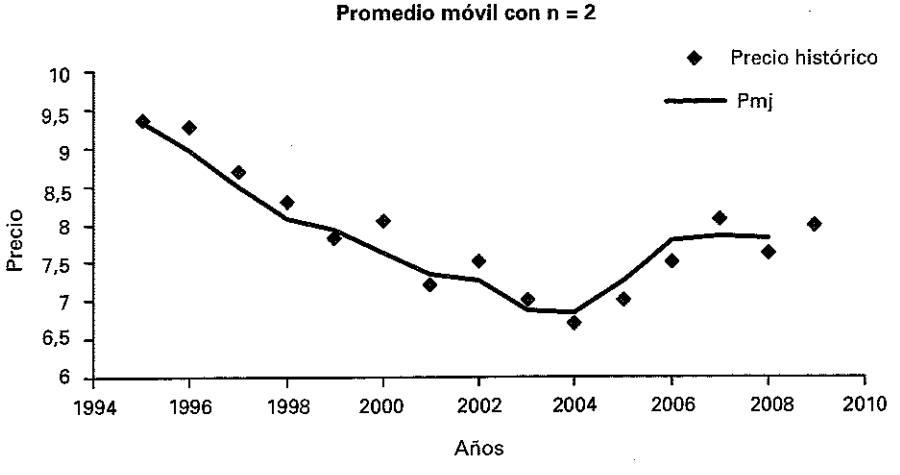
Años	Precio
1995	9,36
1996	9,27
1997	8,68
1998	8,3
1998	7,83
2000	8,05
2001	7,2
2002	7,5
2003	7,01
2004	6,7
2005	7
2006	7,5
2007	8,07
2008	7,63
2009	8

Respuesta

a.1) Si tomamos $n = 2$ tenemos que:

Años	Precio	Total móvil (2)	P_{mj}	Errores Cuadráticos
1995	9,36	18,63	**9,315**	0,0020
1996	9,27	17,95	**8,975**	0,0870
1997	8,68	16,98	**8,490**	0,0361
1998	8,3	16,13	**8,065**	0,0552
1999	7,83	15,88	**7,940**	0,0121
2000	8,05	15,25	**7,625**	0,1806
2001	7,2	14,70	**7,350**	0,0225
2002	7,5	14,51	**7,255**	0,0600
2003	7,01	13,71	**6,855**	0,0240
2004	6,7	13,70	**6,850**	0,0225
2005	7	14,50	**7,250**	0,0625
2006	7,5	15,57	**7,785**	0,0812
2007	8,07	15,70	**7,850**	0,0484
2008	7,63	15,63	**7,815**	0,0342
2009	8			
Total				0,7285
			ECM =	0,0520

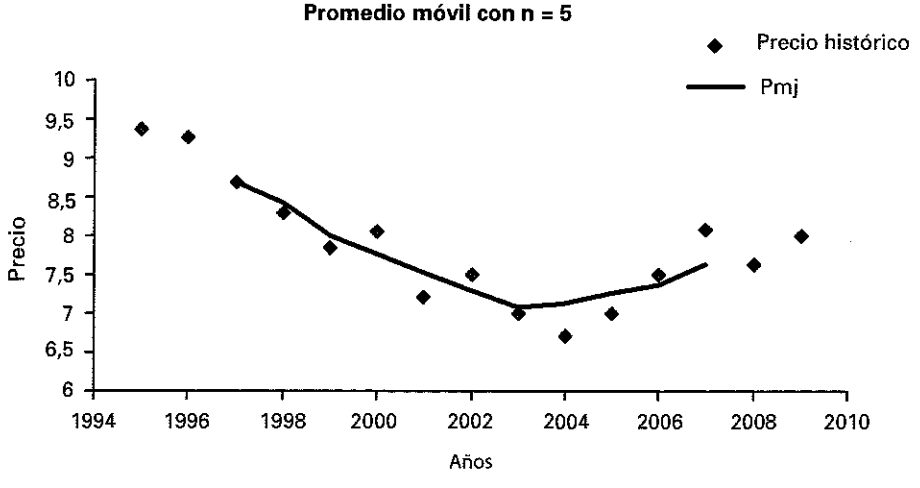
$$ECM = \frac{0,7285}{15-1} \cong 0,0520$$



a.2) Para $n = 5$:

Años	Coste	Total móvil (5)	P_{mj}	Errores Cuadráticos
1995	9,36			
1996	9,27			
1997	8,68	43,44	8,688	0,0001
1998	8,3	42,13	8,426	0,0159
1999	7,83	40,06	8,012	0,0331
2000	8,05	38,88	7,776	0,0751
2001	7,2	37,59	7,518	0,1011
2002	7,5	36,46	7,292	0,0433
2003	7,01	35,41	7,082	0,0052
2004	6,7	35,71	7,142	0,1954
2005	7	36,28	7,256	0,0655
2006	7,5	36,9	7,38	0,0144
2007	8,07	38,2	7,64	0,1849
2008	7,63			
2009	8			
Total				0,7339
			ECM	0,0667

$$ECM = \frac{0,7339}{15-4} \cong 0,0667$$



b) El ECM para $n = 2$ es de 0,0520 y para $n = 5$ es de 0,0667, lo que significa que tomar un promedio móvil a partir de una mayor cantidad de datos implica aumentar el nivel de error que se comete en los pronósticos. Cuando se toman menos datos, la serie suavizada se parece más a la histórica, pero es mas difícil de evidenciar la tendencia de la misma.

c) La observación de los gráficos muestra que en el segundo caso (cuando tomamos $n = 5$) se puede ver mejor la tendencia a largo plazo que cumple la serie, pero el primer caso pone en evidencia con mayor precisión las variaciones cíclicas de la serie.

Ejercicio 7.8. *La siguiente tabla muestra la cantidad de reservas recibidas por una empresa hotelera durante los años 2004-2009.*

a) *Halle un índice estacional por medio del método de porcentaje medio.*

b) *Realice la representación gráfica correspondiente.*

	Ene.	Febr.	Mar.	Abr.	Ma.	Jun.	Jul.	Agos.	Sept.	Octu.	Novi.	Dici.
2004	1.800	1.600	1.640	1.300	1.200	1.000	1.300	1.200	1.250	1.090	1.400	1.600
2005	1.997	1.630	1.600	1.400	1.300	1.150	1.450	1.150	1.320	1.100	1.360	1.550
2006	1.970	1.720	1.000	1.430	1.420	900	1.425	1.230	1.360	1.190	1.450	1.350
2007	2.100	1.920	1.800	1.450	1.360	1.050	1.500	1.320	1.480	1.260	1.420	1.620
2008	2.000	1.880	1.750	1.500	1.230	800	1.550	1.350	1.400	1.320	1.600	1.700
2009	2.060	1.858	1.810	1.350	1.260	850	1.600	1.400	1.600	1.400	1.290	1.830

Respuesta

a) En la siguiente tabla se calcularon los promedios mensuales de cada año. Por ejemplo, para el año 2004 se sumaron las cantidades de todos los meses correspondientes a ese año y se dividió el resultado por 12.

Año	2004	2005	2006	2007	2008	2009
Total suma anual	16.380,0	17.007,0	16.445,0	18.280,0	18.080,0	18.308,0
Promedio mensual	1.365,0	1.417,3	1.370,4	1.523,3	1.506,7	1.525,7

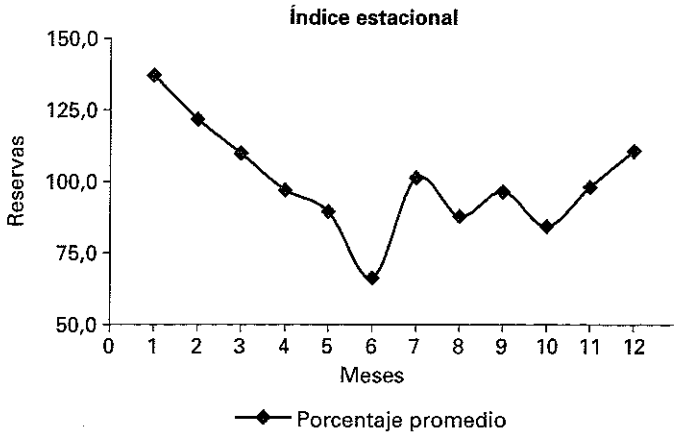
Una vez obtenido los valores promedio mensual de cada año, se dividen los datos históricos de cada mes por el promedio del año correspondiente para obtener la siguiente tabla, donde se expresan los resultados como porcentajes respecto de dichos promedios.

Como la suma de los porcentajes medios obtenidos en la última fila es 1.200 %, es decir que el promedio de los porcentajes es 100 % (1.200% dividido por la cantidad de meses en el año), no es necesario realizar ajustes adicionales.

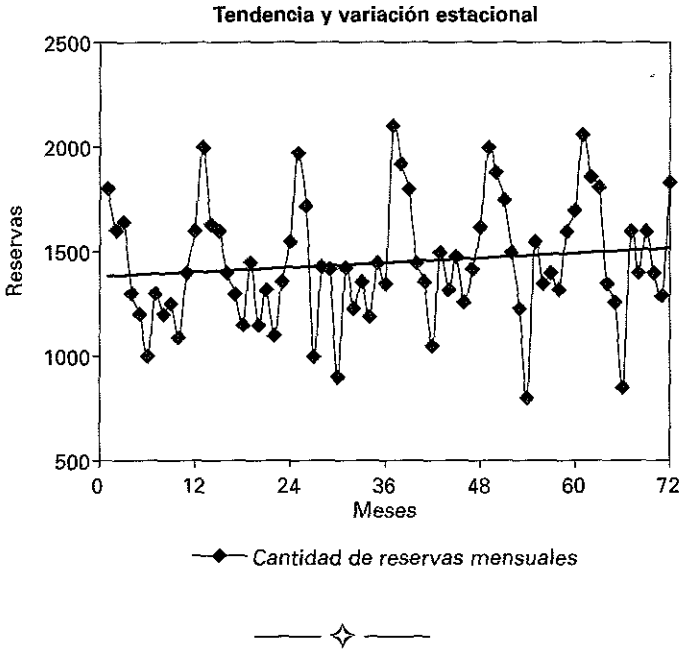
Finalmente, los valores de la fila inferior representan el índice estacional.

	Ene.	Febr.	Mar.	Abr.	Ma.	Jun.	Jul.	Agos.	Sept.	Octu.	Novi.	Dici.
2004	131,9	117,2	120,1	95,2	87,9	73,3	95,2	87,9	91,6	79,9	102,6	117,2
2005	140,9	115,0	112,9	98,8	91,7	81,1	102,3	81,1	93,1	77,6	96,0	109,4
2006	143,8	125,5	73,0	104,3	103,6	65,7	104,0	89,8	99,2	86,8	105,8	98,5
2007	137,9	126,0	118,2	95,2	89,3	68,9	98,5	86,7	97,2	82,7	93,2	106,3
2008	132,7	124,8	116,1	99,6	81,6	53,1	102,9	89,6	92,9	87,6	106,2	112,8
2009	135,0	121,8	118,6	88,5	82,6	55,7	104,9	91,8	104,9	91,8	84,6	119,9
Sumas por mes	822,1	730,3	659,0	581,6	536,8	397,8	607,7	526,8	578,9	506,4	588,3	664,2
Porcent. medio	137,0	121,7	109,8	96,9	89,5	66,3	101,3	87,8	96,5	84,4	98,0	110,7
Sumas porcent.	1.200											

b) El siguiente gráfico representa la variación estacional mediante el índice obtenido:



El siguiente gráfico representa la evolución de la variable a lo largo de los seis (6) años estudiados.



Ejercicio 7.9. *La siguiente tabla muestra los datos correspondientes al número de ocupados, en miles de personas, en España durante el periodo 2005-2009 según la Encuesta de Población Activa (EPA) del Instituto Nacional de Estadística.*

Evolución de los ocupados (miles) en España, periodo 2005-2009

Año	Trimestre	Ocupados	Año	Trimestre	Ocupados
2005	1	18.492,70	2007	3	20.510,60
2005	2	18.894,90	2007	4	20.476,90
2005	3	19.191,10	2008	1	20.402,30
2005	4	19.314,30	2008	2	20.425,10
2006	1	19.400,10	2008	3	20.346,30
2006	2	19.693,10	2008	4	19.856,80
2006	3	19.895,60	2009	1	19.090,80
2006	4	20.001,80	2009	2	18.945,00
2007	1	20.069,20	2009	3	18.870,20
2007	2	20.367,30	2009	4	18.645,90

FUENTE: Encuesta de Población Activa. INE

Obtener el índice estacional usando el método del porcentaje promedio móvil. Realice los gráficos correspondientes

Respuesta

En el siguiente cuadro se realizan una serie de pasos para llegar al índice buscado:

1. En la cuarta columna se calculan promedios móviles para 4 trimestres. Recordemos que estos datos aparecerían centrados en el periodo 2,5, 3,5, 4,5, ..., si bien por convenio, al utilizar únicamente dichas medias móviles se suelen situar en el periodo 2, 3, 4, ...
2. En la quinta columna se calculan los promedios móviles centrados, calculados como promedios móviles de los valores obtenidos en la columna anterior con $n = 2$. De esta manera, el primer promedio aparecería centrado en el periodo 3, ya que los datos utilizados se centraban en 2,5 y 3,5. El segundo se centra en 4 y así sucesivamente.
4. Finalmente se dividen los valores mensuales reales por cada promedio móvil centrado y se expresa el resultado como porcentaje. Por ejemplo, el valor de 100,55 se obtiene de multiplicar por 100 al cociente entre 19.191,10 y 19.086,68. De esta manera se obtiene un nuevo valor de serie expresado como porcentaje sobre la media móvil.

Año	Trimestre	Ocupados (miles)	Promedio móvil 4 trimestres	Promedio móvil centrado	Porcentaje sobre media móvil
2005	1	18.492,70			
2005	2	18.894,90	18.973,25		
2005	3	19.191,10	19.200,10	19.086,68	100,55
2005	4	19.314,30	19.399,65	19.299,88	100,07
2006	1	19.400,10	19.575,78	19.487,71	99,55
2006	2	19.693,10	19.747,65	19.661,71	100,16
2006	3	19.895,60	19.914,93	19.831,29	100,32
2006	4	20.001,80	20.083,48	19.999,20	100,01
2007	1	20.069,20	20.237,23	20.160,35	99,55
2007	2	20.367,30	20.356,00	20.296,61	100,35

(*Continúa*)

(Continuación)

Año	Trimestre	Ocupados (miles)	Promedio móvil 4 trimestres	Promedio móvil centrado	Porcentaje sobre media móvil
2007	3	20.510,60		20.397,64	100,55
			20.439,28		
2007	4	20.476,90		20.446,50	100,15
			20.453,73		
2008	1	20.402,30		20.433,19	99,85
			20.412,65		
2008	2	20.425,10		20.335,14	100,44
			20.257,63		
2008	3	20.346,30		20.093,69	101,26
			19.929,75		
2008	4	19.856,80		19.744,74	100,57
			19.559,73		
2009	1	19.090,80		19.375,21	98,53
			19.190,70		
2009	2	18.945,00		19.039,34	99,50
			18.887,98		
2009	3	18.870,20			
2009	4	18.645,90			

Ahora estamos en condiciones de calcular el índice estacional. El mismo resultará de calcular el porcentaje promedio correspondiente a cada trimestre. Por ejemplo, para el primer trimestre se promedian los valores porcentuales correspondientes a dicho trimestre en los diferentes años.

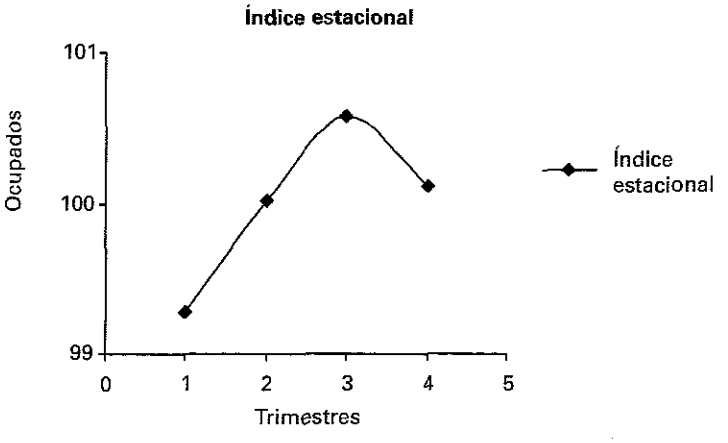
$$\overline{X}_{enero} = \frac{99,55 + 99,55 + 99,85 + 98,53}{4} \Rightarrow$$

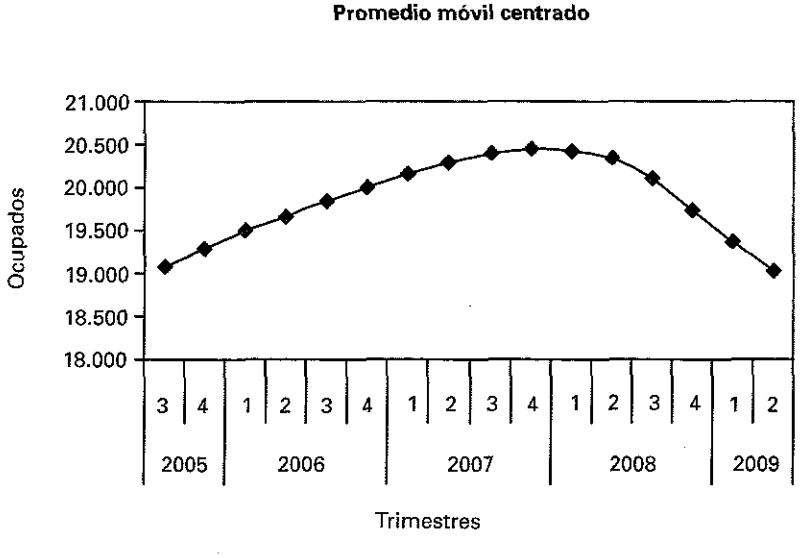
$$\overline{X}_{trimestre1} \cong 99,37\%$$

Como la suma de los porcentajes medios obtenidos en la última fila es diferente a 400%, es decir que el promedio de los porcentajes no es 100% (400% dividido por la cantidad de trimestres en el año), es necesario realizar un ajuste en los porcentajes medios mensuales para llegar a dicha cifra. Dicho ajuste se encuentra en la tercera columna del siguiente cuadro.

En el siguiente cuadro se detallan los índices obtenidos:

Meses	Porcentaje medio	Porcentaje medio ajustado
Trimestre 1	99,37	99,28
Trimestre 2	100,11	100,02
Trimestre 3	100,67	100,58
Trimestre 4	100,20	100,11
Suma	400,36	400





Ejercicio 7.11. *Una empresa usa la suavización exponencial simple con coeficiente de alisado $\alpha = 0,1$ para predecir sus ventas. La predicción para la semana 1 fue de 550 miles de euros, mientras que el valor real es de 500 miles de euros.*

a) *Pronosticar las ventas para la semana 2.*

b) *Suponiendo que las ventas para la semana 2, 3, 4 y 5 son de 600, 500, 550 y 525, pronostique las ventas para las semanas 3, 4 y 5.*

c) *Valore los errores obtenidos.*

Respuesta

Recordemos que el pronóstico para un suavizado exponencial simple se calcula como:

$$\hat{X}_t = S_{t-1}$$

siendo

$$S_t = \alpha X_t + (1 - \alpha)\, S_{t-1}$$

En nuestro caso, tenemos:

$$S_0 = 550 \qquad X_1 = 500$$

a) La predicción para la semana 2 será igual a S_1, que según la fórmula será igual a:

$$\hat{X}_2 = S_1 = 0,1X_1 + 0,9S_0 = 0,1 \cdot 500 + 0,9 \cdot 550 = 545$$

b) Las predicciones para las tres semanas restantes serán:

$$\hat{X}_3 = S_2 = 0,1X_2 + 0,9S_1 = 0,1 \cdot 600 + 0,9 \cdot 545 = 550,5$$

$$\hat{X}_4 = S_3 = 0,1X_3 + 0,9S_2 = 0,1 \cdot 550 + 0,9 \cdot 550,5 = 545,45$$

$$\hat{X}_5 = S_4 = 0,1X_4 + 0,9S_3 = 0,1 \cdot 550 + 0,9 \cdot 545,5 = 545,905$$

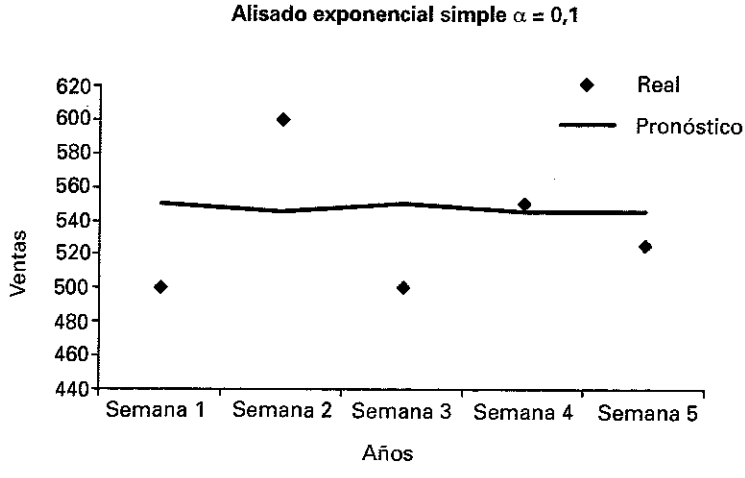
c) Se muestran en la siguiente tabla los valores de ventas reales, los pronosticados por la serie suavizada y los errores obtenidos:

	Real	Pronóstico	Error	%	Error²
Semana 1	500	550	–50	–10,00	2500,00
Semana 2	600	545	55	9,17	3025,00
Semana 3	500	550,5	–50,5	–10,10	2550,25
Semana 4	550	545,5	4,55	0,83	20,70
Semana 5	525	545,9	–20,905	–3,98	437,02
Total					8.532,97
Promedio	535,00	547,37	–12,37	2,82	1706,59

La Raíz del Error Cuadrático Medio toma el valor:

$$RECM = \sqrt{\frac{\sum Error^2}{T}} = \sqrt{\frac{8.532,97}{5}} = \sqrt{1.706,59} = 41,31$$

El gráfico con la serie original y suavizada sería:



Lo idóneo en este caso sería considerar diferentes valores de α, y considerar aquel con el que se obtenga el menor Error Cuadrático Medio y/o el gráfico nos muestre el mejor ajuste.

Capítulo 8

INTRODUCCIÓN A LA PROBABILIDAD

8.1. INTRODUCCIÓN. FENÓMENOS ALEATORIOS Y SUCESOS

En este capítulo desarrollaremos los conceptos básicos de la teoría de la probabilidad; estos conceptos son la base para abordar la inferencia estadística, es decir, para inferir datos a una población a partir de los resultados extraídos de una muestra de la misma, lo que se aborda en el resto de asignaturas de la materia de Métodos Cuantitativos para la Empresa, en la Estadística Empresarial.

Uno de los objetivos de la ciencia consiste en describir y predecir sucesos que ocurren a nuestro alrededor de forma cotidiana. Una manera de hacerlo es mediante la construcción de modelos matemáticos.

Dado el nivel básico de este libro no entraremos en detalles sobre estos modelos matemáticos, conocidos como distribuciones de probabilidad, baste saber que la más habitual de estas distribuciones es la denominada *Distribución Normal* o de Gauss, y que a ella se ajustan buena parte de los fenómenos naturales, económicos y sociales.

En los fenómenos naturales existen patrones de comportamiento que se denominan deterministicos (la hora de salida o puesta del sol, la fuerza con la que cae al suelo un objeto de un Kg. de peso atraído por la gravedad, el resultado de la unión en laboratorio de dos moléculas de oxígeno con una de hidrógeno, etc.; estos comportamientos están determinados por leyes más o menos complejas descubiertas por el hombre (las leyes sobre el movimiento de los astros, la ley de la gravedad, etc.).

En otros casos, no se conoce o no es posible predecir los comportamientos del suceso en cuestión; así, por ejemplo, no nos resulta posible predecir el resultado de un partido de fútbol o saber el resultado del lanzamiento de un dado, etc.

Se dice que un suceso es de carácter determinístico cuando al repetirlo en idénticas condiciones da siempre el mismo resultado y que es de carácter aleatorio cuando al reproducirlo en las mismas condiciones no siempre se obtiene un mismo resultado.

Los resultados de los experimentos o sucesos deterministicos se «predicen» aplicando el modelo o fórmula matemática que los regula; para predecir los sucesos aleatorios, al no existir la ecuación que los determine con total exactitud o ser excesivamente compleja, es necesario un procedimiento que «aproxime» el resultado satisfactoriamente mediante un modelo matemático relativamente sencillo; este es el cometido de la teoría de la probabilidad.

Establecemos a continuación un conjunto de definiciones, necesarias para abordar este cometido[16]:

— **Experimentos o fenómenos aleatorios** son los que al repetirlos bajo análogas condiciones, no se puede predecir el resultado que se va a obtener porque éste depende del azar; se verifican tres condiciones:

1. Se puede repetir indefinidamente en las mismas condiciones.
2. Antes de realizarlo, no se puede predecir el resultado que se va a obtener.
3. El resultado que se obtenga, $e_1, e_2, ..., e_n$ pertenece a un conjunto conocido previamente de resultados posibles E ($e_1, e_2, ..., e \in E$).

A cada uno de los resultados posibles $e_1, e_2, ..., e_n$ se le denomina **suceso elemental;** también se les llama **resultado básico, elemental, comportamiento individual o punto muestral.**

— ***Espacio muestral o espacio de comportamientos,*** es el conjunto formado por todos los sucesos o resultados posibles de un experimento aleatorio. Se designa por E o Ω; para el lanzamiento de un dado, el espacio muestral E, será $E = \{1, 2, 3, 4, 5, 6\}$, para el lanzamiento de una moneda, $E = \{\text{Cara, Cruz}\}$; si E tiene un número finito, n, de elementos, el número de sucesos de E es 2^n.

Se denomina **suceso aleatorio** a cualquier subconjunto de E; suele denotarse con letras mayúsculas $A, B,...$ Al conjunto de todos los sucesos que ocurren en un experimento aleatorio se le llama **espacio de sucesos** y se designa por S.

Ejemplo:

— **Experimento aleatorio:** lanzamiento de dos dados al aire y anotar la suma de las dos puntuaciones obtenidas.

— **Espacio muestral:** $E = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

— **Suceso aleatorio:**

que la suma obtenida sea mayor o igual 8: A= {8, 9, 10, 11, 12}
que la suma sea múltiplo de 3: B= {3, 6, 9, 12}
etc.

Tipos de sucesos aleatorios; pueden ser de varios tipos:

— **Sucesos elementales** son los que están formados por un solo resultado del experimento, Por ejemplo, que el resultado de la suma de los dados sea igual a 10.

[16] El alumno que no tenga suficientes nociones del álgebra de sucesos y de la teoría de la probabilidad debe repasar un Manual en el que se expliquen con mayor detalle estas definiciones y conceptos. Ver, por ejemplo, Casas Sánchez y Santos Peñas *Introducción a la Estadística para Economía y Administración de Empresas*. Ed. Centro de Estudios Ramón Areces. 1995.

— **Sucesos compuestos** son los que están formados por dos o más resultados del experimento, es decir, por dos o más sucesos elementales. Como ejemplo, los sucesos A y B definidos.

— **Suceso seguro** es el que ocurre siempre que se realice el experimento aleatorio. Está formado por todos los resultados posibles del experimento y, por tanto, coincide con el espacio muestral; al tirar los dos dados y anotar la suma, el *suceso seguro E es que el resultado sea mayor o igual a 2 y menor o igual a 12.*

— **Suceso vacío imposible** es el que nunca se verifica. Se representa por Ø. En nuestro ejemplo, cualquier valor menor que 2 y mayor que 12.

***Sistema completo de sucesos*:** Se dice que un conjunto de sucesos $A_1, A_2, \ldots A_n$, constituyen un sistema completo cuando se verifica que:

— $A_1 \cup A_2 \cup \ldots A_n = E$

— $A_1 \cup A_2 \cup \ldots A_n$ son incompatibles 2 a 2.

Operaciones con sucesos

Unión de sucesos	$A \cup B$	Es el suceso formado por todos los elementos de A y todos los elementos de B (suceso que se verifica cuando se realiza A ó B).
Intersección de sucesos	$A \cap B$	Es el suceso formado por todos los elementos que son, a la vez, de A y de B (suceso que se verifica cuando se realizan simultáneamente los sucesos A y B).
Diferencia	A – B	Es el suceso formado por todos los elementos de A que no son de B (suceso que se verifica cuando se verifica A y no se verifica B).
Suceso complementario o Suceso contrario	$\bar{A}$ $\bar{A} = E - A$	Dado un suceso A, denotaremos mediante $\bar{A}$ al suceso que se verifica cuando no se verifica A; por ende, se verifica que $A \cup \bar{A} = E$ y $A \cap \bar{A} = \Phi$.
Sucesos incompatibles	$A \cap B \approx \emptyset$	Dos sucesos A y B, se llaman **incompatibles** cuando no tienen ningún elemento común.
Diferencia simétrica	$A \Delta B$	Dados dos sucesos A y B, denotaremos mediante $A \Delta B$ al suceso que se verifica cuando o bien se verifica A y no se verifica B, o bien se verifica B y no se verifica A.

Propiedades de los sucesos

Propiedad	Unión	Intersección
Conmutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Asociativa	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
Idempotente	$A \cup A = A$	$A \cap A = A$
Simplificación	$A \cup (B \cap A) = A$	$A \cap (B \cup A) = A$
Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Elemento neutro	$A \cup \emptyset = A$	$A \cap E = A$
Absorción	$A \cup E = E$	$A \cap \emptyset = \emptyset$
Leyes de Morgan	El suceso contrario de la unión de dos sucesos es la intersección de sus sucesos contrarios: $\overline{A \cup B} = \overline{A} \cap \overline{B}$	El suceso contrario de la intersección de dos sucesos es la unión de sus sucesos contrarios: $\overline{A \cap B} = \overline{A} \cup \overline{B}$

Álgebra de Sucesos

El álgebra de sucesos S es una familia de sucesos definida mediante los siguientes axiomas:

— **Axioma 1°:** $\Omega \in S$.
— **Axioma 2°:** Si $A, B \in S$, entonces $A \cup B \in S$.
— **Axioma 3°:** Si $A \in S$, entonces $\overline{A} \in S$.

8.2. DEFINICIÓN DE PROBABILIDAD

El concepto y definición de probabilidad ha variado con el desarrollo histórico de la teoría; las tres definiciones históricas de este término son la clásica, la frecuencial y la axiomática; sinteticemos su significado:

a) **Probabilidad clásica o a priori:** es una probabilidad inicial. Por ejemplo, antes de lanzar una moneda al aire, se supone que la probabilidad de que salga cara va a ser igual a ½, de igual forma, podría decirse que si un dado se arroja muchas veces la probabilidad de obtener un *uno* o un *dos* será de 2/6, ya que este suceso puede aparecer 2 veces de 6. Esta probabilidad se define de dos formas equivalentes:

Si un suceso puede ocurrir de n *maneras mutuamente excluyentes e igualmente verosímiles y si* n_A *de éstas poseen un atributo* A, *la probabilidad de* A *es la fracción* n_A/n.

Regla de Laplace: *la probabilidad de un suceso aleatorio* S_i *es el cociente entre el número de casos favorables y el número de casos o elementos posibles del experimento.*

$$P(A) = \frac{\text{Número de resultados favorables a } S_i}{\text{Número de resultados posibles de } E}$$

Un experimento aleatorio se caracteriza porque repetido muchas veces y en idénticas condiciones este cociente tiende a un número fijo. Esta propiedad es conocida como **ley de los grandes números**, establecida por *Jakob Bernouilli.*

Esta definición presenta algunos inconvenientes: el primero es que considera todos los sucesos igual de probables, lo que no siempre ocurre en la naturaleza; el segundo es tenemos que realizar el experimento un gran número de veces para obtener el valor aproximado de la probabilidad; tampoco es posible aplicarla en múltiples sucesos aleatorios, como obtener la probabilidad de que una persona muera antes de una determinada edad, que un determinado individuo sufra un accidente de tráfico, etc.; por ello se avanzó hacia los siguientes conceptos de probabilidad.

b) **Probabilidad a posteriori o frecuencial:** es una probabilidad experimental. Por ejemplo, si sospechamos que un dado no está equilibrado, la única manera de probar esta sospecha es arrojándolo muchas veces y observar si la frecuencia relativa de obtener un uno se aproxima a un sexto.

La probabilidad a posteriori se define como el límite de la frecuencia relativa cuando el número de experimentos realizados tiende a infinito, y se enuncia formalmente de la siguiente manera:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Donde

A: Es el suceso cuya probabilidad se desea obtener.

n: Es el número de veces que se repite el experimento (lanzamiento del dado).

n_A: Es el número de veces que aparece el resultado A.

n_A/n: Es la frecuencia relativa.

lím: Es el límite de la frecuencia relativa a medida que el número de lanzamientos se aproxima a infinito.

Para este concepto de probabilidad hemos tenido que introducir el concepto matemático de límite; con ello se resuelve el segundo de los inconvenientes que aporta la definición anterior, pero no el primero, ya que se siguen considerando todos los sucesos como equiprobables.

c) **Probabilidad axiomática:** este concepto de probabilidad basado en un conjunto de axiomas, que formulamos de la siguiente forma:

Sea S un espacio muestral (conjunto de todos los posibles sucesos de un determinado experimento) y A un determinado suceso de S (cualquier elemento o subconjunto de S), diremos que P es una *función de probabilidad* en el espacio muestral S si se satisfacen los tres axiomas siguientes:

— *Axioma 1.* $P(A)$ es un número real tal que $P(A) \geq 0$ para todo suceso A de S, es decir, la probabilidad de cualquier suceso en un experimento es siempre mayor o igual que 0.

— *Axioma 2.* $P(S) = 1$, es decir, la probabilidad de todos los sucesos posibles de un experimento es igual a 1.

— *Axioma 3.* Si $A, B, C, \ldots$, es una sucesión de sucesos mutuamente excluyentes de S, la probabilidad asociada a la unión de todos ellos (que en un experimento ocurra cualquiera de ellos) es igual a la suma de sus probabilidades.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

De estos tres axiomas se deducen los siguientes teoremas:

— *Teorema 1.* Si definimos el suceso complementario de A, $\bar{A}$, como aquel que está formado por todos los puntos o sucesos del espacio muestral S que no están en A, entonces la probabilidad de $\bar{A}$ será igual a:

$$P(\bar{A}) = 1 - P(A)$$

— *Teorema 2.* Sea A un suceso de S. Siempre se verifica que la probabilidad de que ocurra está comprendida entre 0 y uno:

$$0 \leq P(A) \leq 1$$

— *Teorema 3.* Si ϕ es el suceso nulo, entonces se verifica que: $P(\phi) = 0$, ya que ϕ es el suceso complementario de S.

Estos tres teoremas dan lugar a un conjunto de propiedades que permiten abordar la mayor parte de los problemas que plantea el cálculo de la probabilidad de que ocurra un fenómeno aleatorio; las principales son las siguientes:

1. $P(\overline{A}) = 1 - P(A)$.

2. $P(\varnothing) = 0$.

3. Si A está contenida en B ($A \subset B$) entonces:

   a) $P(B) = P(A) + P(A - B)$.

   b) $P(A) \leq P(B)$

4. Si $A_1, A_2, \ldots, A_k$, son incompatibles dos a dos, entonces:

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$$

5. Propiedad de la Unión:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

6. Si el espacio muestral E es finito y un sucesos es $A = \{x_1, x_2, \ldots, x_K\}$, entonces:

$$P(A) = P(x_1) + P(x_2) + \ldots + P(x_K)$$

Veamos algunos ejemplos de probabilidad:

Ejemplo 8.1. Supongamos que hacemos sucesivas tiradas de un dado de 6 caras y queremos saber la probabilidad de obtener un 5 o un 6.

Tendremos:

Casos favorables (S_i): dos (5 y 6).

Casos posibles (E): seis (1, 2, 3, 4, 5, 6).

$P(S_i)$ del suceso será 2/6 = 1/3 = 33,3%

Es decir que cada 3 veces que tiremos obtendremos o un 5 o un 6 como resultado.

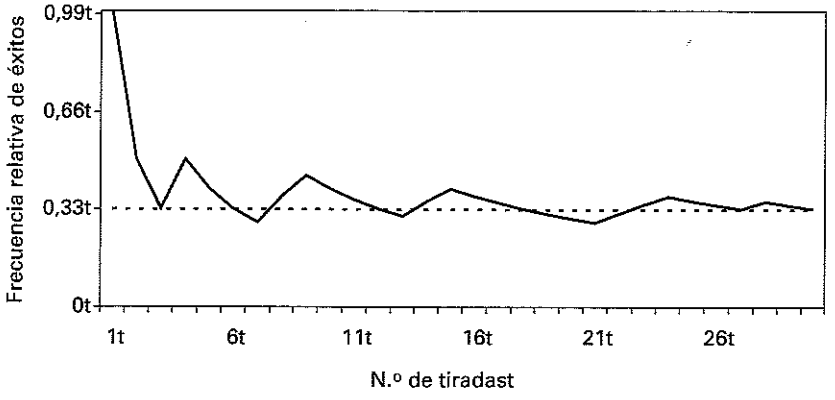
Es evidente que si sólo tiramos 3 veces no tendremos asegurado la obtención de alguno de estos dos números; la probabilidad sólo nos indica que si tiramos el dado muchísimas veces un 33,3% de ellas obtendremos alguno de los números buscados.

Si elaboramos una tabla de resultados con estos 30 intentos, podríamos obtener:

N.º de tiradas	Resultado	N.º de éxitos	Frecuencia relativa del 5 o del 6
1	5	1	1
2	3	1	0,5
3	4	1	0,33333333
4	5	2	0,5
5	2	2	0,4
6	4	2	0,33333333
7	3	2	0,28571429
8	5	3	0,375
9	6	4	0,44444444
10	1	4	0,4
11	1	4	0,36363636
12	3	4	0,33333333
13	4	4	0,30769231
14	5	5	0,35714286
15	6	6	0,4
16	2	6	0,375
17	3	6	0,35294118
18	1	6	0,33333333
19	2	6	0,31578947
20	3	6	0,3
21	4	6	0,28571429
22	5	7	0,31818182
23	6	8	0,34782609
24	5	9	0,375
25	4	9	0,36
26	4	9	0,34615385
27	3	9	0,33333333
28	6	10	0,35714286
29	5	11	0,37931034
30	4	11	0,36666667

Si representamos gráficamente los resultados de 30 intentos obtenemos lo siguiente:

Gráfico 8.1. Probabilidad de éxitos.



Constatándose que los resultados tienden a estabilizarse en torno al 0,33.

Ejemplo 8.2. Considere el experimento de tirar un dado y ver qué número sale. Sean los Sucesos: A = que salga un número par, y B = que salga un número mayor o igual que 3.

Hallar:

a) $P(A)$.

b) $P(B)$.

c) $P(A \cap B)$.

d) $P(A \cup B)$.

Sabemos que:

$A = \{2, 4, 6\}$ números pares del dado, y

$B = \{3, 4, 5, 6\}$ números mayores que 3 del dado.

Para hallar las probabilidades pedidas usaremos la definición clásica.

$$P(A) = \frac{\text{Número de resultados favorables a } A}{\text{Número de resultados posibles}}$$

a) $P(A) = \frac{3}{6} = \frac{1}{2}$.

b) De forma análoga: $P(B) = \frac{4}{6} = \frac{2}{3}$.

c) Como $A \cap B = \{4, 6\}$, entonces $P(A \cap B) = \frac{2}{6} = \frac{1}{3}$

d) Para calcular $P(A \cup B)$ usamos la propiedad de la unión:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

y tendremos:

$$P(A \cup B) = \frac{1}{2} + \frac{2}{3} + \frac{1}{3} \quad \Rightarrow \quad P(A \cup B) = \frac{5}{6}$$

Para verificar este resultado podemos hallar:

$$A \cup B = \{2, 3, 4, 5, 6\}, \quad \text{y de ahí:} \quad P(A \cup B) = \frac{5}{6}$$

8.3. PROBABILIDAD CONDICIONADA. TEOREMA DE BAYES

La probabilidad condicionada permite asignar probabilidades introduciendo informaciones previas acerca del experimento o ciertas creencias subjetivas que se dispongan sobre el mismo.

Veamos varios conceptos y definiciones de interés:

Probabilidad Condicionada

Dados dos sucesos A y B, con $P(B) > 0$, se define la Probabilidad Condicionada de A (Probabilidad de A condicionada a que haya ocurrido el suceso B), como:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Sucesos dependientes e independientes

Dos sucesos A y B se dice que son independientes si $p(A) = p(A/B)$. En caso contrario, $p(A) \neq p(A/B)$, se dice que son dependientes.

Probabilidad de la intersección o probabilidad compuesta:

— Si los sucesos son dependientes $P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$.

— Si los sucesos son independientes $P(A \cap B) = P(A) \cdot P(B)$.

Ejemplo: si al extraer dos cartas de una baraja lo hacemos con devolución tendremos dos sucesos independientes, $P(A \cap B) = P(A) \cdot P(B)$ pero si lo hacemos sin devolución ahora si son dependientes $P(A \cap B) = P(A) \cdot P(B/A)$.

Teorema de la probabilidad total

Sea un **sistema completo de sucesos** (eventos exhaustivos y mutuamente excluyentes $A_1, A_2, A_n \in E$ tales que $A_i \cap A_j = \phi$ si $i \neq j$ y que $P(A_i) > 0$ Para todo «i») y sea un suceso B tal que $P(B/A_i)$ sea conocida, entonces:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots\ldots = \sum P(B \cap A_i)$$

Expresión que es conocida como la Fórmula de **la Probabilidad Total.**

La Regla de Bayes

Sea $A_1, A_2, \ldots, A_n$, un sistema completo de sucesos, tal que:

$$A_i \cap A_j = \phi \quad \forall i, j, \quad i \neq j, \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,n$$

Entonces para cualquier evento B, por el Teorema de la probabilidad total, se tiene que:

$$P(B) = P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) + \ldots + P(A_n)P(B/A_n) \qquad [1]$$

Por otra parte, dado que:

$$P(B/A_i) = \frac{P(A_i \cap B)}{P(A_i)} \Rightarrow P(A_i \cap B) = P(B/A_i) \cdot P(A_i) \qquad [2]$$

Sustituyendo en la expresión siguiente los valores de $P(B)$ y $P(A \cap B)$ por las expresiones derivadas en [1] y en [2] obtenemos:

$$P(A_i/B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B/A_i)P(A_i)}{\sum_{j=1}^{n} P(B/A_j)P(A_j)} \qquad \forall i = 1,\ldots,n$$

Que es la denominada **Regla de Bayes** para el cálculo de la probabilidad condicional.

Ejemplo 8.3. En una determinada región, existen dos empresas de fabricación de turbinas eléctricas, suministrando cada una el 75% y 25% de la demanda respectivamente.

Por otra parte, se sabe que sólo el 80% de las turbinas de la empresa 1 pasan los controles de calidad establecidos, mientras que en el caso de la empresa 2 dicho porcentaje se eleva al 95%.

Conociendo que una determinada turbina cumple las normas de calidad, calcular la probabilidad de que haya sido producida por la empresa 1.

Solución:

Lo primero que haremos será definir los sucesos que entran a formar parte de la probabilidad a calcular. Así, definimos los siguientes:

— A_1: La turbina ha sido producida por la primera empresa.
— A_2: La turbina ha sido producida por la segunda empresa.
— B: La turbina se ajusta a los controles de calidad.

La probabilidad que deseamos calcular es:

$$P(A_1/B)$$

Del enunciado del problema, conocemos:

$$P(A_1) = 0{,}75 \quad P(A_2) = 0{,}25$$
$$P(B/A_1) = 0{,}8 \quad P(B/A_2) = 0{,}95$$

Luego la probabilidad de que una turbina cumpla con los controles de calidad será:

$$P(B) = P(A_1)\,P(B/A_1) + P(A_2)\,P(B/A_2) = 0{,}75 \cdot 0{,}8 + 0{,}25 \cdot 0{,}95 = 0{,}8375$$

Por el teorema de Bayes sabemos que:

$$P(A_1/B) = \frac{P(B/A_1)P(A_1)}{\sum_{i=1}^{2} P(B/A_i)P(A_i)} = \frac{0{,}8 \cdot 0{,}75}{0{,}8375} = 0{,}7164$$

Análogamente, la probabilidad de que fuera de la empresa 2 será:

$$P(A_2/B) = \frac{P(B/A_2)P(A_2)}{\sum_{i=1}^{2} P(B/A_i)P(A_i)} = \frac{0{,}95 \cdot 0{,}25}{0{,}8375} = 0{,}2836 = 1 - P(A_1/B)$$

Las fórmulas de Bayes resultan de utilidad cuando un suceso B puede ocurrir bajo una de las hipótesis $A_1, A_2, \ldots, A_n$, de las cuales tenemos sus probabilidades $P(A_1)$, $P(A_2), \ldots, P(A_n)$ y además bajo las distintas hipótesis A_i, se conoce también la probabilidad $P(B/A_i)$, i = 1, 2, ..., n.

Por tanto, la fórmula de Bayes nos permite el cálculo de las probabilidades de cada una de estas hipótesis A_i, suponiendo que B ha ocurrido.

Ejemplo 8.4. *Se tienen dos urnas con bolas rojas y blancas distribuidas de la siguiente manera:*

A. Urna 1: 3 Rojas y 2 blancas
B. Urna 2: 2 Rojas y 3 blancas

Se extrae una bola de la urna 2 y se introduce en la urna 1. Si al extraer una bola de la urna 1, esta es blanca, ¿Cuál es la probabilidad de que la bola trasladada sea blanca también?

Solución:

Para calcular la probabilidad pedida, definimos los eventos:

B_2: Que la bola trasladada de la urna 2 a la urna 1 sea blanca.
R_2: Que la bola trasladada de la urna 2 a la urna 1 sea roja.
B_1: Que la bola extraída de la urna 1 sea blanca.

Como lo que nos interesa calcular es la probabilidad de que la bola trasladada de la urna 2 a la urna 1 sea blanca, si la bola extraída de la urna 1 es blanca, esta es la probabilidad $P(B_2/B_1)$ que de acuerdo con la definición de probabilidad condicional es:

$$P(B_2/B_1) = \frac{P(B_2 \cap B_1)}{P(B_1)}$$

Para calcular las probabilidades que aparecen en el numerador y denominador de esta expresión, analicemos los datos que suministra el problema.

Las hipótesis que hemos establecido son: que la bola trasladada de la urna 2 a la urna 1 sea blanca (B_2) y que la bola trasladada de la urna 2 a la urna 1 sea roja (R_2).

Bajo esta hipótesis se realiza la extracción de la bola de la urna 1. Las probabilidades de estas hipótesis son:

$$P(B_2)=\frac{3}{5} \quad \text{y} \quad P(R_2)=\frac{2}{5}$$

Para calcular la probabilidad de que la bola extraída de la urna 1 sea blanca, hay que considerar además las probabilidades condicionales $P(B_1/B_2)$ y $P(B_1/R_2)$ ya que trasladar una bola de un color u otro, altera consecuentemente la urna 1.

Calculemos entonces:

$$P(B_1/B_2)=\frac{3}{6}=\frac{1}{2} \quad \text{y} \quad P(B_1/R_2)=\frac{2}{6}=\frac{1}{3}$$

Estamos en condiciones de calcular, utilizando la regla de la probabilidad total, la probabilidad de B_1, que aparece en el denominador de la probabilidad $P(B_2/B_1)$ según pide el problema.

$$P(B_1) = P(B_1/B_2)\,P(B_2) + P(B_1/R_2)\,P(R_2)$$

Sustituyendo los valores correspondientes:

$$P(B_1)=\frac{1}{2}\times\frac{3}{5}+\frac{1}{3}\times\frac{2}{5}=\frac{3}{10}+\frac{2}{15}=0.40$$

Y la probabilidad del numerador será:

$$P(B_2 \cap B_1) = P(B_1/B_2)\,P(B_2) = 0{,}3$$

Luego:

$$P(B_2/B_1)=\frac{P(B_2 \cap B_1)}{P(B_1)}=\frac{0.3}{0.4}=0.75$$

8.4. EJERCICIOS

Sobre probabilidad

Ejercicio 8.1. *Defina el espacio de resultados de los siguientes experimentos:*

a) *Tirar un dado.*

b) *Tirar una moneda dos veces. Llamemos* C = *cara y* S = *cruz.*

c) *Número de llamadas a una centralita telefónica de una compañía aérea durante un día.*

d) *Se extrae una revista de una línea publicitaria de viajes y se la pesa para efectuar un control. Se sabe que el peso no debe exceder de los 250gr.*

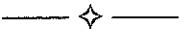
Respuesta

a) $S = \{1, 2, 3, 4, 5, 6\}$.

b) $S = \{CC, CS, SC, SS\}$.

c) $S = \{0, 1, 2, 3, \ldots\}$.

d) $S = \{x \in \Re \,/\, 0 \leq x \leq 250\}$.

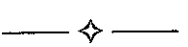


Ejercicio 8.2. *Se está llevando a cabo una auditoría de los servicios de tres nuevas compañías de alquiler de automóviles, seleccionados dentro de cierta área geográfica del país. Cada compañía auditada se marca con una* H *si todas las quejas sobre el servicio fueron resueltas en menos de dos días y se marca con una* T *en caso contrario. Defina el espacio muestral.*

Respuesta

Los resultados posibles son:

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$



Ejercicio 8.3. *Dados los siguientes experimentos aleatorios, indique el espacio muestral para cada uno:*

a) *Se observa la cantidad de vuelos suspendidos durante una semana.*

b) *Se lanza un dado dos veces.*

c) *Se lanza una moneda hasta que salga cara* (c = *cara*, s = *cruz*).

Respuesta

a) $S = \{0, 1, 2, 3, 4, \ldots\}$.

b) $S = \{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)(2,1)(2,2) \ldots (6,1)(6,2)(6,3)(6,4)(6,5)(6,6)\}$.

c) $S = \{c, sc, ssc, sssc, ssssc, sssssc, \ldots\}$.

Ejercicio 8.4. *Identifique los siguientes sucesos:*

a) *Al tirar un dado:* A = *que salga un número par;* B = *que salga un número mayor o igual que 3.*

b) *En el resultado del ejercicio 8.2.:* A = *se obtiene exactamente una* H; B = *no se obtiene ninguna* H.

Respuesta

a) $A = \{2, 4, 6\}$.

$B = \{3, 4, 5, 6\}$.

b) $A = \{HTT, THT, TTH\}$.

$B = \{TTT\}$.

Ejercicio 8.5. *Al instalar sistemas de ordenadores, se ha observado que el 16% de los equipos recién fabricados presentan exactamente un defecto, el 4% tiene exactamente dos defectos y el 1% tiene exactamente tres o más defectos. ¿Cuál es la probabilidad de que un equipo seleccionado al azar no tenga ningún defecto?*

Respuesta

El suceso «0 defectos», tiene por complementario el suceso «1 o más defectos». Luego

$$P(\text{0 defectos}) = 1 - P(\text{1 o más defectos})$$

Por otro lado:

$$P(\text{1 o más defectos}) = P(\text{exactamente 1 defecto} \cup$$
$$\cup \text{ exactamente 2 defectos} \cup \text{exactamente 3 o más defectos})$$

Como estos sucesos son disjuntos (es decir, tienen intersección vacía), obtenemos:

$$P(\text{1 o más defectos}) = P(\text{exactamente 1 defecto}) +$$
$$+ P(\text{exactamente 2 defectos}) + P(\text{exactamente 3 o más defectos}) =$$
$$= 0{,}16 + 0{,}04 + 0{,}01 = 0{,}21$$

Concluimos que:

$$P(\text{0 defectos}) = 1 - P(\text{1 o más defectos}) = 1 - 0{,}21 = 0{,}79$$

Ejercicio 8.6. *En un grupo de 20 personas hay 8 mujeres y 12 varones; 15 fumadores y 6 personas que no estudian carreras universitarias. Se elige una persona al azar del grupo, halle la probabilidad de elegir:*

a) *Una mujer.*

b) *Un fumador.*

c) *Una persona que estudie una carrera universitaria.*

Respuesta

a) $P(\text{Mujer}) = 8/20 = 2/5$.

b) $P(\text{Fumador}) = 15/20 = 3/4$.

c) $P(\text{Universitario}) = 14/20 = 7/10$.

Ejercicio 8.7. *Considere el experimento aleatorio que consiste en lanzar dos monedas. El espacio muestral es:* S = {CC, CS, SC, SS}. *Se define la variable aleatoria X = número de caras al tirar dos monedas. ¿Qué valores toma* **X***?*

Respuesta

$X(CC) = 2.$

$X(CS) = 1.$

$X(SC) = 1.$

$X(SS) = 0.$

Ejercicio 8.8. *A continuación describa los recorridos de distintas variables aleatorias.*

a) Y*: número de personas con cuenta corriente en cierto banco.*

b) Z*: tiempo transcurrido entre dos llamadas telefónicas consecutivas al centro de reservas de una línea aérea.*

Respuesta

a) $R_Y = \{0, 1, 2, \ldots\}.$

b) $R_Z = \{t \in \Re \,/\, t \geq 0\}.$

Ejercicio 8.9. *Una compañía de alquiler de automóviles posee la siguiente información, correspondiente a la distribución de sus ingresos semanales sobre un total de 500 clientes:*

Ingresos (en euros)	Frecuencia absoluta (n_i)
150-250	40
250-350	80
350-450	230
450-550	100
550-650	30
650-750	20
Total	500

Calcule la probabilidad de que un ingreso por alquiler esté comprendido entre 250 y 350 euros. (Nota: asimile el concepto de probabilidad al de frecuencia relativa).

Respuesta

Ingresos (en euros)	Frecuencia absoluta (n_i)	Frecuencia relativa (h_i)
150-250	40	0,08
250-350	80	0,16
350-450	230	0,46
450-550	100	0,20
550-650	30	0,06
650-750	20	0,04
Total	500	

En consecuencia: P (ganar entre 250 y 350) = 0,16.

Ejercicio 8.10. *La siguiente información corresponde a los alumnos de la Facultad de Ciencias Económicas y Empresariales clasificados por sexo y área de estudio de mayor interés. Esta tabla fue confeccionada a partir de una muestra aleatoria de 837 estudiantes.*

Área de interés	Sexo	
	Varones	Mujeres
Indecisos	51	29
Turismo	268	145
Economía	107	42
Administración	65	21
Impuestos	29	13
Estadística	36	25
Otros	5	1
Total	561	276

Si se selecciona un estudiante al azar, determine cuál es la probabilidad de que:

a) *Sea mujer.*

b) *Le interese el área de Administración.*

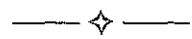
c) *No tenga un área de interés bien definida.*

Respuesta

a) P(mujer) = 276/837 = 0,329.

b) P(le interesa el área de administración) = 86/837 = 0,103.

c) P(no tenga un área definida) = 80/837 = 0,096.



Ejercicio 8.11. *Un curso de turismo internacional tiene 80 alumnos varones y 30 alumnas mujeres. Si se considera el experimento de seleccionar un estudiante al azar de esta clase, determine:*

a) P*(de que sea elegido un varón).*

b) P*(de que sea elegida una mujer).*

c) P*(de que sea elegido un varón o una mujer).*

d) *Si se simboliza como* A = *que sea varón y* B = *que sea mujer, ¿son* A *y* B *sucesos mutuamente excluyentes?*

Respuesta

a) P(varón) = 50/80 = 0,625.

b) P(mujer) = 30/80 = 0,375.

c) P(de que sea elegido un varón o una mujer) = 1.

d) *A* y *B* son mutuamente excluyentes pues si se elige un solo individuo puede ser solamente varón o mujer pero no ambos a la vez.

Ejercicio 8.12. *Un gerente comercial, analizando 90 fichas de clientes, encontró la siguiente información con respecto al hábito de fumar:*

Categoría de fumador	Cantidad de pacientes
Menos de 10	10
Entre 10 y 20	30
Más de 20	50

Sean los sucesos:

A: *Cliente que fuma menos de 10 cigarrillos diarios.*

B: *Cliente que fuma entre 10 y 20 cigarrillos diarios.*

C: *Cliente que fuma más de 20 cigarrillos diarios.*

Calcule:

a) P(A).

b) P(B).

c) P(C).

d) $P(A \cup B)$.

e) $P(A \cup C)$.

f) $P(B \cup C)$.

Respuesta

a) $P(A) = 10/90 = 0{,}111$.

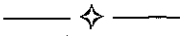
b) $P(B) = 30/90 = 0{,}333$.

c) $P(C) = 50/90 = 0{,}556$.

d) $P(A \cup B) = P(A) + P(B) = 10/90 + 30/90 = 40/90 = 0{,}444$.

e) $P(A \cup C) = P(A) + P(C) = 10/90 + 50/90 = 60/90 = 0{,}667$.

f) $P(B \cup C) = P(B) + P(C) = 30/90 + 50/90 = 80/90 = 0{,}889$.



Ejercicio 8.13. *La inspección de un conjunto de habitaciones seleccionadas por un gerente de recursos humanos de un hotel dio como resultado que 10 habitaciones se ajustaban a las especificaciones requeridas, 4 tenían pequeños defectos y 2 tenían defectos graves.*

Si dentro de ese conjunto, se elige una habitación al azar, ¿cuál es la probabilidad de que:

a) *No tenga defectos.*
b) *Tenga defecto.*
c) *Tenga un defecto grave.*
d) *No tenga defecto o tenga un defecto grave?*

Represente los sucesos de la siguiente manera:

A*: no tenga defecto.*
B*: tenga defectos pequeños.*
C*: tenga defectos graves.*

Respuesta

a) $P(A) = 10/16 = 0{,}625$.

b) $P(B \cup V) = P(B) + P(C) = 4/16 + 2/16 = 6/16 = 0{,}375$.

$P(\text{tenga defectos}) = 1 - P(\text{no tenga defectos}) = 1 - P(A) = 1 - 0{,}625 = 0{,}375$.

c) $P(C) = 2/16 = 0{,}125$.

d) $P(A \cup C) = P(A) + P(C) = 10/16 + 2/16 = 12/16 = 0{,}75$.

Ejercicio 8.14. *Durante un control de calidad de una compañía se observó que en un grupo de 150 empleados analizados 35 presentaban problemas de visión. Si entre todos los empleados, elegimos uno al azar, calcule:*

a) la probabilidad de que presente problemas visuales (V)
b) la probabilidad de que no presente problemas visuales ($\overline{V}$).

Respuesta

a) $P(\text{problemas visuales}) = P(V) = 35/150 = 0{,}233$.

b) $P(\overline{V}) = 1 - P(V) = 1 - 0{,}233 = 0{,}767$.

Ejercicio 8.15. *Una encuesta de opinión, cuyo objetivo era estimar la preferencia de los turistas sobre 3 destinos diferentes durante el verano de 2002 arrojó los siguientes resultados:*

Edad de los encuestados	Destino			
	A	B	C	Total
18-30	58	89	35	182
31-50	80	99	40	219
Más de 50	71	75	30	176
Total	209	263	105	577

Si se selecciona un turista al azar, de entre los 577 encuestados, calcule la probabilidad de que:

a) *El turista prefiera el destino* A.

b) *El turista tenga más de 50 años.*

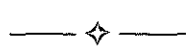
c) *El turista tenga entre 18 y 30 años o prefiera el destino* B.

Respuesta

a) P(prefiera A) = 209 / 577 = 0,362.

b) P(más de 50 años) = 176 / 577 = 0,305.

c) P(tenga entre 18 y 30 años o prefiera B) = P(tenga entre 18 y 30 años) + P(prefiera B) − P(tenga entre 18 y 30 años y prefiera B) = 182 / 577 + 263 / 577 − 89 / 577 = 356 / 577= 0,617.



Ejercicio 8.16. *Un investigador en turismo clasificó a 8.766 clientes considerando su último destino para las vacaciones y el tipo de transporte que utilizó para alcanzar dicho destino. Los resultados obtenidos están resumidos en la siguiente tabla:*

Destino	Montaña	Playa	Otros	Total
T. terrestre (*O*)	983	383	2.892	4.258
T. aéreo (*A*)	679	416	2.625	3.720
T. marítimo (*B*)	134	84	570	788
Total	1.796	883	6.087	8.766

Si en esta muestra se selecciona un cliente al azar, cuál es la probabilidad de que:

a) *Haya usado el transporte tipo* A.

b) *No haya elegido como destino la montaña o la playa.*

c) *Pertenezca al grupo* B *o haya elegido como destino la playa.*

d) *Haya elegido como destino la montaña o pertenezca al grupo* O.

Respuesta

a) $P(A) = 3.720/8.766 = 0{,}424$.

b) $P(\text{otro destino}) = 6.087/8.766 = 0{,}694$.

c) $P(\text{pertenezca al grupo } B \text{ o haya elegido como destino la playa}) = P(\text{pertenezca al grupo } B \cup \text{haya elegido como destino la playa}) = P(\text{pertenezca al grupo } B) + P(\text{haya elegido como destino la playa}) - P(\text{pertenezca al grupo } B \text{ y haya elegido como destino la playa}) = 788/8.766 + 883/8.766 - 84/8766 = 1.587/8.766 = 0{,}181$.

d) $P(\text{haya elegido como destino la montaña o pertenezca al grupo } O) = P(\text{haya elegido como destino la montaña} \cup \text{pertenezca al grupo } O) = P(\text{haya elegido como destino la montaña}) + P(\text{pertenezca al grupo } O) - P(\text{haya elegido como destino la montaña y pertenezca al grupo } O) = 1.796/8766 + 4.258/8766 - 983/8.766 = 5.071/8.766 = 0{,}578$.

Ejercicio 8.17. *Consideremos el siguiente experimento: se tira una moneda, si sale cara se saca una bola de una urna en la que hay 5 bolas negras, 3 azules y 1 verde, si sale cruz se saca una bola de otra urna en la que hay 1 bola negra, 2 azules y 3 verdes.*

a) *Describir el espacio muestral y las probabilidades de cada posible resultado.*

b) *Si se sabe que la bola es azul, ¿cual es la probabilidad de que haya sido obtenida de la primera urna?*

Respuesta

a) Si denominamos C al suceso salir cara al lanzar la moneda y X a salir cruz, y denotamos por N, A y V a los sucesos extraer una bola negra, azul y verde respectivamente:

$$P(N/C) = \frac{P(C \cap N)}{P(C)} \Rightarrow P(C \cap N) = P(C)P(N/C) = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18}$$

De igual modo:

$$P(C \cap A) = P(C)P(A/C) = \frac{1}{2} \cdot \frac{3}{9} = \frac{3}{18}$$

$$P(C \cap V) = P(C)P(V/C) = \frac{1}{2} \cdot \frac{1}{9} = \frac{1}{18}$$

$$P(X \cap N) = P(X)P(N/X) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

$$P(X \cap A) = P(X)P(A/X) = \frac{1}{2} \cdot \frac{2}{6} = \frac{2}{12}$$

$$P(X \cap V) = P(X)P(V/X) = \frac{1}{2} \cdot \frac{3}{6} = \frac{3}{12}$$

b) La solución se deriva de aplicar el Teorema de Bayes. La probabilidad que se pide es la siguiente:

$$P(C/A) = \frac{P(C \cap A)}{P(A)} = \frac{P(C \cap A)}{P(C \cap A) + P(X \cap A)} = \frac{P(C)P(A/C)}{P(C)P(A/C) + P(X)P(A/X)} =$$

$$= \frac{3/18}{3/18 + 2/12} = \frac{1}{2}$$

es decir, si la bola es azul, la probabilidad de haber sido obtenida por ambas urnas es la misma e igual a 0,5.

CONTENIDO DEL DVD

El DVD que complementa este libro pretende que los lectores del libro y, en particular, los alumnos de la asignatura de Introducción a la Estadística de la Licenciatura de Administración y Dirección de Empresas de la UNED dispongan además de información práctica de interés con un doble objetivo:

— Informar de sitios Web donde localizar información estadística y programas freeware de cara a la realización de análisis estadísticos.
— Profundizar desde una perspectiva práctica en los contenidos abordados en el libro, de tal forma que los lectores y/o alumnos puedan realizar los análisis descritos de una manera rápida y sencilla.

Para cubrir el primero de estos objetivos, el DVD incluye un documento con hipervínculos a diferentes páginas web en el directorio «Webs con información y software estadístico— estructurado de la siguiente manera:

— Enlaces a diversas Webs con Información Estadística de interés.
  - Departamentos estadísticos de los Ministerios Españoles.
  - Banco de España.
  - Oficinas estadísticas en comunidades autónomas.
  - Otros organismos e instituciones públicas españoles con información estadística.
  - Otros organismos y empresas con datos para España.
  - Oficinas Estadísticas de la Unión Europea.
  - Oficinas y departamentos estadísticos por Países.
  - Países de la Unión Europea.
  - Resto de Europa.
  - América.
  - Asia.
  - Oceanía.
  - África.
  - Organismos internacionales.
— Enlaces a diversas Webs con Software estadístico.

El segundo de los objetivos se cubre con los siguientes contenidos:

— Por un lado se facilita un programa de manejo muy sencillo (Stdtíska 2009), que se considera de interés principalmente de cara al aprendizaje de las distintas técnicas abordadas en el libro.

— Siete lecciones grabadas en vídeo, donde se muestra de forma visual como realizar las distintas técnicas explicadas en el libro con dos de las herramientas más utilizadas en el ámbito estadístico: la hoja de cálculo Excel de Microsoft y el paquete estadístico SPSS. La visualización de estas lecciones requerirá, obviamente, de un programa de reproducción de video como el Reproductor de Windows Media, Media Player Classic, Real Player, QuickTime, etc. Asimismo, se incluyen en cada lección, los ficheros utilizados como ejemplo en las lecciones, de forma que el lector/alumno pueda practicar las técnicas estudiadas.

El DVD está estructurado en los siguientes directorios:

— Presentación. Incluyendo un vídeo donde se presenta visualmente el contenido del DVD.
— Webs con información y software estadístico.
— Programa Stdístika 2009.
— Lecciones. Este directorio incluye 7 subdirectorios, conteniendo las distintas lecciones.

Por último, de cara a facilitar la navegación por los contenidos del DVD, se ha incluido una aplicación realizada en Macromedia Flash, que se arranca al meter el DVD en la unidad. Esto requiere la instalación previa de Macromedia Flash Player, que puede obtenerse:

http://www.adobe.com/es/shockwave/download/triggerpages_mmcom/flash.html

Si el programa no se ejecutara al insertar el DVD en la unidad, posiblemente por no estar configurado el autoarranque, a través del explorador del sistema operativo, ejecutar el archivo Index.html situado en el directorio raíz.

ESTADÍSTICA PARA ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

Ángel Muñoz Alamillos
Juan Antonio Vicente Vírseda
Azahara Muñoz Martínez

Este manual está dirigido a todos aquellos estudiantes que deban seguir un curso de Introducción a la Estadística o de Estadística Descriptiva en estudios de Economía y de Dirección y Administración de Empresas.

El libro aborda los contenidos teóricos necesarios para comprender y desarrollar los ejercicios planteados y es autosuficiente para superar la asignatura de Introducción a la Estadística en el grado de Administración y Dirección de Empresas siendo el libro recomendado para ello en la Facultad de Económicas de la Universidad Nacional de Educación a Distancia.

Aplica las técnicas estadísticas descriptivas a numerosas situaciones que se plantean en la dirección y gestión de empresas, complementando con un DVD, en el que se incorporan diversos enlaces con páginas Web de interés estadístico, un programa informático de manejo sencillo y la grabación visual de diversas lecciones útiles para aprender como utilizar la hoja de cálculo Excel y el paquete informático SPSS para tratamiento de información estadística; estas lecciones complementan las instrucciones que se dan en el libro para la utilización de estos programas y para la interpretación de los resultados estadísticos obtenidos.

Los autores del libro tienen una dilatada experiencia en la docencia de estas materias y desde hace bastantes años desarrollan su actividad docente en la Universidad Nacional de Educación a Distancia, universidad para la que está especialmente ajustado este manual.

EDICIONES ACADÉMICAS

UNED

ISBN: 978-84-9247

9 788492 477