Manuales docentes de Turismo

Nº 20

Estadística

Eduardo Acosta González

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
Vicerrectorado de Planificación y Calidad

2007

Colección: *Manuales docentes de Turismo*
Nº 20 - Estadística

© del texto:
Eduardo Acosta González

© de la edición:
Vicerrectorado de Planificación y Calidad
Universidad de Las Palmas de Gran Canaria

Primera edición, 2007

Maquetación y diseño:
Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria

ISBN 10: 84-96718-40-9
ISBN 13: 978-84-96718-40-1

Depósito Legal:
GC 39-2007

Impresión:
Servicio de Reprografía, Encuadernación y Autoedición de la ULPGC

Impreso en España. *Printed in Spain*

Índice

Presentación

La Universidad de Las Palmas de Gran Canaria es consciente que la función de una universidad moderna no puede limitar su actividad docente a la enseñanza presencial. Nuestra vocación de servicio en el marco de un contexto geográfico discontinuo y nuestras conexiones con África y América, nos urgen a buscar alternativas para acercar la formación superior a sectores que no pueden cumplir las especificaciones de la enseñanza presencial.

También el Gobierno de Canarias, a través de la Viceconsejería de Turismo, es consciente de la importancia estratégica que tiene para nuestra economía la formación especializada en el sector turístico y que una forma eficaz de lograr ésta es a través de la enseñanza no presencial, dado que permite a muchos profesionales de este sector acceder a estudios superiores de manera compatible con sus horarios laborales.

La formación superior en modalidad no presencial exige materiales docentes de calidad que faciliten los procesos de enseñanza-aprendizaje. Por esta razón, y con la experiencia de 23 manuales editados para la Licenciatura de Psicopedagogía en modalidad no presencial, iniciamos la edición de una colección de manuales docentes que se publican a la vez en formato papel y en soporte electrónico en distintos volúmenes que responden a los contenidos de las asignaturas de Turismo en modalidad no presencial elaborados por profesores de la Universidad de Las Palmas de Gran Canaria.

Estos manuales presentan el mismo diseño instruccional y de publicación que incluye, en primer lugar, la guía académica de la asignatura y desarrolla, posteriormente, cada uno de los módulos con un esquema común que incorpora el índice del módulo, el esquema de la asignatura, los contenidos del módulo, el esquema o mapa conceptual de los contenidos, la exposición de los contenidos, las actividades a desarrollar por los estudiantes, la bibliografía básica para el estudio del módulo y las referencias bibliográficas, los ejercicios de autocontrol y las correspondientes soluciones, un glosario de términos y los anexos.

Queremos expresar nuestro agradecimiento a los autores que han realizado un esfuerzo para elaborar unos materiales rigurosos y adaptados a una nueva forma de enseñar y aprender. Al Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria por su dedicación, diligencia y eficiencia. Y a la colaboración institucional prestada por los departamentos responsables de la docencia en esta titulación y a la Facultad de Ciencias Económicas y Empresariales que han hecho posible la cristalización de este proyecto.

Esperamos que estos manuales docentes sean una herramienta útil para nuestros estudiantes y les ayuden a construir conocimientos significativos. Esta es nuestra apuesta institucional que pretende acercar la formación universitaria a todos los miembros de la sociedad canaria.

Manuel Lobo Cabrera
Rector

Pilar Parejo Bello
Viceconsejera de Turismo

Guía académica

PRESENTACIÓN DE LA ASIGNATURA

La asignatura *Estadística* de la diplomatura en Turismo es una asignatura obligatoria, dirigida a alumnos del 2º curso, con una carga lectiva de 6 créditos. Es la única asignatura de este tipo de contenidos durante la carrera, lo que implica que los instrumentos y herramientas que en ella se explican deben configurar un todo, tanto en su aprendizaje como en su aplicación práctica, ninguna materia puede ser introducida en este curso, pensando en un posterior desarrollo en cursos posteriores. De esta manera, teniendo en cuenta el tiempo de que se dispone, se enfatiza en una aplicación práctica de los conocimientos que se van adquiriendo, a través de metodologías fácilmente comprensibles, que no necesiten elevados mecanismos de abstracción para ser entendidas. Estos desarrollos prácticos ayudan a la compresión de las materias tratadas, gracias a la resolución de ejercicios numéricos y presentación y discusión de casos y aplicaciones.

OBJETIVOS GENERALES DE LA ASIGNATURA

1. El objetivo principal de la asignatura es dotar a los alumnos de unos conocimientos estadísticos que posteriormente puedan utilizar en la toma de decisiones en el desarrollo de su actividad laboral y profesional.
2. Se pretende introducir al estudiante en la parte de la Estadística que se relaciona con la organización y análisis de datos, a través del estudio de una y dos variables, con el fin de que aprenda a describir propiedades del colectivo del que se han tomado gracias al uso de estadísticos descriptivos. Para lograr este objetivo se inicia al estudiante en la recopilación de los datos así como en la forma de ordenar, clasificar y representarlos para, por último, concluir con el estudio de las principales características del colectivo.
3. La última parte de la asignatura se centra en el estudio de las series temporales y los números índices. Con su estudio se persigue que el estudiante sea capaz de determinar los esquemas de dependencia a lo largo del tiempo de variables temporales y como comparar su evolución en el tiempo.

CONTENIDOS

Los contenidos de la asignatura están estructurados en seis módulos. En cada uno de ellos se estudian aspectos relacionados con la estadística desde un punto de vista eminentemente práctico y con una orientación hacia el sector turístico. La denominación de los módulos es la siguiente:

Módulo 1. Introducción a la estadística aplicada al sector turístico.

Módulo 2. Descripción univariante: Tabulación, representación gráfica y medidas de posición.

Módulo 3. Descripción univariante: Medidas de dispersión, de forma y de concentración. Tipificación de una variable.

Módulo 4. Descripción bivariante. Regresión lineal simple.

Módulo 5. Series temporales.

Módulo 6. Números índices.

MÓDULO 1. INTRODUCCIÓN A LA ESTADÍSTICA APLICADA AL SECTOR TURÍSTICO

Se trata de un módulo introductorio. En él se estudia la definición de la *Estadística*, sus usos y el método estadístico. Se presentan conceptos y definiciones claves que se utilizarán en el desarrollo de la asignatura. La diferencia entre *población* y *muestra*, entre *variables cuantitativas* y *variables cualitativas*.

Una clasificación importante que, desde un principio, debe quedar clara al alumno, es la de los diferentes tipos de datos con los que se puede trabajar, ya que en muchas ocasiones las técnicas estadísticas a utilizar están condicionadas a los mismos. El módulo termina con dos epígrafes relacionados con la recogida de información estadística en el ámbito del sector turístico y las fuentes de datos más ampliamente utilizados en esta actividad, tanto a nivel autonómico, nacional e internacional. El contenido del módulo se relaciona a continuación:

1. La Estadística: Definiciones, usos y método.
2. Clasificación de la Estadística.
3. Población y muestra.
4. Variables cuantitativas y variables cualitativas.
5. Clases de datos económicos: corte transversal, serie temporal y datos de panel.
6. Recogida de información en el sector turístico.
7. Fuentes de datos estadísticos del sector turístico.
    7.1. Fuentes nacionales.
    7.2. Fuentes internacionales.
    7.3. Fuentes en al ámbito de la Comunidad Autónoma de Canarias.
    7.4. Obtención de la información mediante encuestas propias.

Módulo 2. Descripción univariante: tabulación, representación gráfica y medidas de posición

Este módulo está dedicado al estudio de cómo describir las características más importantes de la distribución de una determinada variable, especialmente en lo relacionado a sus medidas de posición. Su contenido es el siguiente:

1. Tabulación de una variable.
2. Descripción gráfica.
3. Medidas de posición.
   3.1. Medidas de posición central: la media aritmética y la mediana.
   3.2. Otras medidas de posición: la moda y los cuantiles.

Módulo 3. Descripción univariante: medidas de dispersión, de forma y de concentración

Este módulo, al igual que el segundo, está dedicado al estudio de cómo describir las características más importantes de una determinada variable, pero en esta ocasión, en relación a la dispersión, la forma y concentración de la distribución. Su contenido es el siguiente:

1. Medidas de dispersión: recorrido, varianza, desviación típica y coeficiente de variación.
2. Medidas de forma: el coeficiente de asimetría de Fisher y el coeficiente de curtosis.
3. Medidas de concentración: el índice de Gini y la curva de Lorenz.
4. Tipificación o estandarización de una variable.

Módulo 4. Descripción bivariante. Regresión lineal simple

Este módulo está dedicado a estudiar y evaluar la existencia de relaciones entre dos variables. Su contenido es el siguente:

1. Introducción. Concepto de causalidad.
2. Tabulación. Distribuciones marginales y condicionadas.
3. Representación gráfica.
4. Cuantificación de la relación entre variables.
5. Regresión lineal simple. El método de mínimos cuadrados ordinarios.

Módulo 5. Series temporales

Este módulo está dedicado al estudio de las series temporales. De esta manera se estudiarán las componentes en que se divide el valor de una serie temporal y algunos métodos para su cálculo. El contenido de este módulo es el siguiente:

1. Introducción.
2. Representación gráfica y componentes de una serie temporal.

3. Cálculo de la tendencia. Método analítico.
4. Cálculo de la componente estacional. El método de las medias estacionales sin tendencia.
5. Predicción.
6. Variaciones de las series temporales.

MÓDULO 6. NÚMEROS ÍNDICES

Este módulo está dedicado al estudio de los números índices a través de los cuales se analizan las variaciones, principalmente en el tiempo, de las variables de interés. Su contenido es el siguiente:

1. Introducción. Clasificación de los números índices.
2. Números índices simples.
3. Números índices complejos.
4. Cambio de base y enlace de índices.
5. Deflactación de series temporales.

ESTRATEGIAS DIDÁCTICAS

Guía

El contenido de esta asignatura esta pensado para que el alumno adquiera de forma sencilla, pero sin falta de rigor, técnicas estadísticas que le permitan, en su actual o futura actividad profesional, resolver problemas relacionados con la recopilación y tratamiento de datos cuantitativos. Teniendo en cuanta, que la asignatura de *Estadística* es única en la Diplomatura en Turismo, estas técnicas están desarrolladas hasta un nivel suficiente para ser aplicadas a un entorno real y en ningún caso su aplicabilidad práctica está condicionada a conocimientos que se deban obtener más allá de los que se verán en la asignatura.

Las actividades que el estudiante debe realizar a lo largo del cuatrimestre constituyen un elemento esencial que le permitirá averiguar sus avances en la asimilación de la asignatura. Estas actividades están basadas fundamentalmente en la resolución de casos, donde se plantean determinados problemas y se proponen soluciones relacionadas con los conceptos teóricos estudiados. De esta manera, el estudiante será consciente del potencial de las técnicas estadísticas utilizadas, las más adecuadas en cada caso, y lo límites que las mismas presentan.

MATERIAL DIDÁCTICO

El material básico para la adquisición de los conocimientos requeridos lo constituye el manual preparado para esta asignatura. Además, como complemento se encuentran las referencias bibliográficas relacionadas en el mismo.

Para la realización de gran parte de las actividades y de algunos de los problemas propuestos es necesario el uso de la hoja de cálculo *Excel*. Igualmente, las soluciones a algunos de estos ejercicios se encuentran en este mismo formato. Se ha optado por la utilización de la hoja de cálculo *Excel* para la resolución de este tipo de problemas al tratarse de un programa muy estándar y de uso generalizado. En el módulo 2 se encuentran algunos ejemplos, en los ejercicios de autocontrol, que pueden servir de entrenamiento en el uso de este programa aplicado a la resolución de problemas de estadística. Cada una de las actividades y problemas de autocontrol están señalados mediante los siguientes iconos y significados:

-  →Se recomienda que el ejercicio se haga con el programa *Excel*.
-  →Se recomienda que el ejercicio se haga con lápiz y papel y la ayuda de una calculadora.

En algunas de las actividades planteadas, el estudiante debe conseguir los datos con los que llevar a cabo el trabajo señalado, principalmente a través de las direcciones Webs que se indican en el manual de la asignatura. En estos casos, antes de empezar a trabajar con los datos obtenidos, se debe comunicar al profesor de qué datos se tratan, y obtener su conformidad. En ocasiones puede ocurrir que los datos encontrados no sean válidos para la actividad a desarrollar, y su utilización suponga una pérdida innecesaria de tiempo.

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Lodeiro Hermida, M. J. y Arranz Pérez, M. (1998). *Indicadores estadísticos del sector turístico*. Santiago de Compostela: Asociación Hispalink – Galicia.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: teoría y ejercicios resueltos*. Málaga: Agora.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EVALUACIÓN

La evaluación de la asignatura incluye la asistencia a las sesiones presenciales, la realización de trabajos prácticos, la participación en las actividades en línea (foros de discusión, charlas, aportaciones, iniciativas y propuestas del alumnado) y una prueba escrita presencial:

Parte I. La participación en las actividades en línea y la realización de las tareas programadas aporta el 40% de la nota final y se evaluará a partir de la asistencia a las sesiones presenciales, la participación en las actividades en línea y la realización de los trabajos previstos. La nota oscila entre 0 y 4, y será necesario obtener una nota mínima de 2 puntos para superar esta parte de la asignatura.

Parte II. La parte teórica de la asignatura aporta el 60% de la nota final y se evaluará mediante una prueba escrita que consta de 30 preguntas de respuesta múltiple. La nota del examen oscila entre 0 y 6, y será necesario obtener una nota mínima de 3 puntos para superar esta parte de la asignatura.

La asignatura se considera superada cuando el estudiante ha conseguido los mínimos necesarios (2 puntos en la parte I y 3 puntos en la parte II). Superados los límites señalados anteriormente, la nota final se obtiene de la suma de las puntaciones de la parte I y de la parte II.

Guía

Módulo 1

Introducción a la estadística aplicada al sector turístico

PRESENTACIÓN DEL MÓDULO

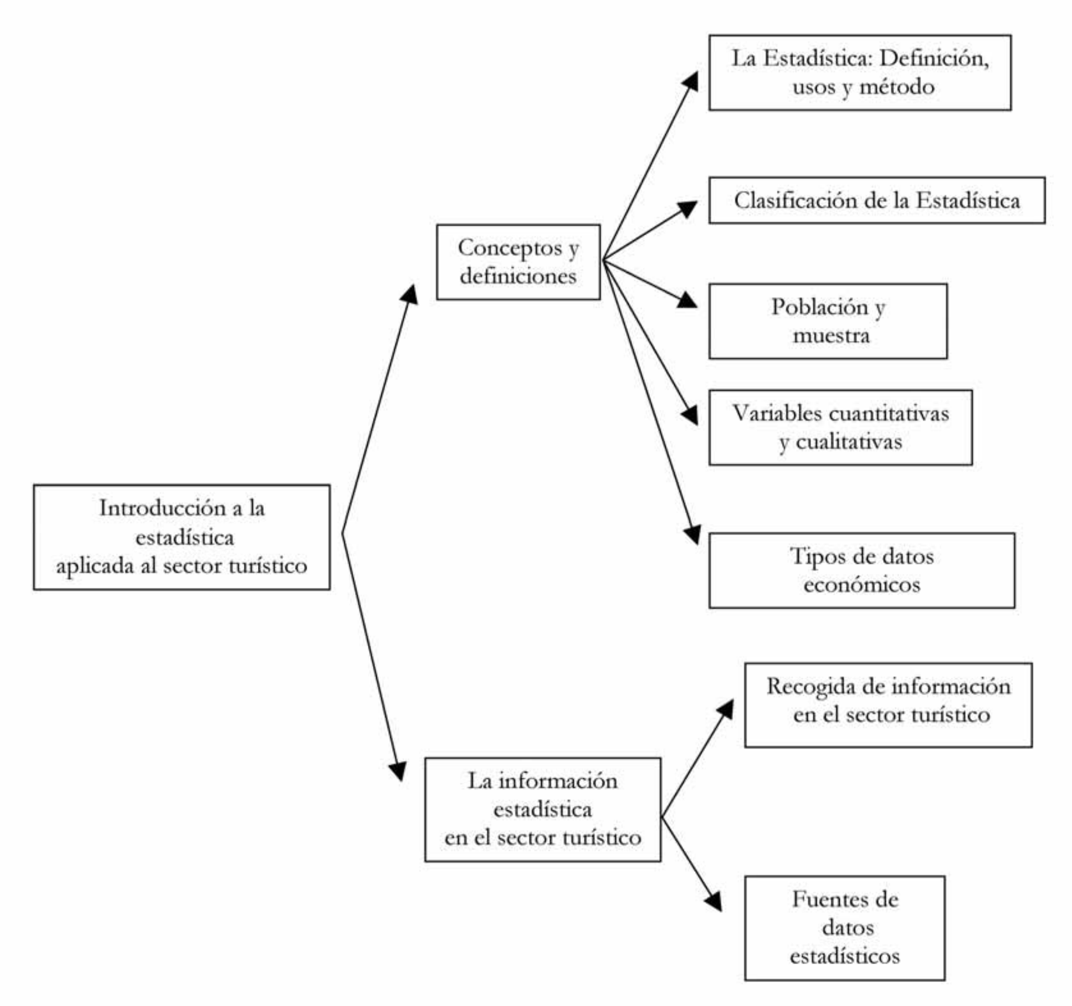
Al tratarse del primer módulo de la asignatura, este módulo tiene un carácter introductorio, en él se verá la definición de Estadística, los usos que en el ámbito de la gestión puede tener esta disciplina, la descripción del método que utiliza, así como la clasificación que, de la misma, habitualmente se realiza. En este módulo también se verán conceptos y definiciones muy importantes para poder entender y contextualizar las herramientas estadísticas, que a lo largo de la asignatura se irán viendo. De este modo, se destacará la diferencia entre los conceptos de población y muestra, la distinción entre variables cuantitativas y cualitativas y los diferentes tipos de datos con los que nos podemos encontrar. Los dos últimos epígrafes del módulo están relacionados con la recogida y fuentes de información en el sector turístico. De esta manera, se estudiarán las características específicas y los problemas que en este sector supone la recogida de información para posteriormente, exponer una relación de las principales fuentes de información que a nivel internacional, nacional y autonómico se pueden consultar a través de Internet. Finalmente, se verán las pautas mínimas imprescindibles que se deben observar en el caso de que la información que se desea obtener se realice directamente mediante encuesta.

OBJETIVOS DEL MÓDULO

En este módulo hay dos objetivos principales que han de alcanzarse para superarlo:

1. Familiarizar al estudiante con conceptos y definiciones básicas de la Estadística, lo que le facilitará la compresión y alcance de las diferentes herramientas que a lo largo de la asignatura se irán viendo.
2. Reconocer la problemática que la recogida de información estadística tiene en un sector como el turístico, al tiempo de ser capaz de planificar y obtener información, tanto de fuentes ajenas como propias, identificando, en el caso de las fuentes ajenas, aquellos organismos que se destacan, tanto a nivel internacional, nacional y autonómico, en la recopilación de este tipo de información. En el caso de las fuentes propias, ha de ser consciente de las características que una encuesta debe tener para que sea susceptible de una posterior explotación con cierta garantía de fiabilidad.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. La Estadística: definiciones, usos y método

Etimológicamente el término Estadística proviene del latín *status*, participio del verbo *stare*, que significa "estar inmóvil", teniendo como sustantivo el significado de "estado, situación".

Según el diccionario de la Real Academia Española, el vocablo Estadística tiene las siguientes acepciones: 1. Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas. 2. Conjunto de estos datos. 3. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

Definir, ¿qué es la Estadística?, supone un elevado riesgo, en tanto en cuanto, implica caracterizar suficientemente el concepto para no confundirlo con otro y, sin embargo, determinar todos aquellos aspectos que lo configuran. Una definición posible sería: Conjunto de métodos y procedimientos necesarios para recoger, representar y resumir datos y hacer inferencia, que con respecto a una realidad concreta medible nos permite el conocimiento cuantitativo de la misma con el fin de determinar su situación.

Esta definición deja abierto el objeto material de la Estadística, en cuanto que sus métodos y procedimientos se aplican "... respecto a una realidad concreta ...", pero sin determinar exactamente cuál. Esto es así, en la medida en que esta disciplina se puede aplicar a diferentes campos como pueden ser el de la Biología, la Meteorología, la Economía, etc., sin embargo, se da una característica de esa realidad, ésta debe ser "medible", por tanto, quedarán fuera del objeto de la Estadística, todas aquellas cuestiones no medibles ya sea directa o indirectamente. También se recoge su objeto formal "...el conocimiento cuantitativo de la misma (realidad) ...", y su fin "...con el fin de determinar su situación". Este fin es genérico y, por tanto, a partir de él, podemos especificar fines más concretos, como la toma de decisiones en ámbitos de incertidumbre, o el de la predicción.

Uno de los usos más importante de la Estadística consiste en servir de herramienta para la toma de decisiones en ambientes de incertidumbre, sin embargo, es importante remarcar, que esta incertidumbre debe poder ser tratada en términos de probabilidades, de tal manera que los diversos acontecimientos o fenómenos que puedan producirse estén unidos a una determinada probabilidad de realización. En un ambiente de incertidumbre total, la Estadística no podría cumplir esta función. Disminuir la incertidumbre en la toma de decisiones, cuantificar las relaciones entre diferentes variables, o la predicción son aspectos en los que la Estadística juega un papel fundamental.

Otro uso de la Estadística está relacionado con la recogida de información. Cualquier actividad empírica que se realice está unida a la utilización de datos, los cuales deben ser obtenidos a partir de la realidad objeto de estudio. Para que la recogida de información sea correcta y, por tanto, los cálculos que se realicen a partir de ella nos sirvan en una posterior toma de decisiones, hemos de utilizar técnicas estadísticas que nos garanticen este resultado. Cada vez más, se incrementa el volumen de información, tanto en el campo de la economía en general como en el de la empresa en particular. Los responsables de la utilización de esta información deben contar con técnicas capaces de manejarla eficientemente, y conseguir, a partir de ella, las características más importantes de la realidad que pretenden estudiar.

Para llevar adelanta cualquier aplicación o uso de la Estadística es muy importante que ésta se lleve a cabo mediante lo que se denomina el método estadístico. Hay que determinar cómo obtener la información necesaria y cómo tratarla rigurosamente para de este modo obtener los aspectos fundamentales que la caracterizan y poder realizar una interpretación adecuada de los resultados. Este proceso se puede describir en las siguientes fases:

1. Objeto de estudio: debe quedar claro cuál es el fenómeno objeto de estudio. Su descripción debe ser precisa, así como la población correspondiente. Por ejemplo, los gestores de un determinado establecimiento turístico pueden tener el objetivo de planificar las actividades de ocio y recreativas entre sus clientes, y para este fin pueden estar interesados en conocer determinadas características de sus clientes como la edad, el sexo, la nacionalidad, etc. En este caso, la población serían los clientes del establecimiento.
2. Selección de las variables: se debe determinar cuáles son las variables que definen el fenómeno objeto de estudio. Para el ejemplo anterior, entre otras, una variable a seleccionar podría ser la edad. El conocimiento de cómo se distribuye esta variable entre los clientes puede ayudar a planificar las actividades de ocio. Las planificación de actividades recreativas y de ocio para clientes de edad muy avanzada, en muchas ocasiones, no serán las mismas que para clientes jóvenes.

3. Obtención de los datos: hay que obtener los valores de las variables que se han seleccionado en el paso anterior. Estos datos se pueden encontrar en bases ya confeccionadas, que se utilizarán como fuente de información o, por el contrario, si éste no es el caso, tendrán que obtenerse directamente de la población, normalmente a partir de la selección de muestras. La selección de las variables, así como la obtención de datos, si éstos se tienen que obtener directamente, son dos fases que deben ser estudiadas con sumo cuidado debido a los costes que implican. La forma en que se selecciona la muestra y el tipo de variables elegidas condicionarán los resultados. Una muestra no significativa implicará resultados erróneos, que más que ayudar a tomar decisiones u obtener conclusiones acertadas, ayudarán a todo lo contrario. Para el caso del ejemplo que estamos siguiendo, la variable *edad de los clientes* puede estar recogida previamente en bases de datos, con lo que ya se tendría solucionada esta parte del proceso. En caso contrario, habrá que realizar encuestas entre los clientes.
4. Resumen de los datos: una vez se han obtenido los datos de las variables de interés, la siguiente fase consiste en realizar un resumen de los mismos, que permita, a partir del cálculo de determinados estadísticos, tener una idea respecto a su distribución. Se estudian aspectos como sus valores centrales, la dispersión, asimetría y relaciones con otras variables. En esta fase se utilizarán principalmente técnicas propias de la Estadística Descriptiva.
5. Inferencia: el Cálculo de Probabilidades aplicado a la Estadística permite hacer inferencia sobre la población de donde se ha obtenido la muestra. En esta fase se procederá a realizar contrastes de hipótesis a partir de las estimaciones de los estadísticos, los cuales deberán cumplir determinadas propiedades para que puedan desempeñar un propósito inferencial.
6. Diagnosis final: a partir de los resultados de los pasos anteriores, se obtendrán las conclusiones finales, las cuales se utilizarán para alcanzar los fines propuestos. En el ejemplo que se ha seguido sobre las actividades de ocio, se determinaría qué actividades son las más adecuadas teniendo en cuenta, entre otras consideraciones, la distribución de la edad de los clientes.

2. Clasificación de la Estadística

Se puede realizar la siguiente clasificación de la Estadística: (1) La Estadística Descriptiva, (2) la Estadística Matemática y (3) la Inferencia Estadística.

La Estadística Descriptiva

El objetivo de la Estadística Descriptiva consiste en la utilización de determinados métodos y técnicas que permiten al investigador, por un lado, recoger información procedente de la realidad que se desea estudiar y, por otro, resumir y describir dicha información mediante la utilización de determinadas medidas y representaciones gráficas.

Los primeros antecedentes de la Estadística los podemos encontrar en los recuentos y recogida de datos que desde muy temprano en la historia se producen en determinadas culturas (Mesopotamia, Egipto), ante la necesidad de conocer la situación de aspectos específicos que afectaban a los miembros de estas sociedades en su vida diaria, como el número de familias, armas, ganado, etc.. Normalmente, los datos recogidos no eran tratados con el objetivo de obtener conclusiones o análisis que fueran más allá de la mera recopilación y recuento de los mismos.

La Estadística Matemática

La Estadística Matemática nace de la integración entre la Estadística y el Cálculo de Probabilidades[1] y estudia el modo de tratar los valores de determinadas variables por medio de modelos probabilísticos. Así, el Cálculo de Probabilidades se configura como una herramienta fundamental en esta parte de la Estadística. Es interesante conocer que la mayoría de los autores consideran que el inicio del Cálculo de Probabilidades está unido a los juegos de azar.

La Inferencia Estadística

Utilizando la Estadística Matemática, la Inferencia Estadística desarrolla un conjunto de métodos y técnicas dirigidas principalmente a la estimación y contrastes de hipótesis, así como a la obtención de muestras, con el ánimo de extender (inferir) los resultados, que de las misma se obtengan, a la población en su globalidad.

3. Población y muestra

La población está constituida por todos los individuos (elementos) que forman parte de la realidad que se pretende estudiar. Por el contrario, la muestra es un subconjunto de la población.

A la hora de realizar un estudio estadístico lo ideal sería trabajar con la población, sin embargo, en la mayoría de ocasiones, como consecuencia fundamentalmente del elevado coste que supone la obtención de datos de toda la población, o porque la propia obtención de un dato implica la eliminación del elemento observado (por ejemplo, vida útil de un bombilla), es necesario realizar el estudio sobre una muestra de la población.

La muestra no se puede obtener de cualquier manera. Es muy importante seguir determinados procedimientos y normas para que el proceso se pueda clasificar como correcto, estadísticamente hablando. Tal y como se vio en la definición de la Inferencia Estadística, una de las funciones de esta parte de la Estadística consiste en elaborar métodos y técnicas que permitan obtener muestras a partir de la población. La forma en que se obtiene una muestra es crucial si se quiere tener seguridad en los resultados de la inferencia. Existen varios procedimientos para la obtención de muestras, y es muy importante saber cuál es el más adecuado en cada caso. Una muestra recogida de forma incorrecta, puede llevar a conclusiones erróneas cuando extrapolamos los resultados obtenidos a la población en su conjunto. Así, por ejemplo, si se desea estudiar determinados hábitos de consumo en un determinado país, y para ello se realiza una encuesta telefónica, si el país en cuestión no tiene el suficiente nivel económico como para que el uso del teléfono esté generalizado, la muestra estará sesgada al no poder ser seleccionada una parte importante de la población.

Uno de los procedimientos más sencillos a la hora de obtener una muestra es el denominado Muestreo Aleatorio Simple. Su característica principal consiste en que cada uno de los elementos de la población tiene la misma probabilidad de ser elegido para formar parte de la muestra.

1 Algunos autores sitúan esta integración a mediados del siglo XIX.

4. Variables cuantitativas y variables cualitativas

La variable es la característica de la población que se desea estudiar. Estas variables pueden ser **cuantitativas** o **cualitativas**. Las variables cuantitativas se expresan mediante un valor numérico, y a su vez se pueden clasificar en **discretas** y **continuas**. Son variables cuantitativas discretas aquellas en las que se puede elegir dos valores entre los cuales no existe ningún otro. Por ejemplo, el número de hermanos en la familia. En este caso, los valores posible de la variable son: 0,1, 2, ..., pero en ningún caso la variable puede tomar valores entre el 1 y el 2. No es posible, por ejemplo, tener 1.5 hermanos. Sin embargo, las variables cuantitativas continuas siempre pueden tomar valores entre dos valores cualesquiera. Ejemplo de variables cuantitativas continuas son: altura de los españoles, edad de los clientes de un hotel, etc.

Por otro lado, las variables cualitativas no son medibles, con lo que a priori no tienen asignado un valor numérico[2], no hace falta asignarles un número para describirlas. Como ejemplo de estas variables tenemos el estado civil, fumar o no fumar, país de procedencia del turista, etc. Las variables cualitativas se expresan en modalidades o atributos mutuamente excluyentes, de esta manera, por ejemplo, la variable cualitativa *sexo* tiene dos modalidades; hombre y mujer. A la variable cualitativa *estado civil* se le puede definir las modalidades: Soltero, casado, divorciado y viudo. O a la variable cualitativa *máximo nivel académico alcanzado* se le puede definir las modalidades: Estudios primarios, estudios secundarios y estudios superiores.

Las variables cualitativas pueden ser **nominales** y **ordinales**. Las variables cualitativas nominales se caracterizan porque no tienen orden, de los ejemplos anteriores, éste sería el caso de las variables *sexo* y *estado civil*. Sin embargo, en las variables cualitativas ordinales sí se puede establecer un orden, éste sería el caso de la variable *máximo nivel académico alcanzado*.

5. Clases de datos económicos: corte transversal, serie temporal y datos de panel

Datos de corte transversal

Todas las observaciones están referenciadas a un mismo instante del tiempo. Por ejemplo, porcentaje de ocupación en el año 2005 de cada uno de los hoteles de cinco estrellas de la isla de Gran Canaria. En este caso, la población estará constituida por todos los hoteles de cinco estrellas de la isla de Gran Canaria, y los datos consistirían en el porcentaje de ocupación de cada uno de ellos. La característica que define a estos datos como de corte transversal es que todos están referenciados al mismo instante del tiempo, en este caso el año 2005, tal y como quedan recogidos en la tabla 1.

2 Es muy habitual, a la hora de trabajar con variables cualitativas, asignarles a sus atributos o modalidades valores numéricos para facilitar su tratamiento estadístico. Sin embargo, la asignación de estos valores numéricos, respetando algunas normas, es bastante arbitrario, y en ningún caso responden al carácter intrínseco que esta asignación supone para el caso de las variables cuantitativas.

Tabla 1. Corte transversal. Porcentaje de ocupación de los hoteles de cinco estrellas de la isla de Gran Canaria

Año	Hotel 1	Hotel 2	...	Hotel N
2005	$P_{1,05}$	$P_{2,05}$	...	$P_{N,05}$

$P_{i,j}$ representa la variable *porcentaje de ocupación* del hotel i en el año j, donde i = 1, 2,…N, siendo N el cantidad de hoteles de cinco estrella en la isla de Gran Canaria. Así, por ejemplo, $P_{1,05}$ corresponde con el porcentaje de ocupación del hotel 1 en el años 2005.

Series temporales

En este caso se tendría la evolución a lo largo del tiempo de una determinada variable. Así, siguiendo con el mismo ejemplo que para el caso de datos de corte transversal, una serie temporal sería el porcentaje de ocupación anual de uno de los hoteles de cinco estrellas de la Isla de Gran Canaria entre el año 1980 y 2005. Evidentemente, para cada uno de los hoteles se podría crear una serie temporal con sus correspondientes porcentajes de ocupación. Una característica muy importante de una serie temporal es que sus valores tiene un orden predeterminado que lo fija el tiempo, característica que no poseía los datos de corte transversal. Siguiendo con el ejemplo de los hoteles de cinco estrellas de la isla de Gran Canaria, en la tabla 2 se presenta la serie temporal para el porcentaje de ocupación del Hotel 2.

Tabla 2. Serie temporal. Porcentaje de ocupación del Hotel 2

Año	Hotel 2
1980	$P_{2,80}$
1981	$P_{2,81}$
⋮	⋮
2005	$P_{2,05}$

Datos de panel

En un panel de datos se tienen tanto datos de corte transversal como de serie temporal. Para el ejemplo que se viene utilizando, el panel de datos contendría los porcentajes de ocupación de cada uno de los hoteles de cinco estrellas de la isla de Gran Canaria para el periodo entre 1980 y 2005. En la tabla 3 se presenta este caso.

Tabla 3. Panel de datos. Porcentaje de ocupación de los hoteles de cinco estrellas de la isla de Gran Canaria

Año	Hotel 1	Hotel 2	…..	Hotel N
1994	$P_{1,80}$	$P_{2,80}$	…..	$P_{N,80}$
1995	$P_{1,81}$	$P_{2,81}$	…..	$P_{N,81}$
…..	…..	…..	…..	
2003	$P_{1,05}$	$P_{2,05}$	…..	$P_{N,05}$

Es importante determinar el tipo de datos con el que se está trabajando, ya que en ocasiones, las técnicas estadísticas que se utilizan son específicas para algunos de ellos. De esta manera, por ejemplo, las técnicas que se verán en la lección 4, tal y como su propio nombre indica, sólo son aplicables a series temporales. Los números índices predominantemente se utilizan en series temporales, aunque pueden existir algunas aplicaciones a datos de corte transversal. Por otro lado, una explotación exhaustiva de datos de panel requiere el uso de técnicas específicas para este tipo de datos.

6. Recogida de información en el sector turístico

Módulo 1

Tal y como se señaló en el apartado 2, la estadística descriptiva, entre otras finalidades, tiene la de utilizar métodos y técnicas que permitan recoger información procedente de la realidad que se desea estudiar. Este objetivo se vuelve especialmente complicado, en una actividad como la turística, debido a la dificultad que supone su medición. Así, por ejemplo, el consumo que realiza un cliente en un determinado restaurante sería considerado turístico cuando el cliente es un turista y como no turístico cuando no lo sea. De esta manera, el carácter de la actividad viene determinado por el tipo de cliente que tenga el negocio en cuestión y evidentemente, en muchos casos, este puede ser mixto.

A esta dificultad se añade la necesidad de buscar significados concretos a conceptos como viajero, turista, residente, excursionista, sobre todo si se desea recoger información de forma homogénea en diferentes partes y por diferentes organismos. Por esta razón, instituciones como la Organización Mundial del Turismo (OMT) y las Organización de Naciones Unidas (ONU) realizan esfuerzos en este sentido, para de esta manera, poder realizar estadísticas que permitan describir la situación de un determinado país en relación a su realidad turística. La importancia, cada vez mayor, que la actividad turística está teniendo en la economía de muchos países y, por tanto, en el escenario mundial, hace necesario disponer de indicadores estadísticos que permitan cuantificar la situación económica real de estos países. Sin embargo, esta no es la situación que se está dando. Otros sectores económicos, a pesar de tener, en muchos casos, una dimensión menor que el turístico, disponen de unas estadísticas que permiten una descripción de los mismos más cercana a la realidad que representan. Esa situación, no sólo se debe al reciente desarrollo que ha experimentado el sector turístico en muchas partes del mundo, sino igualmente a la dificultad que entraña una actividad como la turística, donde intervienen tantos sectores distintos, y con unas características específicas que la diferencia, en muchos aspectos, a otros sectores económicos más tradicionales.

En un intento de construir una estructura conceptual básica a la hora de clasificar las actividades turísticas, la OMT y la ONU, como organismo responsable, han promovido la creación de la Clasificación Internacional Uniforme de Actividades Turísticas (CIUAT)[3]. La CIUAT se enmarca dentro de la Clasificación Internacional Industrial Uniforme de todas las actividades económicas (CIIU)[4], desarrollando dentro de esta clasificación aquellas actividades consideradas turísticas, e indicando los criterios para determinar si una actividad se dedica total o parcialmente al turismo. De esta manera, actividades como la de barcos de crucero (6110-1) o la de agencias de viajes (6304-1) se consideran con una dedicación total a la actividad turística, sin embargo, actividades como la del taxi (6022-1), o la venta al por menor de equipos de caza y pesca (5239-4), se les entienden como actividades de dedicación parcial al turismo.

Igualmente, la importante intervención que en la actividad turística tienen otras actividades de la economía, ha llevado a que la ONU acordara en marzo de 2000 la creación de la Cuenta Satélite del Turismo (CST). El concepto de cuenta satélite se desarrolló con el objetivo de medir actividades económicas que no son reconocidas como industrias en las cuentas nacionales como, por ejemplo, el turismo. De esta manera, mediante la homogenización de concepto y definiciones, se pueden realizar comparaciones de la actividad turística con otras industrias de la economía, o con actividades turísticas y tradicionales de terceros países. La idea fundamental que subsiste en la CST, es identificar y relacionar la demanda y la oferta de bienes y servicios turísticos dentro de una determinada economía, teniendo en cuenta, principalmente, el hecho que de la actividad turística se caracteriza porque es el consumidor quien le da carácter turístico o no a la adquisición de bienes y servicios y, por tanto, una misma actividad puede ser considerada turística o no dependiendo del tipo de consumidor que en cada ocasión demande su producción. De esta manera, se podrá cuantificar para un determinado país, por ejemplo, la contribución del turismo al Producto Interior Bruto, el número de puestos de trabajos que se deben a esta actividad, el consumo turístico, o la recaudación de impuestos que genera.

7. Fuentes de datos estadísticos del sector turístico

Las fuentes de datos estadísticos pueden ser **propias** y **ajenas**. Las propias se obtienen por el interesado directamente mediante encuesta, acudiendo a toda la población cuando es posible, y en caso contrario, mediante la selección de muestras. Tienen una mayor flexibilidad, dado que la recogida de información se puede programar en función de los objetivos de se pretenden obtener. Como contrapartida, su coste suele ser elevado cuando se realiza mediante encuesta específica. En otras ocasiones, sin embargo, las propias actividades empresariales y de gestión generan determinados datos que pueden ser susceptibles de explotación estadística sin grandes costes añadidos. Este tipo de base de datos es cada vez más habitual gracias al desarrollo que los ordenadores y las aplicaciones informáticas han experimentado en los últimos tiempos.

3 En ingles: *Standard International Classification of Tourism Activities* (SICTA). Una relación detallada de esta clasificación de actividades se puede encontrar en la página web de la OMT http://www.world-tourism.org/

4 En ingles: *Internacional Standard Industrial Classification* (ISIC). Una relación detallada de esta clasificación de actividades se puede encontrar en la página web http://unstats.un.org/ de la división de estadística de la organización de las Naciones Unidas.

Por otro lado, la información de las fuentes de datos ajenas suele ser recopilada por instituciones públicas y privadas. Su coste suele ser menor para el usuario final que en el caso de las fuentes propias, y en muchas ocasiones se puede acceder a ellas sin pago alguno. Sin embargo, para bases de datos muy específicas, normalmente es necesario afrontar un canon o precio. Este tipo de bases de datos tienen una menor flexibilidad que las propias, en la medida en que generalmente se obtienen bajo premisas generalistas y, por tanto, sin consideraciones particulares o específicas. A continuación se presentan las principales fuentes de datos que sobre el sector turístico se pueden encontrar en el ámbito nacional e internacional de uso más frecuente.

7.1. Fuentes nacionales

Como fuentes a nivel de todo el territorio nacional se pueden destacar el Instituto Nacional de Estadística (INE) y el Instituto de Estudios Turísticos (IET).

El Instituto Nacional de Estadística (INE)

Sus antecedentes se remontan a mediados del siglo XIX. No es un organismo dedicado exclusivamente a la recogida de información estadística proveniente de la actividad turística, sino que tiene una vocación generalista, sin embargo, entre sus fondos se encuentran importantes bases de datos relacionados con esta actividad. Esta información se puede encontrar tanto en la base INEbase y en la denominada Tempus, si bien desde abril de 2000 ésta última se ha integrado en el sistema de almacenamiento de la primera. Las encuestas de INEbase relacionadas con la actividad turística se relacionan a continuación[5]:

1. Encuesta de Ocupación Hotelera: esta encuesta sustituye desde 1999 a la Encuesta de Movimiento de Viajeros en Establecimientos Hoteleros.
2. Encuesta de Ocupación en Acampamentos Turísticos: este tipo de alojamiento hace referencia a lo que se conoce como *campings*, donde se pueden utilizar como residencia caravanas, tiendas de campaña o similares. Tienen cuatro categorías: lujo, primera, segunda y tercera. Esta encuesta ha sustituido a la Encuesta de Movimientos de Viajeros en Acampamentos. Contiene información sobre la oferta y demanda de este tipo alojamientos.
3. Encuesta de Ocupación de Apartamentos Turísticos: recoge el servicio de alojamiento que presta este tipo de alojamientos.
4. Encuesta de Ocupación en Alojamientos de Turismo Rural: recoge el servicio de alojamiento que presta este tipo de alojamientos.
5. Índice de precios hoteleros e índice de ingresos hoteleros: el índice de precios hoteleros recoge la evolución de los precios que los empresarios hoteleros en España aplican a sus distintos clientes. Mientras que el índice de ingresos hoteleros recoge la evaluación mensual de los ingresos.
6. Encuesta sobre la estructura de las empresas hoteleras: recoge información estructural y económica que caracteriza al sector hotelero.

5 Para más información se puede visitar la página web del INE http://www.ine.es/

7. Encuesta sobre la estructura de empresas de agencias de viajes: recoge información estructural y económica que caracteriza al sector de las agencias de viajes.
8. Entrada de viajeros: dentro de este apartado se encuentran las siguientes estadísticas:
   - Total visitantes.
   - Total turistas.
   - Turistas: Carretera.
   - Turistas: Aeropuertos.
   - Turistas: Ferrocarril.
   - Turistas: Puertos.
   - Excursionistas: Carretera.
9. Ingresos y pagos por turismo: dentro de este apartado se encuentran las siguientes estadísticas:
   - Saldo/Turismo y viajes.
   - Cuenta corriente/Turismo y viajes/Ingresos.
   - Cuenta corriente/Turismo y viajes/Pagos.
10. Albergues y ciudades de vacaciones: dentro de este apartado se encuentran las siguientes estadísticas:
    - Pernoctaciones causadas en albergues juveniles por país y año.
    - Albergues juveniles por CCAA/Provincias, albergues/Plazas y tipo. Año 2000.
    - Edificaciones por ciudades de vacaciones. Año 2001.

Tal y como se ha comentado, el INE también cuenta con una base de datos denominada Tempus donde se pueden encontrar estadísticas relacionadas con el ámbito turístico, concretamente en el Boletín Mensual de Estadística se presentan los siguientes subcapítulos:

- Entrada de visitantes en España.
- Encuesta de ocupación hotelera.
- Personal empleado según categoría hotelera.
- Índices de precios hoteleros por comunidades autónomas.
- Ingresos y pagos por turismo.

Instituto de Estudios Turísticos (IET)[6]

El IET, integrado en la Secretaría General de Turismo, de la Secretaría de Estado de Turismo y Comercio del Ministerio de Industria Turismo y Comercio, fue creado en 1962. En este instituto destacan las siguientes encuestas: Movimientos Turísticos en Frontera (FRONTUR), Encuesta Gasto Turístico (EGATUR) y Movimientos Turísticos de los Españoles (FAMILITUR).

Tanto la encuesta FRONTUR como la EGATUR recogen información relacionado con el turismo receptor. El inicio de la encuesta FRONTUR se sitúa en 1995, y su objetivo es la cuantificación

6 Para más información se puede visitar su página web: http://www.iet.tourspain.es

y caracterización de los flujos de entrada de visitantes por las fronteras españolas. En esta encuesta se puede encontrar la siguiente información:

- Entrada de visitantes por tipología, tipo de dato y años.
- Entrada de turistas por vías de acceso, tipo de dato y años.
- Entrada de turistas por país de residencia, tipo de dato y años.
- Entrada de turistas por CCAA de destino principal, tipo de dato y años.
- Entrada de turistas por motivo de la visita, tipo de dato y años.
- Entrada de turistas por tipo de alojamiento utilizado, tipo de dato y años.
- Visitantes según vía de entrada (Series mensuales).

Por otro lado, en la encuesta EGATUR se puede encontrar la siguiente información:

- Gasto de los visitantes por vías de acceso, tipo de visitante, tipo de valoración, tipo de gasto y tipo de dato.
- Gasto de los turistas por país de residencia, tipo de valoración, tipo de gasto y tipo de dato.
- Gasto de los turistas por destino principal, tipo de valoración, tipo de gasto y tipo de dato.
- Gasto de los turistas por motivo del viaje, tipo de valoración, tipo de gasto y tipo de dato.
- Gasto de los turistas por tipo de alojamiento, tipo de valoración, tipo de gasto y tipo de dato.

La encuesta FAMILITUR esta relacionada con el turismo nacional y recoge datos sobre los viajes realizados por los españoles tanto en España como en el extranjero. Esta encuesta cuenta con la siguiente información:

- Viajes turísticos por forma de turismo, medio de transporte utilizado, tipo de dato y años.
- Viajes turísticos por número y pernoctaciones, CCAA de destino, tipo de dato y años.
- Viajes turísticos por forma de turismo, motivo de la visita, tipo de dato y años.
- Viajes turísticos por forma de turismo, alojamiento utilizado, tipo de dato y años.
- Viajes turísticos por forma de turismo, forma de organización, tipo de dato y años.

7.2. Fuentes internacionales

Organización Mundial del Turismo (OMT)

La OMT tiene sus antecedentes en el Congreso Internacional de Asociaciones Oficiales de Tráfico Turístico, celebrado en 1925 en La Haya. En 1969 la Asamblea General de las Naciones Unidas aprueba una resolución en virtud de la cual la Unión Internacional de Organizaciones Oficiales de Turismo (UIOOT) se transformaba en una organización intergubernamental. Posteriormente la UIOOT cambia su nombre por el de Organización Mundial de Turismo (OMT). En 1976 la secretaria de la OMT se traslada de Ginebra a Madrid.

Dentro de la OMT existe la Sección de Estadística y Evaluación Económica del Turismo. Su objetivo principal es determinar, mediante datos estadísticos, la importancia del turismo y su incidencia en la economía mundial. Igualmente, tuvo un papel protagonista en la creación de la Cuenta Satélite del Turismo, mencionada en el apartado anterior. La OMT es un organismo fundamental a la hora de determinar las normas internacionales de evaluación de datos relacionados con la actividad turística, de tal manera que sus recomendaciones son adoptadas por las Naciones Unidas en 1993, con el objetivo de homogenizar las estadísticas, y con ello, poder llevar a cabo análisis comparativos entre diferentes destinos turísticos.

Actualmente, la OMT cuenta con las bases de datos estadísticos de la actividad turística más completas a nivel mundial. Estas bases de datos estadísticos tienen acceso desde Internet[7], ofrece información sobre más de 200 países, desde el año 1985. Normalmente, la disponibilidad de estas bases de datos se realiza mediante suscripción, y entre ellas se pueden destacar las siguientes: El Anuario Estadístico de Turismo y el Compendio de Estadísticas de Turismo. Otras publicaciones de interés son: La Cuenta Satélite de Turismo como un proceso continuo: pasado, presente y futuros desarrollos, documentos Enzo Paci sobre la evaluación de la importancia económica del turismo, vol 1 y 2 y la evaluación del gasto de los visitantes en el turismo receptor.

7.3. Fuentes en al ámbito de la Comunidad Autónoma de Canarias

La Comunidad Autónoma de Canarias es en la actualidad uno de los destinos turísticos más importantes, tanto a nivel nacional como internacional, con una entrada de visitantes homologables a otros destinos mundiales de referencia[8]. La principal fuente de información pública de la actividad turística la constituye el Instituto Canario de Estadística (ISTAC)[9]. También se puede encontrar estadísticas del sector turístico de las Islas Canarias en la página web de la Consejería de Turismo del Gobierno de Canarias[10].

Las estadísticas del sector turístico con las que cuenta el ISTAC están clasificadas según los siguientes temas: Infraestructura turística, indicadores de actividad turística, pasajeros y turistas, alojamiento turístico, expectativas hoteleras y gasto turístico.

7.4. Obtención de la información mediante encuestas propias

Es ocasiones podemos estar interesados en el estudio de un fenómeno o tema de interés y, sin embargo, no tener la información ajena necesaria para llevar a cabo dicho estudio. Este tipo de situaciones pueden ser muy habitual en un contexto de explotaciones empresariales y particularmente en el caso de explotaciones turísticas. Así, podríamos estar interesados en lo que piensan nuestros clientes en relación a los productos y servicios que les ofrecemos, cómo se nos valora respecto a destinos alternativos, qué servicios son susceptibles de mejora, cuál es la procedencia de nuestros clientes, etc. En otras ocasiones, nuestro objetivo de interés puede ir más allá del ámbito de una única explotación y estar interesado en el comportamiento de un sector (restaurantes, actividades recreativas, etc.), de una determinada área geográfica. En estos y otros muchos casos, donde la información

7 Dirección página web: http://www.world-tourism.org/

8 En 2002 la entrada visitantes en las Islas Canarias, según el Instituto de Estudios Turísticos, fue de 10.8 millones. Esta misma cifra para las Islas Hawai fue de 6.4 millones según el *Department of Business, Economic Development and Tourism* (Estado de Hawai).

9 Dirección página web: http://www.gobiernodecanarias.org/istac/

10 Dirección página web: http://www.gobcan.es/turismo/

normalmente no puede ser obtenida directamente de las fuentes que ya hemos comentado en los apartados anteriores, la encuesta propia se convierte en un instrumento imprescindible.

Una de las cuestiones más importantes a la hora de realizar una encuesta es determinar con claridad cuál es la población objeto de estudio. En ocasiones la población es demasiado extensa como para poder entrevistar a todos y cada uno de los elementos que la componen, en estos casos se opta por la selección de muestras. Es muy importante, si se quiere que los resultados de la muestra se infieran (extrapolen) al resto de la población con garantías de éxito, que ésta sea representativa de dicha población. Para que esto sea así, es necesario que las técnicas de muestreo utilizadas sean correctas.

Entre los diferentes tipos de muestreo podemos destacar los denominados probabilísticos, éstos se caracterizan por el hecho de que se conoce la probabilidad de que cada uno de los elementos de la población sea seleccionado para formar parte de la muestra, lo que implica la posibilidad de realizar inferencia a partir de los resultados muestrales obtenidos. Dentro de este tipo de muestreo podemos a su vez distinguir entre:

1. Muestreo aleatorio simple: todos los individuos de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra. Este tipo de muestreo se puede llevar a cabo con reemplazamiento y sin reemplazamiento. En el primer caso, una vez que un individuo es elegido para formar parte de la muestra, vuelve a la población de la que fue extraído. En el segundo caso, sin embargo, una vez extraído un individuo de la población, éste no vuelve a la misma, y por tanto, aumenta la probabilidad de ser seleccionado para los individuos que quedan en la población. En la práctica, cuando la población es muy grande no existe prácticamente diferencia entre un caso y otro.

2. Muestreo estratificado: en este tipo de muestreo se divide previamente la población en conjuntos o estratos a partir de un determinado criterio (nacionalidad, estado civil, sexo, etc.), de tal manera que se mantiene cierta homogeneidad dentro de los estratos y heterogeneidad entre ellos. Con este tipo de muestreo se consigue información adicional para cada uno de los estratos y, por otro lado, si las proporciones de los individuos, dentro de cada estrato, es igual al que se encuentra en la población, lo que debe ser un objetivo en este tipo de muestreo, disminuyen los errores muestrales y aumenta la precisión de las estimaciones.
3. Muestro por conglomerado: este tipo de muestreo suele realizarse cuando el acceso individualizado a cada uno de los elementos de la población es muy complicado o costoso si se aplicara directamente un muestreo aleatorio simple. Para evitar este inconveniente se divide la población en conglomerados, los cuales están caracterizados porque los individuos que los componen están cerca entre sí. Una vez se realiza esta división, se seleccionan los conglomerados mediante muestreo aleatorio simple, y se entrevista a los individuos que los componen.

Una vez se ha determinado el criterio a seguir para la selección de la muestra, el siguiente elemento fundamental a la hora de realizar una encuesta es el cuestionario. En este documento se encuentran las preguntas que debe contestar el encuestado. Estas preguntas pueden ser cerradas o abiertas. En el caso de las cerradas, al entrevistado se le ofrece una serie de alternativas concretas sobre las que ha de elegir, en las abiertas, por el contrario, no cuenta con estas alternativas y puede responder todo aquello que considere. También, dependiendo el tipo de respuesta, las preguntas se pueden clasificar en cuantitativas, cuando las respuestas derivan en una variable con este carácter, y cualitativas, cuando las repuestas derivan en variables cualitativas. Un tipo de pregunta muy habitual en los cuestionarios es la que recoge variables medidas en escalas.

El diseño del cuestionario es uno de los aspectos más importantes a la hora de realizar una encuesta. Su configuración puede condicionar la respuesta de los entrevistados, con lo que éste se debe confeccionar teniendo en cuenta que las preguntas no sean ambiguas y supongan contestaciones claras y concisas.

ACTIVIDADES

A partir de las fuentes de información del sector turístico que aparecen relacionadas en este módulo, elige algunas variables que consideres de interés y confecciona y obtén una base de datos para una variable de corte transversal, otra para una serie temporal y otra para un panel de datos. (Los resultados pueden ser presentados en hojas de cálculo *Excel*).

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Lodeiro Hermida, M. J. y Arranz Pérez, M. (1998). *Indicadores estadísticos del sector turístico*. Santiago de Compostela: Asociación Hispalink – Galicia.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EJERCICIOS DE AUTOCONTROL

Determina cuáles de las siguientes afirmaciones son verdaderas. Ten en cuenta que en cada una de las preguntas pueden haber más de una afirmación verdadera.

1. La Estadística: definición, usos y método:
   a. La Estadística es un conjunto de métodos y procedimientos necesarios para recoger, representar y falsear datos.
   b. La predicción es un fin de la Estadística.
   c. Cualquier realidad que intente estudiar la Estadística no necesariamente debe ser medible.
   d. La Estadística es una disciplina que nos puede ayudar a la hora de afrontar situaciones de total incertidumbre.

2. La Estadística: definición, usos y método:
   a. La recogida de información es un de los usos de la Estadística.
   b. La utilización del método estadístico nos garantiza tomar decisiones correctas en ambientes de incertidumbre.
   c. Cuando vamos a utilizar métodos estadísticos para estudiar una determinada realidad, el objetivo del estudio no es necesario especificarlo al inicio del trabajo, sino que a medida que éste avanza se va concretando.
   d. La selección de las variables que nos ayudarán a estudiar un determinado fenómeno es de vital importancia si queremos aplicar herramientas estadísticas en nuestro estudio.

3. Clasificación de la Estadística:
   a. La Estadística Descriptiva es una de las disciplinas en las que se puede clasificar la Estadística.
   b. Cuando en la Estadística Descriptiva se utiliza un gran contenido matemático, ésta pasa a denominarse Estadística Matemática.
   c. La Inferencia Estadística sirve principalmente para estimar, hacer contrastes de hipótesis y obtener muestras.
   d. Los primeros antecedentes de la Estadística los podemos encontrar a lo largo del siglo XX cuando se crean los primeros institutos de estadística.

4. Población y muestra:
   a. La población está constituida por todos los elementos que forman parte de la realidad que se pretende estudiar.
   b. La muestra está constituida por todos los elementos que forman parte de la realidad que se pretende estudiar.
   c. La población española se utiliza para todos los estudios que, relacionados con España, se puedan realizar.
   d. Si es posible, es preferible trabajar con la población que con una muestra de la población.

5. Variables cuantitativas y cualitativas:
   a. Las variables cuantitativas se expresan mediante un valor numérico.
   b. Las variables cuantitativas se pueden clasificar en discretas y continuas.
   c. La variable *número de hijos de las familias* es una variable cuantitativa discreta.
   d. Las variables cuantitativas discretas se caracterizan porque el entrevistador debe ser discreto con la contestación de los entrevistados, por el carácter personal de las preguntas.

6. Variables cuantitativas y cualitativas:
   a. A una variable estadística se le dice discreta cuando entre dos valores próximos puede tomar a lo sumo un número finito de valores.
   b. Cuando el estudio que estamos realizando afecta a un fenómeno de naturaleza cualitativa la idea de atributo sustituye al de variable.
   c. A una variable estadística se le dice continua cuando las conclusiones que sacamos de su estudio no son concluyentes y, por tanto, pueden continuar en el futuro.
   d. Los estudios estadísticos sobre variables de naturaleza cualitativa no son tan rigurosos como los realizados sobre variables cuantitativas.

7. Variables cuantitativas y cualitativas:
   a. Las variables cualitativas se dividen en nominales y ordinales.
   b. Las variables cualitativas nominales son aquellas en las que el encuestado es identificado.
   c. Las variables cualitativas ordinales son aquellas en las que el encuestado no es identificado.
   d. Bajo determinadas circunstancias la variable *sexo* puede ser considerada como una variable cualitativa nominal y en otras como una variable cualitativa ordinal.

8. Clase de datos económicos:
   a. Los datos de corte transversal se caracterizan porque todos los valores de la variable están referenciados al mismo instante del tiempo.
   b. Los datos de corte transversal se caracterizan porque ninguno de los valores de la variable superan un corte o cota establecida.
   c. Los valores de una serie temporal siempre están ordenados de menor a mayor.
   d. Un panel de datos contiene tanto datos de corte trasversal como de serie temporal.

9. Recogida de información en el sector turístico:
   a. Entre otros objetivos, la Organización Mundial de Turismo se dedica a la elaboración de indicadores estadísticos que permiten cuantificar la situación del sector turístico a nivel mundial.
   b. Uno de los problemas de la recogida de información en el sector turístico es que no existe una estructura conceptual básica a la hora de clasificar las actividades turísticas.
   c. Gracias a los cambios conceptuales que está proponiendo la Organización Mundial del Turismo, hoy en día la actividad que realiza un restaurante es considerada turística.

d. Las Cuentas Satélites se desarrollan con el objetivo de medir actividades económicas que no son consideradas como industrias en las cuentas nacionales.

10. Fuentes de datos estadísticos del sector turístico:

a. En España, el organismo público que en exclusiva tiene las competencias para la recogida de información del sector turístico es el Instituto de Estudios Turísticos (IET). Las competencias que sobre esta materia tenía el Instituto de Nacional de Estadística (INE) ha sido transferidas al IET.

b. En el IET destacan las siguientes encuestas: (1) Movimientos de turísticos en frontera (FRONTUR), (2) Encuesta del gasto turístico (EGATUR) y (3) Encuesta de los alojamientos turísticos (EATUR).

c. En el muestreo aleatorio simple, cada uno de los elementos de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra.

d. El muestreo estratificado sólo es posible cuando se realizan encuestas atendiendo al nivel de renta de las familias.

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

1. b
2. a, d
3. a, c
4. a, d
5. a, b, c
6. a, b,
7. a
8. a, d
9. a, b, d
10. c

GLOSARIO DE TÉRMINOS

Corte transversal: conjunto de datos referenciados al mismo instante de tiempo.

Cuestionario: instrumento donde se encuentran las preguntas y se recogen las respuestas de las encuestas.

Encuesta: recogida de información tanto de elaboración propia como ajena.

Estadística: conjunto de métodos y procedimientos necesarios para recoger, representar y resumir datos y hacer inferencia, que con respecto a una realidad concreta medible nos permite el conocimiento cuantitativo de la misma con el fin de determinar su situación.

Estadística descriptiva: disciplina de la Estadística dirigida principalmente a la recogida, resumen y descripción de la información.

Estadística matemática: disciplina de la Estadística que estudia el modo de tratar la información por medio de modelos probabilísticos.

Estadístico: medida que se obtiene como función de datos muestrales. Ejemplo de estadísticos son la media, la mediana, etc.

Estimación: asignar como valor de un parámetro poblacional el obtenido a partir de una muestra.

Inferencia: extrapolar los resultados obtenidos de una muestra a toda la población.

Muestra: subconjunto de elementos de una población.

Muestreo aleatorio simple: muestreo en el que todos los individuos (elementos) de la población tienen la misma probabilidad de ser elegidos para formar parte de la muestra.

Muestreo estratificado: tipo de muestreo donde previamente a la elección de los individuos (elementos) se divide la población en conjuntos o estratos.

Muestreo por conglomerado: tipo de muestreo donde los individuos se incluyen en conglomerados y éstos son seleccionados por muestreo aleatorio simple. Todos los individuos de los conglomerados seleccionados forman parte de la muestra.

Panel de datos: conjunto de datos de corte transversal y de serie temporal para un mismo grupo de individuos.

Población: todo el conjunto de individuos que tienen la característica objeto de estudio.

Serie temporal: conjunto de datos de una variable ordenados en el tiempo.

Variable cualitativa: son variables no medibles, con lo que a priori no tienen asignado un valor numérico.

Variable cualitativa nominal: son variables cualitativas que no tienen un orden predeterminado.

Variable cualitativa ordinal: son variables cualitativas que tienen un orden predeterminado.

Variable cuantitativa: son variables medibles por lo que se expresan mediante un valor numérico.

Variable cuantitativa continua: son variables que siempre pueden tomar valores entre dos valores cualesquiera.

Variable cuantitativa discreta: son variables en las que se pueden elegir dos valores entre los cuales no existe ningún otro.

Módulo 2

Descripción univariante: Tabulación, representación gráfica y medidas de posición

PRESENTACIÓN DEL MÓDULO

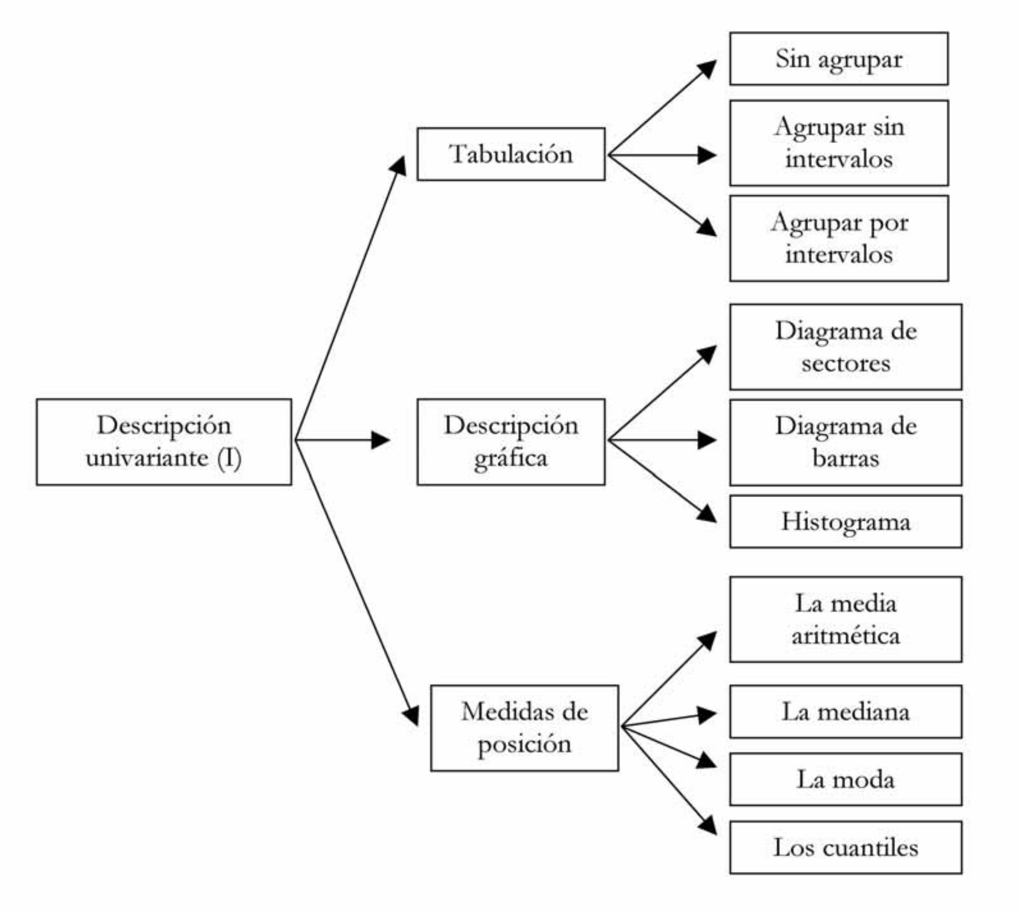
Cuando nos enfrentamos a una relación de valores de una determinada variable, como puede ser: el porcentaje de ocupación de un establecimiento turístico, la entrada de turistas a través de un aeropuerto, la oferta de camas de los hoteles de una ciudad, etc., difícilmente podremos sacar conclusiones globales sobre la misma si no contáramos con instrumentos estadísticos que nos permitiesen describir y resumir la información. Este proceso descriptivo se realiza mediante representaciones gráficas y la obtención de **estadísticos** o **medidas** a partir de los valores que toma la variable. Ejemplos de estos estadísticos o medidas son la media aritmética y la mediana. En este módulo veremos cómo tabular los datos de una variable, cómo representarla gráficamente y cómo obtener **estadísticos** descriptivos de posición.

OBJETIVOS DEL MÓDULO

Fundamentalmente, existen tres objetivos en este módulo que han de alcanzarse para que sea superado. Concretamente, estos objetivos son:

1. Tabular los valores de una determinada variable mediante su agrupación por frecuencias absolutas.
2. Describir la distribución de una variable de forma gráfica.
3. Describir la posición de la distribución de una variable mediante la obtención de estadísticos descriptivos.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. TABULACIÓN DE UNA VARIABLE

Antes de iniciar el cálculo de los estadísticos que describen a una determinada variable, es conveniente tabular los valores que la misma toma. La tabulación permite alcanzar dos objetivos:

1. Estudiar las frecuencias o distribución de la variable.
2. Facilitar el cálculo de los estadísticos descriptivos.

Obtención de las frecuencias

Supongamos que estamos interesados en estudiar la edad de los clientes que visitan un determinado hotel para planificar determinadas actividades de ocio. Para llevar a cabo este estudio se entrevistan a 50 clientes seleccionados de forma aleatoria. Los resultados de esta encuesta se presentan en la tabla 1, donde se relacionan las edades de los 50 clientes entrevistados. Al número de individuos u observaciones de la variable se le denomina por la letra N, en este caso N=50.

Tabla 1. Edad de los clientes

19	25	32	44	51	28	32	20	60	45
54	23	25	36	40	45	33	22	55	48
25	40	23	45	33	19	36	22	48	32
25	36	28	19	25	22	32	44	45	55
32	25	60	40	45	33	22	28	44	45

TABULACIÓN DE LOS DATOS MEDIANTE AGRUPACIÓN SIN INTERVALOS

Tabla 2. Agrupación por frecuencias

X_i	n_i	f_i	N_i	F_i
19	3	0.06	3	0.06
20	1	0.02	4	0.08
22	4	0.08	8	0.16
23	2	0.04	10	0.20
25	6	0.12	16	0.32
28	3	0.06	19	0.38
32	5	0.10	24	0.48
33	3	0.06	27	0.54
36	3	0.06	30	0.60
40	3	0.06	33	0.66
44	3	0.06	36	0.72
45	6	0.12	42	0.84
48	2	0.04	44	0.88
51	1	0.02	45	0.90
54	1	0.02	46	0.92
55	2	0.04	48	0.96
60	2	0.04	50	1.00
Total	N=50	1.00		

La tabla 1 recoge directamente los datos que proporciona la encuesta. La tabulación de los mismos, a partir de sus frecuencias, quedaría tal y como se presenta en la tabla 2, donde X_i representa los valores de la variable para i = 1, 2, ... , J, siendo J el número de valores que toma la variable *edad* distintos entre sí, concretamente en nuestro ejemplo J=17[1]. Tal y como se puede observar, la variable X_i está ordenada de menor a mayor. Para el cálculo de determinados estadísticos, el orden en que se encuentra la variable no afecta, sin embargo, para otros sí. Por tanto, y para evitar problemas, siempre es conveniente desde el principio tener la costumbre de ordenar la variable de menor a mayor. El resto de la notación que aparece en la tabla 2 representa a las siguientes definiciones:

La **frecuencia absoluta** viene representada por n_i y se define como el número de veces que se repite un determinado valor de la variable. Así, por ejemplo, el número de clientes encuestados con 19 años de edad es de tres, o con 23 años de edad es de dos. Estos resultados se pueden comprobar en la tabla 1. La suma de todas las frecuencias absolutas es igual a N, o sea,

$$\sum_{i=i}^{J} n_i = N \tag{2.1}$$

La **frecuencia relativa** viene representada por f_i y se obtiene mediante la siguiente expresión:

$$f_i = \frac{n_i}{N} \tag{2.2}$$

La suma de las frecuencias relativas es igual a la unidad. Es habitual presentar la frecuencia relativa multiplicada por cien. En este caso se define como el porcentaje de individuos que toma un determinado valor de la variable. Así, por ejemplo, el 12% de los clientes entrevistados tienen una edad de 45 años.

La **frecuencia absoluta acumulada** se representa mediante N_i, para su cálculo es necesario que la variable esté ordenada de menor a mayor. Se obtiene como la suma acumulada de la frecuencia absoluta hasta un determinado valor de i

$$N_i = n_1 + n_2 + ... + n_i \tag{2.3}$$

Por ejemplo, para i=3, tendremos que $N_3 = n_1 + n_2 + n_3$, que para el ejemplo de la edad de los clientes será: $N_3 = 3 + 1 + 4 = 8$. Se puede comprobar fácilmente que el valor de la última frecuencia absoluta acumulada (N_J) es igual a N.

La **frecuencia relativa acumulada** se representa por F_i. Tiene las mismas características que la frecuencia absoluta acumulada y, por tanto, es necesario ordenar previamente los valores de la variable de menor a mayor. Se obtiene a partir de la frecuencia relativa de la siguiente manera:

$$F_i = f_1 + f_2 + ... + f_i \tag{2.4}$$

Para i=3, tendremos que $F_3 = f_1 + f_2 + f_3$, que para el ejemplo de la edad de los clientes será: $F_3 = 0.06 + 0.02 + 0.08 = 0.16$. Se puede comprobar fácilmente que el valor de la última frecuencia relativa acumulada (F_J) es siempre igual a 1.

1 En el caso de que los datos no estén agrupados i = 1,2,…, N.

TABULACIÓN DE LOS DATOS SIN AGRUPAR

La tabulación que representa la tabla 2 puede realizarse igualmente sin necesidad de agrupar las observaciones por sus frecuencias. Así, por ejemplo, la tabla 2 quedaría tal y como se recoge en la tabla 3. El inconveniente de utilizar una tabla como la 3, es que cuando el número de individuos es elevado, la tabla se hace excesivamente grande. Sin embargo, aún aunque el número de individuos sea elevado, si el número de valores distintos entre sí son pocos, una tabla como la 2 nos permite tener una tabulación más pequeña y, por tanto, más manejable. En nuestro ejemplo, como sabemos, el número de observaciones (individuos) es N=50, y el número de valores distintos entre sí es J=17. Además, la distribución de la variable es más difícil de intuir en una tabla como la 3 al no existir una agrupación de las frecuencias. Otro inconveniente de esta tabla se produce a la hora de calcular los estadísticos descriptivos de la variable sin la ayuda de un ordenador. En este caso, los cálculos se simplifican enormemente cuando se trabaja con una tabla como la 2. Obviamente este inconveniente no surge cuando los estadísticos descriptivos se calculan mediante la utilización de *softwares* informáticos. Cuando un *software* lo permite, es indiferente introducir los datos agrupados, como en la tabla 2, o sin agrupar, como en la tabla 3. En cualquier caso, es importante tener claro, que a efectos del resultado final, es indistinto utilizar una u otra tabla.

Tabla 3. Tabulación sin agrupación

X_i	n_i	f_i	N_i	F_i
19	1	0.02	1	0.02
19	1	0.02	2	0.04
19	1	0.02	3	0.06
20	1	0.02	4	0.08
22	1	0.02	5	0.10
22	1	0.02	6	0.12
22	1	0.02	7	0.14
22	1	0.02	8	0.16
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
55	1	0.02	47	0.94
55	1	0.02	48	0.96
60	1	0.02	49	0.98
60	1	0.02	50	1
Total	N=50	1		

TABULACIÓN DE LOS DATOS MEDIANTE LA AGRUPACIÓN POR INTERVALOS

En ocasiones, aún cuando los datos se agrupen tal y como se recoge en la tabla 2, el tamaño de la misma es excesivamente larga debido a que J es grande. En estos casos, la obtención de las frecuencias se puede facilitar mediante la agrupación por intervalos de los valores de la variable. De esta manera se consigue reducir la dimensión de la tabla. Sin embargo, este tipo de agrupación supone pérdida de información.

Supongamos que estamos interesados en estudiar al personal empleado en las provincias españolas en el sector turístico. Para ello se cuenta con los datos correspondientes a diciembre de 2002 recogidos en la tabla 4.

Tabla 4. Empleados en el sector turístico por provincias españolas en diciembre de 2002

Álava	427	Castellón/Castelló	1144	Lleida	1788	Segovia	627
Albacete	359	Ciudad Real	497	Lugo	620	Sevilla	3642
Alicante/Alacant	6591	Córdoba	1141	Madrid	10745	Soria	338
Almería	1466	Coruña (A)	1965	Málaga	9058	Tarragona	1713
Asturias	1861	Cuenca	367	Murcia	1665	Teruel	637
Ávila	437	Girona	2290	Navarra	937	Toledo	852
Badajoz	939	Granada	2884	Ourense	421	Valencia/ Val ncia	3226
Balears (Illes)	4967	Guadalajara	368	Palencia	390	**Valladolid**	**952**
Barcelona	9550	Guipúzcoa	1249	Palmas (Las)	13204	Vizcaya	1164
Burgos	802	Huelva	857	Pontevedra	1841	Zamora	299
Cáceres	1051	Huesca	1248	Rioja (La)	840	Zaragoza	1429
Cádiz	3002	Jaén	714	Salamanca	901	Ceuta	156
Cantabria	1377	**León**	**952**	Santa Cruz de Tenerife	13273	Melilla	91

Fuente: Instituto Nacional de Estadística

Tal y como se puede comprobar en la tabla 4, el único valor que se repite es el de las provincias de León y Valladolid, esto significa que una agrupación como la que se obtuvo en la tabla 2 no tendrá los efectos esperados en la reducción de la dimensión de la tabla, dado que N=52 y J=51. En este caso sólo es posible reducirla mediante la agrupación por intervalos, tal y como se observa en la tabla 5.

Tabla 5. Empleados en el sector turístico por provincias españolas agrupados por intervalos

X_i	n_i	f_i	N_i	F_i
[0, 1000)	25	0.48.08	25	48.08
[1000, 2000)	15	0.28.85	40	76.92
[2000, 3000)	2	0.03.85	42	80.77
[3000, 4000)	3	0.05.77	45	86.54
[4000, 5000)	1	0.01.92	46	88.46
[6000, 7000)	1	0.01.92	47	90.38
[9000, 10000)	2	0.03.85	49	94.23
[10000, 11000)	1	0.01.92	50	96.15
[13000, 14000)	2	0.03.85	52	100.00
Total	52	1	52	100.00

A la hora de especificar un intervalo es importante conocer las siguientes definiciones:

1. **Extremos del intervalo**: cada intervalo consta de un extremo inferior y otro superior, que para el intervalo i se representa por E_{i-1} y E_i, respectivamente. Así, por ejemplo, el extremo inferior y superior del intervalo tercero (i=3) de la tabla 5 son iguales a $E_2 = 2000$ y $E_3 = 3000$, respectivamente. Tal y como se puede comprobar, el extremo superior de un determinado intervalo coincide con el extremo inferior del siguiente. Esto supone un problema a la hora de determinar en qué intervalo iría aquella observación que coincida con un extremo. Para evitar esta situación, consideraremos que los intervalos están cerrados por el lado inferior y abiertos por el superior. Así, en el ejemplo que estamos siguiendo, en el caso de que el número de empleados en el sector turístico de una determinada provincia fuera de 2000, este dato se incluiría en el tercer intervalo y no en el segundo.
2. **Amplitud del intervalo**: es la diferencia entre el extremo superior e inferior del intervalo. Se representa por A_i. Por tanto, $A_i = E_i - E_{i-1}$. La amplitud de los intervalos puede ser constante o variable. En el caso de los datos tabulados en la tabla 5, los intervalos son de amplitud constate, concretamente $A_i = 1000$ para todo i.
3. **Marca de clase**: es el valor que divide en dos partes iguales el intervalo. La marca de clase del intervalo i se denota por a_i y se obtiene mediante la expresión:

$$a_i = \frac{E_{i-1} + E_i}{2} \qquad (2.5)$$

La marca de clase se utiliza como valor de la variable a la hora de obtener las diferentes estadísticas o medidas descriptivas.

El número de intervalos depende de la amplitud de éstos, así como del recorrido de la variable[2]. Además, la pérdida de información que implica la agrupación por intervalos, tal y como se ha comentado, depende igualmente de la amplitud de éstos. Entre mayor es la amplitud, mayor es la pérdida de información en la que se incurre.

2 Se entiende como por recorrido de una variable a la resta ente el valor máximo y mínimo que toma la misma.

2. Descripción gráfica

Las representaciones gráficas nos ayudan a describir el comportamiento de las variables, principalmente en relación a su distribución, y por tanto, son un instrumento complementario al uso de los estadísticos descriptivos que estudiaremos a lo largo del curso. Existen una gran variedad de tipos de gráficos que pueden utilizarse, sin embargo, para simplificar la exposición de este epígrafe nos centraremos en los más habituales, concretamente en el diagrama de sectores, en el diagrama de barras y en el histograma. Las dos últimas representaciones dependerán de si se agrupan o no por intervalos los valores de la variable. En el caso de no agrupar por intervalos se utilizaría el diagrama de barras, en caso contrario se utilizará el histograma.

Diagrama de sectores

Normalmente se dibuja como un círculo con tantos sectores como valores toma la variable que se estudia. Cada uno de los sectores tiene una extensión proporcional a la frecuencia del valor de la variable que representa. Supongamos que se está realizando un estudio sobre la nacionalidad de los turistas que visitan un determinado destino turístico. Si los resultados de la encuesta que se realiza son: 20 alemanes, 12 británicos, 7 suecos y 3 de otras nacionalidades, el diagrama de sectores quedaría tal y como se representa en la figura 1[3].

Figura 1. Nacionalidades

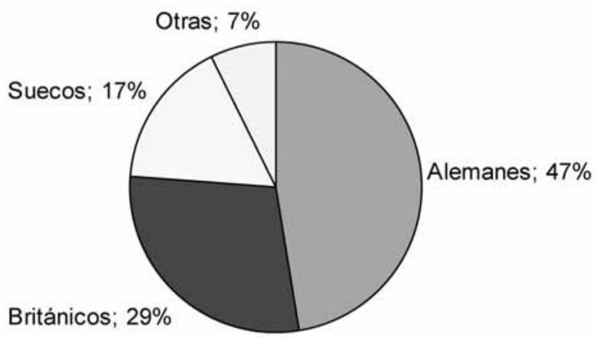
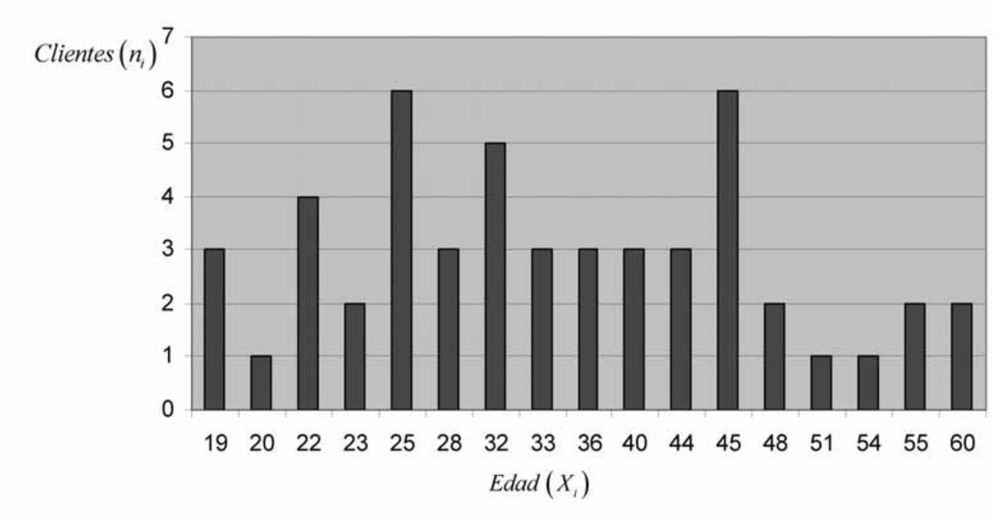


Diagrama de barras

En este tipo de representaciones gráficas se utilizan cuando los valores de la variable no están agrupados por intervalos. En el eje de abscisas se representan los valores de la variable, mientras que en el eje de ordenadas, los valores de las frecuencias absolutas o relativas, según se considere más conveniente. En la figura 2 se presenta el diagrama de barras para los datos de la variable *edad* recogidos en la tabla 2.

3 El cálculo del área que de un determinado sector i, del diagrama de sectores, se obtiene teniendo en cuenta que los grados del ángulo de éste debe ser igual a: $x_i^0 = 3600 \cdot f_i$. Hoy en día es habitual que este tipo de representaciones gráficas se realice con la ayuda de *softwares* informáticos diseñados al efecto.

Figura 2. Diagrama de barras de la variable edad recogida en la tabla 2



HISTOGRAMA

Se utiliza cuando los datos de la variable se agrupan por intervalos. En el eje de abscisas se representan los valores de la variable agrupados en intervalos y en el eje de ordenadas la altura de cada uno de los rectángulos que forman los intervalos, de tal manera que la superficie de cada uno de ellos sea proporcional al valor de la frecuencia absoluta. En la figura 3 se representa el histograma de la variable *ocupados en el sector turístico*, cuyos valores están recogidos en la tabla 5. Tal y como se puede apreciar en dicha figura, la altura de los rectángulos se ha representado mediante h_i. Su valor se obtiene teniendo en cuenta que la superficie de un rectángulo es igual al producto entre su base y su altura, de esta manera se tiene que:

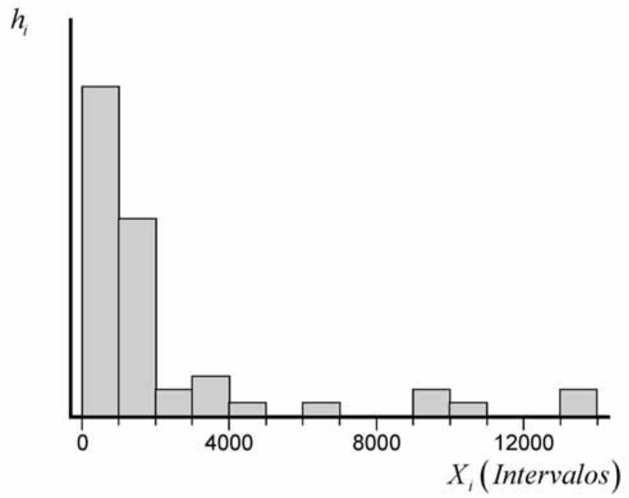
$$n_i = A_i \times h_i$$

donde, si se recuerda,representa la amplitud del intervalo i, que coincide con la base del rectángulo. De la expresión anterior se deduce que la altura se obtiene mediante la expresión:

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

Al igual que para el caso del diagrama de barras, el histograma se puede confeccionar utilizando la frecuencia relativa en lugar de la absoluta. En este caso la superficie de cada una de los rectángulos sería proporcional a la frecuencia relativa.

Figura 3. Histograma de la variable "ocupados en el sector turístico" recogida en la tabla 5



Todos los intervalos de la figura 3 tienen la misma amplitud, recordemos, sin embargo, que esto no tiene que ser así necesariamente. Cuando los intervalos tienen la misma amplitud, la proporcionalidad del histograma se mantiene, aunque se utilicen las frecuencias absolutas o relativas en el eje de ordenadas, en lugar de la altura. Por esta razón, en este caso, es habitual utilizar estas frecuencias a la hora de realizar esta representación gráfica.

Módulo 2

3. Medidas de posición

Para determinar las características básicas que resumen la información de la variable que se pretende estudiar se utilizan determinadas medidas o estadísticos. Cada una de estas medidas o estadísticos nos proporcionan información sobre características concretas. De esta manera, se tendrán medidas que indican posiciones centrales de la distribución de la variable alrededor de los cuales se encuentran sus valores. A estas medidas se les denomina de posición central, entre las que se encuentran la media aritmética y la mediana.

3.1. Medidas de posición central: la media aritmética y la mediana

La media aritmética

Es la medida de posición más utilizada. Se utiliza para determinar la posición central de la distribución. Se obtiene como la suma de todos los valores que ha tomado la variable dividida por el número total de observaciones. Cuando los datos no están agrupados su expresión es:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} \qquad (2.6)$$

o alternativamente en el caso de datos agrupados:

$$\bar{X} = \frac{\sum_{i=1}^{J} X_i n_i}{N} = \frac{X_1 n_1 + X_2 n_2 + X_3 n_3 + \ldots + X_J n_J}{N} \tag{2.7}$$

Además, teniendo en cuenta que $f_i = \frac{n_i}{N}$, igualmente la media aritmética se puede obtener a partir de la siguiente expresión:

$$\bar{X} = \sum_{i=1}^{J} X_i f_i = X_1 f_1 + X_2 f_2 + X_3 f_3 + \ldots + X_J f_J \tag{2.8}$$

Propiedades de la media aritmética

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

$$\sum_{i=1}^{J} \left(X_i - \bar{X} \right) n_i = 0$$

2. Si se crea una nueva variable a partir de la original, sumando una constante a todos sus valores, la media de la nueva variable es igual a la media de la variable original más la constante.

$$Si \quad Z_i = C + X_i \quad \Rightarrow \quad \bar{Z} = C + \bar{X}$$

3. Si se crea una nueva variable a partir de la original, multiplicando por una constante a todos sus valores, la media de la nueva variable es igual a la media de la variable original multiplicada por la constante.

$$Si \quad Z_i = C \times X_i \quad \Rightarrow \quad \bar{Z} = C \times \bar{X}$$

La mediana

La mediana es el valor de la variable que una vez ordenados sus valores de menor a mayor divide el número de observaciones o individuos en dos partes iguales. Supongamos que tenemos los datos de la tabla 6, correspondientes al número de clientes que han entrado en el buffet de desayuno de un determinado hotel a lo largo de una semana.

Tabla 6. Número de clientes que han entrado al buffet de desayuno

Día	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Clientes	30	27	30	31	27	28	28

Si los datos no están agrupados por frecuencias absolutas, la mediana se obtendría mediante los siguientes pasos:

1. Se ordenan los valores de la variable de menor a mayor, tal y como aparecen en la tabla 7.

Tabla 7. Número de clientes que han entrado al buffet de desayuno ordenados de menor a mayor

Día	Martes	Viernes	Sábado	Domingo	Lunes	Miércoles	Jueves
Clientes	27	27	28	28	30	30	31

2. En la tabla 7 se busca el valor de la variable que deja el mismo número de observaciones (en este caso días) a su derecha y a su izquierda. Por tanto, la mediana es en este caso 28 clientes. Tal y como se puede ver el valor 28 del Domingo deja a su derecha e izquierda 3 observaciones, respectivamente. Dicho de otra manera: La mitad de los días acudieron al buffet 28 o menos clientes y la otra mitad 28 a más clientes.

En el caso de que los datos estuvieran en una tabla agrupados por sus frecuencias absolutas, el procedimiento sería el siguiente:

1. Se ordenan los valores de la variable de menor a mayor, pero, en esta ocasión, en una tabla agrupada por frecuencias. Para el ejemplo del buffet de desayuno esta ordenación se encuentra en la tabla 8.

Tabla 8. Número de clientes que han entrado al buffet de desayuno

Clientes	n_i	N_i	F_i
27	2	2	0.286
28	2	4	0.571
30	2	6	0.857
31	1	7	1

2. Se obtiene la frecuencia absoluta acumulada y se busca dentro de sus valores el primero que sea mayor a $N/2$. Este valor nos indicará el valor mediano en la columna de los valores de la variable.

Concretamente, en nuestro ejemplo, $N/2 = 7/2 = 3.5$, siendo el primer valor mayor a 3.5 en la columna N_i el 4, con lo que la mediana es 28 clientes.

Alternativamente a la utilización de la frecuencia absoluta acumulada, se puede utilizar la frecuencia relativa acumulada (F_i) para obtener la mediana. En este caso, habrá que buscar el primer valor superior a 0.5 en la columna F_i, éste nos indicará el valor de la mediana en la columna de los valores de la variable. Utilizando la tabla 8, se puede obtener la mediana siguiendo este procedimiento cuyo valor (28 clientes) coincide, como no podía ser de otra manera, con el procedimiento anteriormente expuesto.

Si existe un valor de la frecuencia absoluta acumulada o, en su caso de la frecuencia relativa acumulada, que coincida con $N/2$ ó 0.5, respectivamente, la mediana se obtendrá sumando el valor correspondiente de la variable más el siguiente y dividiendo la suma por dos. Supongamos que, para nuestro ejemplo, los datos recogidos sólo han sido los que van desde el lunes hasta el sábado, tal y como se recogen en la tabla 9 o para datos agrupados en la tabla 10. En este caso, si nos fijamos en la tabla 9, no existe ninguna observación (día) que divida en dos partes iguales la muestra, al tratarse de un número par de observaciones, con lo que la mediana se obtendrá como media de los dos valores más próximos a la posición central, en este caso los valores 28 y 30. Si, por el contrario, se desea calcular la mediana a partir de una tabla agrupada por frecuencias como la tabla 10, se procede de la siguiente manera: En este caso $N/2 = 6/2 = 3$, que coincide con un determinado valor en la columna de las frecuencias absolutas acumuladas de la tabla 10. Igualmente, el valor 0.5 aparece en la columna de las frecuencias relativas acumuladas de esta misma tabla. En esta situación la mediana se obtendría igualmente como:

Módulo 2

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

Tabla 9. Número de clientes que han entrado al buffet de desayuno ordenados de menor a mayor

Día	Martes	Viernes	Sábado	Lunes	Miércoles	Jueves
Clientes	27	27	28	30	30	31

Tabla 10. Número de clientes que han entrado al buffet de desayuno

Clientes	n_i	N_i	F_i
27	2	2	0.286
28	1	3	0.50
30	2	5	0.833
31	1	6	1

En el caso de tener los datos en una tabla agrupados en intervalos, la mediana se encuentra en el intervalo mediano, que se obtiene de la misma manera que se obtenía la mediana en el caso de los datos no agrupados por intervalos.[4]

Al igual que la media aritmética, la mediana es una medida que intenta determinar la posición central de la distribución de la variable. La mediana tiene la ventaja, frente a la media aritmética, que no se ve influenciada por valores extremos.

3.2. Otras medidas de posición: la moda y los cuantiles

La moda

Es el valor de la variable que más veces se repite. La representaremos por M_0. La distribución de una variable puede tener más de una moda. Así, por ejemplo, la variable *edad de los clientes*, recogida en la tabla 2, cuenta con dos modas, las edades de 25 y 45 años. En el caso de trabajar con una variable agrupada en intervalos la moda se encuentra en el intervalo modal, que corresponderá con aquel que tiene una mayor frecuencia absoluta. Igualmente, una variable puede tener más de un intervalo modal[5].

Cuantiles

Módulo 2

Son valores de la variable que dividen las observaciones en partes iguales. Como ejemplo de cuantiles tenemos los siguientes: cuartiles, deciles y percentiles.

- Cuartiles: son los tres valores de la variable que dividen en cuatro partes iguales el número de observaciones.
- Deciles: son los nueve valores de la variable que dividen las observaciones en 10 partes iguales.

4 Una vez determinado el intervalo mediano, la mediana se obtiene mediante la siguiente expresión:

$$Me = E_{i-1} + \frac{N/2 - n_{i-1}}{n_i} a_i$$

Sin embargo, para no complicar sobre manera los cálculos para la obtención de esta medida, teniendo en cuenta que la utilización de la agrupación por intervalos es cada vez menos utilizada, debido a la facilidad con que los ordenadores manejan los datos sin necesidad de realizar este tipo de agrupaciones, y considerando, además, que la agrupación por intervalos supone pérdida de información; a lo largo de la asignatura sólo será necesario identificar el intervalo mediano cuando lo datos estén agrupados por intervalos.

5 Una vez determinado cuál es el intervalo modal, la moda se obtiene a partir de la siguiente expresión:

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

Por la misma razón que se exponía para el caso de la mediana (ver nota a pie de página anterior), a lo largo de la asignatura, en el caso de que los datos estén agrupados por intervalos, sólo será necesario identificar el intervalo modal.

- Percentiles: son los noventa y nueve valores de la variable que dividen las observaciones en 100 partes iguales.

El procedimiento para el cálculo de los diferentes cuantiles es muy parecido al de la mediana. En el caso de la mediana se busca el valor de la variable que divide el número de observaciones en dos partes iguales. En el caso de los cuartiles se buscan los valores de la variable que dividen las observaciones en cuatro partes iguales, en diez partes iguales para los deciles y en cien partes iguales para los percentiles.

Como ejemplo del cálculo de los cuartiles se utilizarán los datos de la tabla 2 de las edades de los clientes de un determinado complejo turístico. Para este propósito se pueden utilizar las frecuencias absolutas acumuladas o las frecuencias relativas acumuladas. En el caso de las frecuencias absolutas acumuladas hay que identificar en esta columna los primeros valores inmediatamente superiores a:

$$N/4, \quad N/4 \times 2 \quad y \quad N/4 \times 3 \tag{2.9}$$

que denominaremos valores de referencia o indicadores. En el caso concreto que nos ocupa serán:

$$N/4 = 50/4 = 12.5\,,\ N/4 \times 2 = 50/4 \times 2 = 25 \text{ y } N/4 \times 3 = 50/4 \times 3 = 37.5\,.$$

Módulo 2

Observando la tabla 2, y más concretamente en la columna N_i, los primeros valores inmediatamente superiores a los de referencia son el 16, el 27 y el 42, respectivamente, a partir de los cuales se obtienen en la columna de los valores de la variable (X_i) el primer, el segundo y el tercer cuartil, que son las edades de 25, 33 y 45 años, respectivamente.

En el caso de que los valores indicadores o de referencia señalados en (2.9) coincidan con algún valor concreto de los que se encuentra en la columna de la frecuencia absoluta acumulada (N_i), la obtención del cuartil correspondiente se realizará, como en el caso de la mediana, tomando el valor de la variable que corresponde a dicha frecuencia absoluta acumulada más el siguiente y dividiendo esta suma entre dos.

Si en vez de utilizar la frecuencia absoluta acumulada para el cálculo de los cuartiles, se utiliza la frecuencia relativa acumulada, habrá que actuar de la siguiente manera: Los valores de referencia son:

$$1/4 = 0.25,\ 1/4 \times 2 = 0.50 \ y\ 1/4 \times 3 = 0.75$$

A partir de estos valores se localizan en la columna de la frecuencia relativa acumulada (F_i) los primeros valores inmediatamente superiores a los de referencia, que en este caso son: 0.32, 0.54 y 0.84, los cuales indicarán el valor del primer, del segundo y del tercer cuartil en la columna de los valores que toma la variable (X_i), que en este caso son las edades de 25, de 33 y de 45 años, respectivamente.

Al igual que para el caso de la frecuencia absoluta acumulada, si uno de los valores de referencia coincide exactamente con alguno de los valores de la frecuencia relativa acumulada (F_i), la obtención del cuartil correspondiente se realizará mediante la suma del valor de la variable que señala la frecuencia relativa acumulada más el siguiente y dividiendo esta suma entre dos.

El cálculo de los deciles y los percentiles se realiza de la misma manera que para los cuartiles, la única diferencia será la expresión que se utiliza para obtener los valores de referencia. En el caso de los deciles y los percentiles, y utilizando las frecuencia absoluta acumulada, los valores de referencia son:

$$N/10,\ N/10\times 2\ ,\ N/10\times 3,\cdots,N/10\times 9 \text{ para los deciles y}$$

$$N/100,\ N/100\times 2\ ,\ N/100\times 3,\cdots,N/100\times 99 \text{ para los percentiles}$$

Si se utiliza la frecuencia relativa acumulada los valores de referencia son:

$$1/10, 1/10\times 2, 1/10\times 3,\cdots,1/10\times 9 = 0.10, 0.20, 0.30,\cdots,0.90 \text{ para los deciles y}$$

$$1/100, 1/100\times 2, 1/100\times 3,\cdots,1/100\times 99 = 0.01, 0.02, 0.03,\cdots, 0.99 \text{ para los percentiles}$$

ACTIVIDADES

ACTIVIDAD 1

En el fichero *actividad_2_1.xls* se encuentran los datos del porcentaje de ocupación de un determinado hotel durante 50 días. Con esta información realiza las siguientes cuestiones[6]:

a) Tabular los datos mediante una agrupación sin intervalos.

b) En la misma tabla incorporar las siguientes columnas: (1) La frecuencia absoluta, (2) la frecuencia absoluta acumulada, (3) la frecuencia relativa, y (4) la frecuencia relativa acumulada. Explica qué significado tiene cada una de estas columnas, utilizando para ello ejemplos de los resultados concretos que has obtenido para el caso del porcentaje de ocupación.

c) Utilizando la tabulación realizada en el punto a), obtener los siguientes estadísticos descriptivos del porcentaje de ocupación: (1) La media aritmética, (2) la mediana, (3) la moda y (4) los cuartiles. Interpreta el resultado de cada uno de los estadísticos que has obtenido.

6 Al tratarse de la primera actividad que se realiza, y teniendo en cuenta, igualmente, que por primera vez se utiliza el programa *Excel* para su elaboración, antes de llevarla a cabo sería conveniente que se realizara el ejercicio 2 de autocontrol, que con otros datos, reproduce en bastantes puntos la actividad 1.

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: Teoría y ejercicios resueltos*. Málaga: Agora.

EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

Determina cuáles de las siguientes afirmaciones son verdaderas. Ten en cuenta que en cada una de las preguntas pueden haber más de una afirmación verdadera.

1. La frecuencia absoluta:
   a. Es el número de individuos que presentan un mismo valor de la variable.
   b. Es el número de veces que se repite un determinado valor de la variable.
   c. El número total de observaciones debe ser igual a la suma de las frecuencias absolutas.
   d. En ocasiones la frecuencia absoluta puede ser negativa.

2. Frecuencia relativa y frecuencia relativa acumulada:
   a. La frecuencia relativa es la proporción de individuos que presentan un mismo valor de la variable.
   b. La suma de las frecuencias relativas es igual al número total de observaciones.
   c. La frecuencia relativa acumulada difiere de la frecuencia absoluta en que ésta última siempre es superior.
   d. La suma de las frecuencias relativas acumuladas debe ser igual a uno.

3. Agrupaciones por intervalos:
   a. La marca de clase se define como el punto medio del intervalo.
   b. La marca de clase se define como el producto de $1/2$ por la amplitud del intervalo.
   c. La marca de clase se define como el producto de $1/2$ por la suma de los extremos del intervalo.
   d. La amplitud del intervalo puede ser distinto para cada uno de los intervalos.

4. Descripción gráfica:
   a. Si queremos obtener información sobre el comportamiento descriptivo de una determinada variable, las diferentes representaciones gráficas de la misma sustituye la necesidad de calcular sus diferentes estadísticos descriptivos.
   b. En el diagrama de barras las variables no se agrupan por intervalos mientras que en el histograma sí.
   c. El histograma de frecuencias nos proporciona la historia gráfica de una determinada variable.
   d. La altura de las diferentes barras del histograma siempre coincide con el valor de la frecuencia absoluta.

5. La media aritmética:
   a. Nunca puede tomar valores negativos.
   b. Sirve para determinar la dispersión de una variable.
   c. Se ve afectada por la existencia de valores extremos.
   d. Si a cada uno de los valores de una variable se la suma una misma constante, la media aritmética de la variable no varía.

6. La mediana:
   a. La mediana es aquel individuo que deja por encima y por debajo el mismo número de estos.
   b. La mediana es aquel valor de la variable, que una vez ordenada de menor a mayor, deja por encima y por debajo el mismo número de individuos.
   c. La mediana es una medida de posición central.
   d. Una distribución puede tener más de una mediana.

7. La moda:
   a. La distribución de una variable sólo puede tener una moda.
   b. La distribución de una variable puede tener más de una moda.
   c. La moda es el valor de la variable que más veces se repite.
   d. La moda es la frecuencia absoluta que más veces se repite.

8. Los cuantiles:
   a. Los cuantiles son valores de la variable que recogen la cuenta de la suma de las frecuencias absolutas.
   b. Para conocer con más detalle la distribución de una variable, es preferible calcular los cuartiles que los deciles.
   c. Los percentiles son los 100 valores de una variable que dividen las observaciones en 100 partes iguales.
   d. El segundo cuartil coincide con el valor de la mediana.

EJERCICIO 2

En el fichero *ejercicio_2_2.xls* se encuentran los datos del porcentaje de ocupación de un determinado hotel durante 50 días. Con esta información realiza las siguientes cuestiones:

a) Tabular los datos mediante una agrupación sin intervalos (asegúrate que los valores de la variable estén ordenados de menor a mayor).

b) En la misma tabla incorporar las siguientes columnas: (1) La frecuencia absoluta, (2) la frecuencia absoluta acumulada, (3) la frecuencia relativa, y (4) la frecuencia relativa acumulada.

c) Utilizando la tabulación realizada en el punto a), obtener los siguientes estadísticos descriptivos del porcentaje de ocupación: (1) La media aritmética, (2) la mediana, (3) la moda y (4) los cuartiles.

EJERCICIO 3

En la siguiente tabla se presenta el precio de diferentes menús (en euros) que se sirven en un determinado restaurante. A partir de esta información obtén:

a) La frecuencia relativa, la frecuencia absoluta acumulada y la frecuencia relativa acumulada.

b) La media aritmética, la mediana y la moda. Interpreta los resultados.

c) Interpreta el valor f_2, N_2 y F_2 de la frecuencia relativa, la frecuencia absoluta acumulada y la frecuencia relativa acumulada, respectivamente.

X_i	n_i
10.0	3
12.5	20
15.0	10
25.0	7

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

1. a, b, c
2. a
3. a, c, d
4. b
5. c
6. b, c
7. b, c
8. d

EJERCICIO 2

La solución del ejercicio se encentra en el fichero *sol_ejercicio_2_2.xls*.

EJERCICIO 3

Módulo 2

a)

	X_i	n_i	N_i	f_i	F_i	$X_i\, f_i$
	10	3	3	0.08	0.08	0.75
	12.5	20	23	0.50	0.58	6.25
	15	10	33	0.25	0.83	3.75
	25	7	40	0.18	1	4.38
Total		40		1		15.13

b)

Media = 15.13
Mediana = 12.50
Moda = 12.50

Interpretación: el precio medio de los menús es igual a 15.13 euros. El 50% de los menús que se han servido tienen un precio inferior o igual a 12.50 euros mientras que el otro 50% tiene un precio superior o igual a 12.50 euros. El menú que más se sirve es el de 12.50 euros.

c) $f_2 = 0.50$ → El 50 % de los menús que se han servido han tenido un precio de 12.50 euros.

$N_2 = 23$ → De los 40 menús que se han servido, 23 han tenido un precio de 12.50 euros o menos, mientras que los restantes han tenido un precio superior.

$F_2 = 0.58$ → El 58% de los menús que se han servido han tenido un precio de 12.50 euros o menos y el resto (42%) han tenido un precio superior.

GLOSARIO DE TÉRMINOS

Amplitud del intervalo: diferencia entre el extremo superior e inferior del intervalo.

Cuantiles: valores de una variable que dividen en partes iguales a las observaciones (individuos) de la muestra.

Cuartiles: tres valores de la variable que divide en cuatro partes iguales el número de observaciones de la muestra.

Deciles: nueve valores de la variable que divide en diez partes iguales el número de observaciones de la muestra.

Diagrama de sectores: representación gráfica de las frecuencias de un variable.

Estadístico: medida que se obtiene como función de datos muestrales. Ejemplo de estadísticos son la media, la mediana, etc.

Frecuencia absoluta: número de veces que se repite un determinado valor de una variable.

Frecuencia relativa: proporción en que las observaciones toman un mismo valor de una variable.

Histograma: representación gráfica de las frecuencias para datos agrupados por intervalos.

Marca de clase: valor que divide en dos partes iguales un intervalo.

Media aritmética: promedio de los valores de una variable.

Mediana: una vez ordenados los valores de una variable de menor a mayor, la mediana es el valor de la variable que divide en dos partes iguales el número de observaciones.

Medidas de posición: indican valores de las variables alrededor de las cuales se sitúa un grupo de observaciones.

Moda: es el valor de una variable que más veces se repite.

Percentiles: noventa y nueve valores de la variable que divide en cien partes iguales el número de observaciones de la muestra.

Módulo 3

Descripción univariante: Medidas de dispersión,
de forma y de concentración.
Tipificación de una variable

PRESENTACIÓN DEL MÓDULO

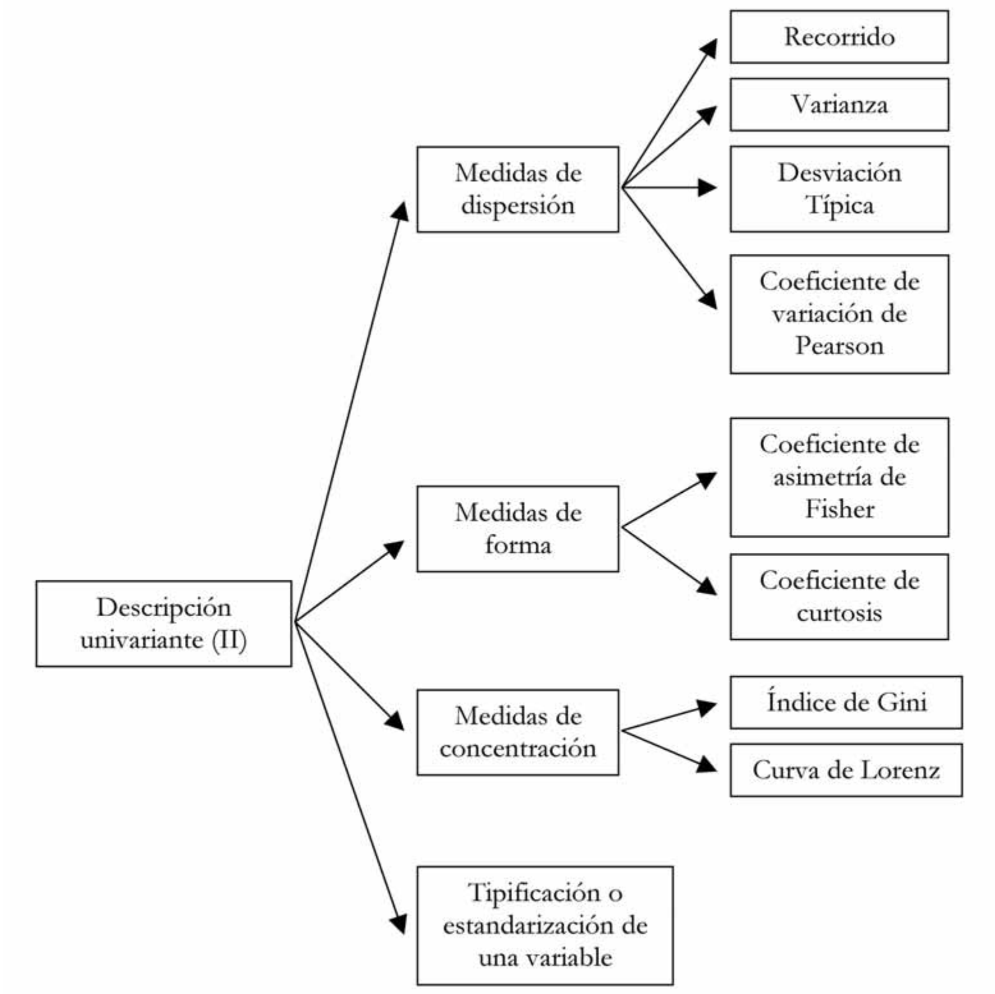
Al igual que en el módulo 2, en este módulo se seguirán estudiando diferentes estadísticos o medidas que nos servirán para describir la distribución de una variable. En concreto se verán medidas de dispersión, de forma y de concentración. Las medidas de dispersión nos servirán para determinar el grado de variabilidad con que cuenta la variable respecto a un valor de referencia, normalmente la media aritmética, lo que nos dará una idea de la representatividad de este estadístico. Las medidas de forma nos ayudarán a determinar la forma de la distribución de la variable en relación a su asimetría y su apuntamiento. Por último, las medidas de concentración nos indicarán el grado de reparto o equidistribución de la variable, característica especialmente relevante cuando se trabaja con variables de riqueza o renta. El módulo finaliza con el concepto de tipificación o estandarización de una variable. Con este procedimiento seremos capaces de detectar observaciones atípicas ("raras") en la distribución de una variable.

OBJETIVOS DEL MÓDULO

La superación de este módulo supone alcanzar los siguientes objetivos:

1. Saber interpretar y describir la dispersión, forma y concetración de la distribución de una variable mediante la obtención de los estadísticos descriptivos correspondientes.
2. Saber tipificar y detectar la existencia de observaciones atípicas en la distribución de una variable.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. MEDIDAS DE DISPERSIÓN: RECORRIDO, VARIANZA, DESVIACIÓN TÍPICA Y COEFICIENTE DE VARIACIÓN

Las medidas de dispersión intentan cuantificar la variabilidad o dispersión que presenta la variable. Una medida inmediata que cumple esta función es la diferencia entre el valor máximo y mínimo que toma la variable. Sin embargo, en este caso sólo se utilizan dos valores de la variable, con lo que se pierde la información que proporciona el resto de las observaciones. Otras medidas de dispersión cuantifican la diferencia entre todas las observaciones y un valor de referencia, que suele ser la media aritmética. Un problema que surge a la hora de proceder de esta manera es que las diferencias pueden tener diferentes signos, y por tanto, compensarse las desviaciones positivas con las negativas y dar una impresión falsa sobre la variabilidad. Para evitar este efecto se puede utilizar las diferencias en valor absoluto, sin embargo, lo más habitual es elevar las diferencias al cuadrado.

El recorrido

El recorrido se obtiene de la resta entre el valor máximo y mínimo de la variable tal que:

$$R = X_N - X_1 \tag{3.1}$$

donde previamente se han ordenado los valores de la variable de menor a mayor. Esta medida tiene el inconveniente de que sólo utiliza dos de los distintos valores de la variable, con lo que una parte importante de la información que se posee no se utiliza. Cuando en la distribución se tienen valores excesivamente extremos, aunque sean muy pocos, afectan a la medida enormemente.

La varianza y desviación típica

La varianza de una variable es la media del cuadrado de las desviaciones entre cada una de sus observaciones y su media aritmética. Esta medida da una idea de la representatividad de la media aritmética. Entre mayor es el valor de la varianza de una variable mayor es su dispersión alrededor de su media. La expresión de la varianza para datos agrupados es la siguiente:

$$\mathrm{Var}(X) = \frac{\sum_{i=1}^{J} \left(X_i - \overline{X}\right)^2 n_i}{N} \tag{3.2}$$

Otras expresiones equivalentes son:

$$Var(X) = \sum_{i=1}^{J} \left(X_i - \overline{X}\right)^2 f_i; \quad Var(X) = \frac{\sum_{i=1}^{J} X_i^2 n_i}{N} - \overline{X}^2; \quad Var(X) = \sum_{i=1}^{J} X_i^2 f_i - \overline{X}^2 \tag{3.3}$$

Un inconveniente que presenta la varianza es su unidad de medida. Ésta no coincide con la de la variable, al tener que elevar al cuadrado las diferencias entre los valores observados y su correspondiente media aritmética. Para corregir este hecho se suele utilizar como medida de dispersión la desviación típica, que se obtiene como la raíz cuadrada de la varianza, tal que:

$$Dt(X) = \sqrt{Var(X)} \tag{3.4}$$

Supongamos que se tienen las estadísticas del envío diario de ropa (en decenas de kilogramos) que se recibe en la lavandería A de un determinado establecimiento turístico, tal y como aparece en la primera columna de la tabla 1.

Tabla 1. Ropa recibida por la lavandería A (unidad: decenas de kilogramos)

X_i	n_i	$X_i n_i$	$(X_i - \overline{X})^2 n_i$
10	3	30	75
12	5	60	45
15	12	180	0
17	10	170	40
20	2	40	50
	32	480	210

En este caso, $\overline{X} = 15$. Además, teniendo en cuenta las expresiones (3.2) y (3.4), el cálculo de la varianza y la desviación típica será igual a

$$Var(X) = \frac{210}{32} = 6.56 \quad \Rightarrow \quad Dt(X) = \sqrt{6.56} = 2.56$$, respectivamente.

De esta manera, podemos decir que la dispersión de la variable alrededor de su media es de 25.6 kilogramos aproximadamente. Supongamos que para una segunda lavandería (lavandería B), y para la misma variable, se tienen los datos reflejados en la tabla 2.

Tabla 2. Ropa recibida por la lavandería B (unidad: decenas de kilogramos)

X_i	n_i	$X_i n_i$	$(X_i - \overline{X})^2 n_i$
10	8	80	200
12	6	72	54
15	2	30	0
17	7	119	28
20	9	180	225
	32	481	507

En esta ocasión $\overline{X} = 15.03$, mientras que la varianza y la desviación típica son:

$$Var(X) = \frac{507}{32} = 15.84 \quad \Rightarrow \quad Dt(X) = \sqrt{15.84} = 3.98$$, respectivamente.

Tal y como se puede comprobar, aún cuando ambas lavanderías reciben en media aproximadamente los mismos kilogramos de ropa (150 kg.), sin embargo, las desviaciones típicas de ambos establecimientos son muy distintas. Mientras que la desviación típica de la primera lavandería es igual a 25.6 kilogramos, la de la segunda es de 39.8 kilogramos, un 55% superior. En el caso de que se pretenda destinar recursos de personal y material a las lavanderías, en función de la media de kilogramos de ropa que diariamente se reciben, la menor dispersión con que cuenta la lavandería A hace que sea más fácil realizar esta asignación en ella. En la lavandería B se corre un mayor riesgo de que en determinados días los recursos asignados no puedan hacer frente a las necesidades del servicio, mientras que en otros, sin embargo, los recursos se quedan ociosos.

PROPIEDADES DE LA VARIANZA

1. La varianza siempre es positiva.
2. Si se crea una nueva variable a partir de la original, sumando una constante a todos sus valores, la varianza de la nueva variable es igual a la varianza de la variable original.

$$Si \quad Z_i = C + X_i \;\Rightarrow\; Var\,(Z) = Var(X)$$

3. Si se crea una nueva variable a partir de la original, multiplicando una constante a todos sus valores, la varianza de la nueva variable es igual al cuadrado de la constante multiplicada por la varianza de la variable original.

$$Si \quad Z_i = C \times X_i \;\Rightarrow\; Var\,(Z) = C^2 \times Var(X)$$

El coeficiente de variación de Pearson

Un inconveniente de la desviación típica es su dimensionalidad. Una vez se conoce su valor para una determinada distribución, no es inmediato determinar si es grande o pequeño, salvo que se compare con un valor de referencia, por ejemplo, la media aritmética. Además, y por esta misma razón, la desviación típica varía para una misma distribución si se cambia la unidad de medida de la variable. Así, por ejemplo, en el caso del envío de ropa a la lavandería, visto en el ejemplo anterior, si la unidad de medida en vez de decenas de kilogramos fuera de gramos, las desviaciones típicas obtenidas quedarían multiplicadas por 10000. Para evitar estos inconvenientes se puede utilizar el coeficiente de variación de Pearson, que se obtiene como[1]

$$CV = \frac{Dt(X)}{\bar{X}} \tag{3.5}$$

1 No tiene sentido calcular el coeficiente de variación cuando la variable toma valores positivos y negativos al mismo tiempo. Piénsese en que estas circunstancias una media aritmética de cero sería posible, con lo que el coeficiente de variación sería infinito. Sólo cuando todos los valores de la variable son positivos, o cuando todos los valores de la variable son negativos, tendrá sentido el uso de esta medida de dispersión.

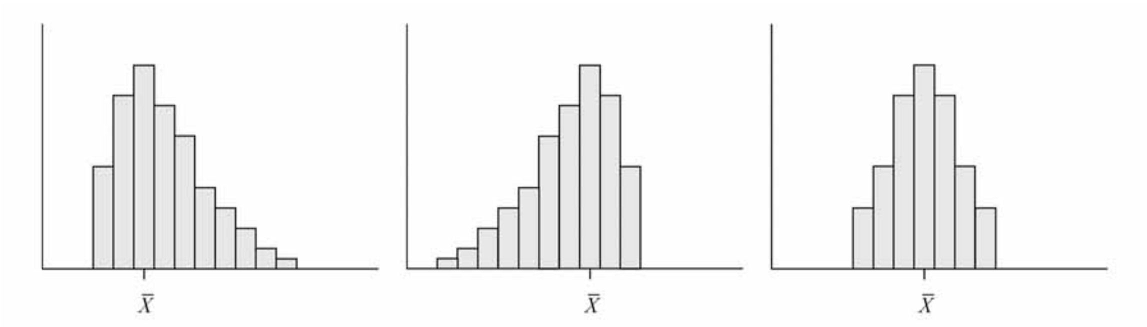
Cuando CV es menor que uno en valor absoluto[2], la desviación típica es inferior al valor de la media, con lo que se considera que la variabilidad no es muy elevada. Por el contrario, cuando CV es superior a la unidad en valor absoluto se empieza a hablar de una desviación grande, ya que en este caso la desviación típica supera el valor de la media. Si se desea el CV puede presentarse en porcentajes, simplemente multiplicándolo por 100.

2. Medidas de forma: el coeficiente de asimetría de Fisher y el coeficiente de curtosis

El coeficiente de asimetría de Fisher

Gráficamente, el histograma es una herramienta que permite visualizar el grado de asimetría de una distribución. En la figura 1 se presentan diferentes tipos de asimetrías. En el caso (a) se dice que hay asimetría a la derecha, en el caso (b), por el contrario, que hay asimetría a la izquierda y, por último, en el caso (c), que hay simetría.

Figura 1. Diferentes tipos de asimetrías



(a) Asimetría a la derecha. **(b)** Asimetría a la izquierda. **(c)** Simetría.

El coeficiente de Fisher permite medir el grado de asimetría que presenta una distribución sin necesidad de utilizar el histograma de frecuencias, sino simplemente obteniendo el valor de dicho coeficiente. Éste se obtiene mediante la expresión:

$$CA = \frac{\frac{1}{N}\sum_{i=1}^{J}\left(X_i - \bar{X}\right)^3 n_i}{Dt(X)^3} \tag{3.6}$$

2 Recordemos que para un determinado valor negativo "*a*", su valor absoluto se obtiene tal que |-*a*|= *a*. Como sabemos, la desviación típica siempre es positiva, sin embargo, la media aritmética de una variable puede ser negativa. Por esta razón el CV puede ser negativo, de ahí la necesidad de interpretar su valor en términos de valor absoluto.

La interpretación de los posibles valores del CA es la siguiente:

Si $CA<0$ hay asimetría a la izquierda (asimetría negativa).

Si $CA=0$ hay simetría.

Si $CA>0$ hay asimetría a la derecha (asimetría positiva).

Es interesante tener en cuenta que cuando todos los valores de una determinada variable se multiplican por una misma constante, el coeficiente de asimetría de Fisher no varía.

Siguiendo con el caso de las lavanderías de la tabla 1 y 2, el desarrollo necesario para el cálculo del coeficiente de asimetría de la lavandería A y B se presentan en la tabla 3, a partir de los cuales obtenemos sus valores tal que:

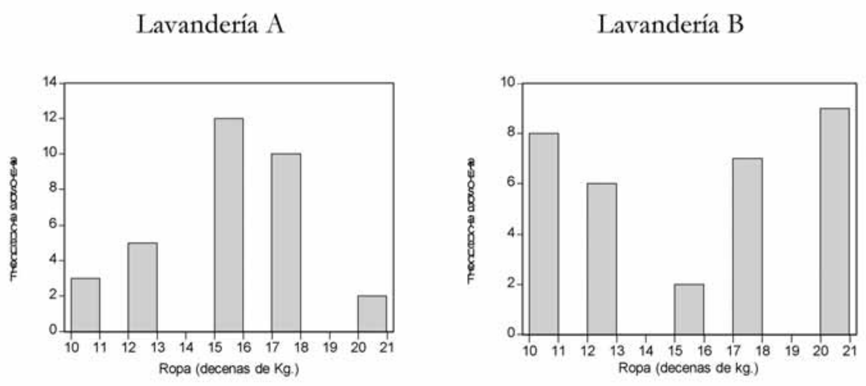
Lavandería A — Lavandería B

$$CA = \frac{\frac{1}{32}(-180)}{2.56^3} = -0.33 \,; \quad CA = \frac{\frac{1}{32}(-28.53)}{3.98^3} = -0.01$$

Tabla 3. Cálculos para la obtención del coeficiente de asimetría y de curtosis

Lavandería A		Lavandería B	
$(X_i - X)^3\, n_i$	$(X_i - X)^4\, n_i$	$(X_i - X)^3\, n_i$	$(X_i - X)^4\, n_i$
-375	1875	-1018.87	5126.18
-135	405	-167.12	506.57
0	0	0.00	0.00
80	160	53.42	105.16
250	1250	1104.04	5485.69
-180	3690	-28.53	11223.60

Figura 2. Histogramas de la ropa recibida en las lavanderías



Tal y como se puede observar, el signo del coeficiente de asimetría para ambas lavanderías es negativo lo que indica una asimetría hacia el lado izquierdo. Sin embargo, estas asimetrías son muy suaves debido a que en ambos casos los coeficientes están muy próximos a cero. En la figura 3.2 se pueden observar distribuciones bastantes simétricas para este ejemplo.

El coeficiente de curtosis

Al igual que para el caso de la asimetría, observando el histograma se puede percibir el grado de apuntamiento que presenta la distribución de una determinada variable. En la figura 3 se presenta la clasificación habitual de los tipos de apuntamientos. La distribución normal[3] sirve de referencia para esta clasificación. Esta distribución está representada en los dibujos de la figura 3 por la línea continua en forma de campana. En esta figura el dibujo (a) representa una distribución con un apuntamiento igual que el de una distribución normal, se le denomina *mesocúritica*. El dibujo (b) presenta un apuntamiento mayor que el correspondiente a una distribución normal, en este caso se dice que la distribución es *leptocúrtica*. Finalmente, en el dibujo (c) se exhibe una distribución con un apuntamiento menor que la correspondiente a la distribución normal, en este caso la distribución recibe el nombre de *planticúrtica*.

El coeficiente de curtosis sirve para determinar el grado de apuntamiento de una distribución, su valor se obtiene mediante la siguiente expresión:

Módulo 3

$$K = \frac{\frac{1}{N}\sum_{i=1}^{J}\left(X_i - \bar{X}\right)^4 n_i}{Dt(X)^4} \qquad (3.7)$$

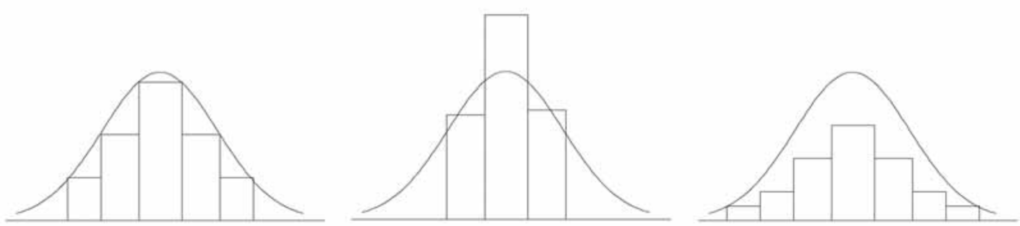
Tal y como se puede observa, el coeficiente de curtosis es siempre positivo. La interpretación de este coeficiente responde al siguiente esquema:

Si K=3 la distribución es mesocúrtica.

Si K > 3 la distribución es leptocúrtica.

Si K < 3 la distribución es planticúrtica.

Figura 3. Tipos de apuntamientos



(a) Mesocúrtica (b) Leptocúrtica (c) Planticúrtica

3 En la disciplina de la Estadística existen muchas distribuciones teóricas que se utilizan con diversos propósitos. La distribución más utilizada es la que se conoce con el nombre de "normal" o "gaussiana".

Es interesante tener en cuenta que cuando todos los valores de una determinada variable se multiplican por una misma constante, el coeficiente de curtosis no varía.

En la tabla 3 se han desplegado las columnas segunda y cuarta, necesarias para calcular el coeficiente de curtosis para ambas lavanderías a partir de la ecuación (3.7). Su resultado es:

Lavandería A Lavandería B

$$K = \frac{\frac{1}{32} \times 3690}{2.56^4} = 2.68\,; \quad K = \frac{\frac{1}{32} \times 11223.60}{3.98^4} = 1.40$$

Tal y como se observa, en ambos casos el coeficiente de curtosis tiene valores inferiores a 3, lo que implica que las distribuciones son más achatadas que la correspondiente a una distribución normal, concretamente se trata de dos distribuciones *planticúrticas*.

3. Medidas de concentración: el índice de Gini y la curva de Lorenz

Las medidas de concentración nos indican el grado de reparto o equidistridribución de la distribución de una variable. Son especialmente útiles cuando estamos interesados en estudiar el reparto que se hace de variables como la renta, la tenencia de propiedades, el reparto de subvenciones, etc.

El índice de Gini

El valor de este índice sirve para determinar el grado de concentración de una variable. Una vez ordenados los valores de la variable de menor a mayor, el índice de Gini se obtiene mediante la siguiente expresión:

$$I_G = \frac{\sum_{i=1}^{J-1}(p_i - q_i)}{\sum_{i=1}^{J-1} p_i} \qquad (3.8)$$

Este índice está acotado en el intervalo [0,1]. Los valores extremos tienen la siguiente interpretación: Un valor igual a uno indica máxima concentración o mínima equidistribución o reparto, mientras que un valor igual a cero indica mínima concentración o máxima equidistribución o reparto. Supongamos que estamos estudiando el reparto de los salarios entre los empleados de un determinado establecimiento turístico. En el caso de $I_G = 1$, un trabajador se queda con toda la masa salarial y, por tanto, el resto no recibe nada. Por otro lado, si $I_G = 0$ todos los trabajadores tienen el mismo salario. Evidentemente, esto dos valores extremos difícilmente se van a dar en la realidad, pero nos sirven de referencia para todos los casos intermedios. De esta manera, cuando el índice está próximo a uno estaremos ante situaciones de alta concentración, mientras que valores próximos a cero indicarán un reparto más equitativo.

En la tabla 4 se encuentran los salarios anuales de los empleados de un determinado establecimiento turístico. En la misma tabla se presentan los cálculos necesarios para calcular el índice de Gini, donde:

$$u_i = X_1 n_1 + X_2 n_2 + \cdots + X_i n_i; \qquad p_i = \frac{N_i}{N} \times 100; \qquad \text{y} \qquad q_i = \frac{u_i}{u_N} \times 100 \tag{3.9}$$

Tal y como se puede comprobar, P_i coincide con la frecuencia absoluta acumulada multiplicada por cien *($F_i \times 100$)*.

A partir de los datos calculados en la tabla 4 se tiene que:

$$I_G = \frac{\sum_{i=1}^{J-1} (p_i - q_i)}{\sum_{i=1}^{J-1} p_i} = \frac{58.34}{548.18} = 0.11$$

con lo que el reparto de la masa salarial entre los empleados de este establecimiento se puede clasificar de bastante equitativa, teniendo en cuenta que el valor del índice de Gini está muy próximo a cero.

Módulo 3

Tabla 4. Cálculo del índice de Gini para el salario anual de los trabajadores de un determinado complejo turístico

X_i	n_i	N_i	$X_i n_i$	u_i	p_i	q_i	$p_i - q_i$
9000	30	30	270000	270000	21.90	13.87	8.03
12000	50	80	600000	870000	58.39	44.68	13.71
15000	35	115	525000	1395000	83.94	71.65	12.29
18000	10	125	180000	1575000	91.24	80.89	10.35
24000	5	130	120000	1695000	94.89	87.06	7.83
30000	5	135	150000	1845000	98.54	94.76	3.78
42000	1	136	42000	1887000	99.27	96.92	2.35
60000	1	137	60000	1947000	100	100	0
Total	137		$u_N = 1947000$		548.18*		58.34

* Suma de todos los valores excepto el último.

En la tabla 5 y 6 se presentan dos ejemplos para el caso de máxima y mínima concentración:

Tabla 5. Cálculo del índice de Gini para el caso de concentración máxima ($I_G = 1$)

X_i	n_i	N_i	$X_i n_i$	u_i	p_i	q_i	$p_i - q_i$
0	136	136	0	0	99.27	0.00	99.27
9000	0	136	0	0	99.27	0.00	99.27
12000	0	136	0	0	99.27	0.00	99.27
15000	0	136	0	0	99.27	0.00	99.27
18000	0	136	0	0	99.27	0.00	99.27
24000	0	136	0	0	99.27	0.00	99.27
30000	0	136	0	0	99.27	0.00	99.27
42000	0	136	0	0	99.27	0.00	99.27
60000	1	137	60000	60000	100.00	100.00	0.00
Total	137		60000		794.16*		794.16

* Suma de todos los valores excepto el último

$$I_G = \frac{\sum_{i=1}^{J-1}(p_i - q_i)}{\sum_{i=1}^{J-1} p_i} = \frac{794.16}{794.16} = 1$$

Módulo 3

Tabla 6. Cálculo del índice de Gini para el caso de concentración mínima ($I_G = 0$)

X_i	n_i	N_i	$X_i n_i$	u_i	p_i	q_i	$p_i - q_i$
0	0	0	0	0	0.00	0.00	0.00
9000	137	137	1233000	1233000	100.00	100.00	0.00
12000	0	137	0	1233000	100.00	100.00	0.00
15000	0	137	0	1233000	100.00	100.00	0.00
18000	0	137	0	1233000	100.00	100.00	0.00
24000	0	137	0	1233000	100.00	100.00	0.00
30000	0	137	0	1233000	100.00	100.00	0.00
42000	0	137	0	1233000	100.00	100.00	0.00
60000	0	137	0	1233000	100.00	100.00	0.00
Total	137		1233000		700.00*		0.00

* Suma de todos los valores excepto el último

$$I_G = \frac{\sum_{i=1}^{J-1}(p_i - q_i)}{\sum_{i=1}^{J-1} p_i} = \frac{0.00}{700.00} = 0$$

Curva de Lorenz

La curva de Lorenz tiene el mismo propósito que el índice de Gini, determinar la mayor o menor concentración de una variable. Esta curva se construye a partir de los valores de p_i y q_i que se calculan para el índice de Gini en la ecuación (3.9). En la figura 4 se encuentra la curva de Lorenz para el caso de la distribución de los salarios de los trabajadores de un determinado establecimiento turístico, presentados en la tabla 4. Tal y como se puede observar, en esta figura aparece una bisectriz en trazado de puntos, que nos facilitará la interpretación del gráfico. La curva de Lorenz está situada siempre en el lado inferior de la bisectriz, y se observa en la figura (9) en trazado continuo. A medida que la curva de Lorenz se acerca a la bisectriz, el índice de Gini tiende a cero. De esta manera, si la curva de Lorenz se confunde totalmente con la bisectriz, el índice de Gini es cero (figura 5). Por el contrario, a medida que la curva de Lorenz se aleja de la bisectriz, el índice de Gini tiende a uno. En el extremo, cuando la curva de Lorenz se confunde con el eje p_i el índice de Gini será igual a 1 (figura 6). Por tanto, lo lejos o cerca que se encuentre la curva de Lorenz de la bisectriz o del eje p_i indicará la mayor o menor equidistribución o concentración de la variable, respectivamente. Por ejemplo, para el caso que nos ocupa del reparto salarial de los empleados, recogido en la tabla 4, la curva de Lorenz (figura 4) se encuentra bastante cerca de la bisectriz, con lo que el reparto de los salarios se puede clasificar de bastante equitativa, tal y como ya se había confirmado con el cálculo del índice de Gini.

Módulo 3

Figura 4. Curva de Lorenz (I_G = 0.11)

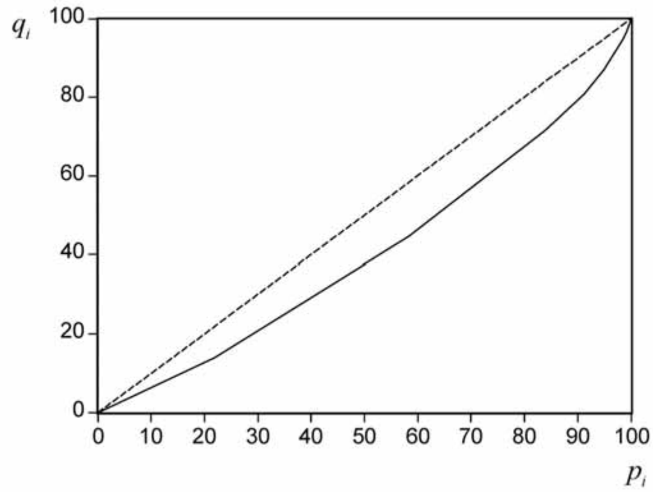


Figura 5. Curva de Lorenz ($I_G = 0$)

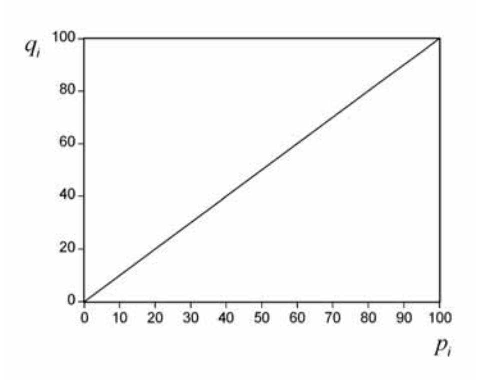
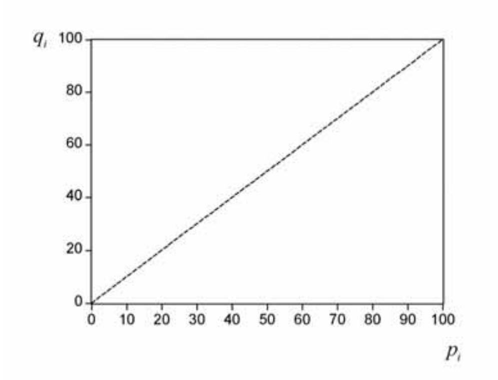


Figura 6. Curva de Lorenz ($I_G = 1$)



4. Tipificación o estandarización de una variable

La tipificación o estandarización de una variable se obtiene restando a ésta su media y dividiendo el resultado por su desviación típica. Si denominamos Z_i a la variable tipificada de X_i, tendremos que:

$$Z_i = \frac{X_i - \bar{X}}{Dt(X)} \qquad (3.10)$$

La media y la desviación típica de una variable tipificada es siempre cero y uno respectivamente. La tipificación de una variable permite de forma sencilla determinar que observaciones de la distribución de una variable se pueden considerar *atípicas*[4].

Se considera que una observación es atípica cuando su valor difiere del resto de las observaciones de forma significativa. Por esta razón también se podrían considerar observaciones *raras*, *extrañas* o *extremas*. El valor de la media y la desviación típica son un referente que sirven para determinar si una observación se puede considerar atípica. Cuanto más se aleja una observación de la correspondiente media aritmética de la variable, mayor es la probabilidad de considerarla atípica. Sin embargo, la desviación típica también juega un papel importante en esta consideración. Dos observaciones de dos distribuciones distintas pueden estar alejadas de sus respectivas medias aritméticas la misma distancia y, solamente una de ellas considerarse atípica. Si este fuera el caso, la observación atípica sería la correspondiente a la variable con menor desviación típica. Por esta razón, para una variable tipificada, la distancia de cada una de las observaciones a la media aritmética se mide en términos de sus respectivas desviaciones típicas.

Veamos un ejemplo, supongamos que el Hotel A y el Hotel B han tenido un 65% de ocupación en el último mes. Además, se tiene la siguiente información sobre el porcentaje de ocupación mensual, en el pasado reciente, para ambos hoteles:

4 La utilidad de la tipificación de una variable en Estadística va más allá de los que se verá en este manual. Por ejemplo, esta operación es muy utilizada a la hora de obtener la distribución de probabilidades de variables aleatorias con distribuciones normales o gaussianas.

	Hotel A	Hotel B
Media	60	50
Desviación típica	10	5

En este caso, el valor estandarizado del 65% de ocupación será igual a:

$$Z_{Hotel\ A} = \frac{65-60}{10} = 0.5 \qquad Z_{Hotel\ B} = \frac{65-50}{5} = 3$$

De esta manera, el porcentaje de ocupación del último mes para del Hotel A y el Hotel B se alejan de su media 0.5 y 3 veces sus respectivas desviaciones típicas, respectivamente. Esto implica que el porcentaje de ocupación del 65% es menos habitual en el Hotel B que en el Hotel A. La cuestión a plantearse a continuación sería la siguiente: ¿Podemos considerar el porcentaje de ocupación del Hotel B atípico? ¿El hecho de que tenga un valor tipificado de 3 significa que lo debemos considerar como atípico?, o ¿este valor se puede encontrar con cierta asiduidad en un distribución estandarizada y, por tanto, considerarlo normal? La respuesta a esta cuestión depende de cómo se distribuyen las observaciones de la variable, sin embargo, existen determinados límites que nos permiten responder a esta pregunta, aunque sea de forma aproximada[5]. De esta manera, usaremos los siguientes valores de referencia para calificar una determinada observación de la variable en relación a su atipicidad.

Si $2 \leq |Z_i| < 2.5 \rightarrow$ Observación algo atípica.

Si $2.5 \leq |Z_i| < 3 \rightarrow$ Observación atípica.

Si $|Z_i| \geq 3 \rightarrow$ Observación extremadamente atípica.

En definitiva, atendiendo a lo anterior se tendrá que cuando los valores absolutos estandarizados se encuentran por encima de 2 y por debajo de 2.5 las observaciones pueden empezar a considerarse atípicas, para valores absolutos por encima a 2.5 e inferiores a 3 serían atípicas y para valores absolutos por encima de 3 serían extremadamente atípicas. Además, el signo del valor estandarizado nos indica el tipo de desviación, positiva o negativa. Supongamos que nos encontramos con una observación atípica. Si el signo del valor estandarizado es positivo, se trata de una observación anormalmente superior a la gran mayoría del resto de las observaciones, por el contrario, en el caso de que su valor estandarizado fuera negativo, se trata de una observación anormalmente inferior a la gran mayoría del resto de las observaciones.

De esta manera, y volviendo al ejemplo del porcentaje de ocupación de los hoteles A y B, un porcentaje de ocupación del 65% en el hotel A sería algo bastante normal, tal y como indica su

5 En las distribuciones teóricas se pueden determinar intervalos con probabilidades de que una observación caiga dentro de los mismos, en el caso de distribuciones desconocidas, la desigualdad de Tchebyshec puede ser de utilidad para este fin. Un estudio pormenorizado de estos aspectos queda fuera de los objetivos de esta asignatura, por lo que solamente nos centraremos en indicar aquellos valores que nos pueden servir de referencia para determinar si una observación se puede considerar atípica o no.

valor estandarizado (0.5), sin embargo, este mismo porcentaje de ocupación (65%) en el hotel B sería algo muy extraño (extremadamente atípico) ya que su valor estandarizado es de 3. Dicho de otra manera, un 65% de ocupación es algo normal, habitualmente hotel en el hotel A. Normalmente este hotel debe de tener porcentajes de ocupación muy cercanos al 65%. Sin embargo, para el hotel B, este mismo porcentaje de ocupación es extraño, es muy raro que alcance el 65%. Además, teniendo en cuenta que el signo de su valor estandarizado es positivo, este nivel de ocupación es extremadamente alto para el hotel B.

ACTIVIDADES

ACTIVIDAD 1

Utilizando el mismo fichero de la actividad 1 del módulo 2 (*actividad_2_1.xls*), y a partir de los resultados de la misma, realiza las siguientes actividades:

a) Obtén los siguientes estadísticos descriptivos: (1) El recorrido, (2) la varianza y la desviación típica, (3) el coeficiente de variación de Pearson, (3) el coeficiente de asimetría de Fisher, y (4) el coeficiente de curtosis.

b) Tipifica cada uno de los valores de la variable y determina si existe alguna observación que podamos considerar atípica.

ACTIVIDAD 2

Módulo 3

En el fichero *actividad_3_2.xls* se encuentran los datos del salario anual (en euros) que perciben los 137 empleados en un determinado establecimiento turístico. A partir de esta información calcula el índice de Gini y dibuja la curva de Lorenz.

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: teoría y ejercicios resueltos*. Málaga: Agora.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

Determina cuáles de las siguientes afirmaciones son verdaderas. Ten en cuenta que en cada una de las preguntas pueden haber más de una afirmación verdadera.

1. La varianza y la desviación típica:
   a. La varianza y la desviación típica son las medidas de dispersión más utilizadas.
   b. Dado que la desviación típica es la raíz cuadrada de la varianza, no podremos hallar la desviación típica de aquellas variables cuya varianza es negativa.
   c. Cuando a una variable le sumamos una constante, la varianza de la variable resultante no varía.
   d. La varianza siempre es superior a la desviación típica.

2. El coeficiente de variación de Pearson:
   a. El coeficiente de variación queda definido por el cociente de la desviación típica y la mediana.
   b. Si a una variable le sumamos una constante, el coeficiente de variación no varía.
   c. Si a una variable la multiplicamos por una constante, el coeficiente de variación no varía.
   d. El coeficiente de variación queda definido por el cociente de la desviación típica y la media.

3. De la distribución de los precios de los menús de diferentes restaurantes, sabemos que la media es igual a 8.40 euros y el coeficiente de variación de Pearson es igual al 0.12 esto significa que:
   a. Hay poca dispersión alrededor de la media.
   b. Hay mucha dispersión alrededor de la media.
   c. Que los precios sólo varían un 12% respecto al valor de la media.
   d. Que la mediana también será 8.40 euros.

4. ¿Cuál de las siguientes distribuciones sobre los precios de las habitaciones de un hotel tiene menos variabilidad?:
   a. Una que tiene de media 18 euros y desviación típica 3 euros.
   b. Una que tiene de media 18 euros y coeficiente de variación 42%.
   c. Una que tiene de media 18 euros y varianza 38.46 euros.
   d. Una que tiene de media 18 euros y desviación típica 1.80 euros.

5. Medidas de forma. Asimetría:
   a. El coeficiente de simetría de Fisher sólo puede tomar valores en el intervalo [0,1].
   b. Si a todos los valores de una variable se multiplican por una misma constante, el coeficiente de asimetría de Fisher no varía.
   c. Una asimetría a la izquierda es equivalente a una asimetría positiva.
   d. Una asimetría negativa es equivalente a una asimetría a la derecha.

6. Medidas de forma. Apuntamiento:
   a. Una distribución con un elevado apuntamiento se le considera leptocúrtica.
   b. En una distribución mesocúrtica el coeficiente de curtosis es igual a 3.
   c. Si a todos los valores de una variable se multiplican por una misma constante, el coeficiente de curtosis varía.
   d. El coeficiente de curtosis no puede tomar valores negativos.

7. Índice de Gini y curva de Lorenz:
   a. Un valor del índice de Gini negativo implica mínima concentración.
   b. Para el cálculo del índice de Gini es necesario previamente la ordenación de la variable de menor a mayor.
   c. Un valor del índice de Gini de cero implica mínima concentración.
   d. La curva de Lorenz sólo se utiliza cuando el índice de Gini no se puede calcular.

8. Tipificación de una variable:
   a. La media de una variable tipificada puede tomar cualquier valor en el intervalo [0,1].
   b. La varianza de una variable tipificada siempre es igual que su desviación típica.
   c. Una observación con valor tipificado de -4 se puede calificar de normal para la distribución de una variable.
   d. El coeficiente de variación de una variable tipificada no tiene sentido.

EJERCICIO 2

Utilizando el mismo fichero del ejercicio 2, del módulo 2 (*ejercicio_2_2.xls*), y a partir de los resultados del mismo, realiza las siguientes cuestiones:

a) Obtén los siguientes estadísticos descriptivos: (1) El recorrido, (2) la varianza y la desviación típica, (3) el coeficiente de variación de Pearson, (3) el coeficiente de asimetría de Fisher, y (4) el coeficiente de curtosis.

b) Tipifica cada uno de los valores de la variable y determina si existe alguna observación que podamos considerar atípica.

EJERCICIO 3

En el fichero *ejercicio_3_3.xls* se encuentran los valores de una determinada variable (X_i) y las frecuencias absolutas con que se repiten. A partir de esta información calcula el índice de Gini y dibuja la curva de Lorenz.

EJERCICIO 4

A partir de los datos presentados en la tabla del ejercicio 3, de los ejercicios de autocontrol del módulo 2, realiza los siguientes apartados:

c) Calcula los siguientes estadísticos descriptivos: (1) El recorrido, (2) la varianza y la desviación típica, (3) el coeficiente de variación de Pearson, (3) el coeficiente de asimetría de Fisher, y (4) el coeficiente de curtosis.

d) Tipifica cada uno de los valores de la variable. Comprueba que la media y la varianza de la variable son cero y uno, respectivamente.

EJERCICIO 5

Módulo 3

En la siguiente tabla se encuentran los datos correspondientes a las subvenciones (en miles de euros) que un determinado organismo público ha concedido a 86 establecimientos turísticos para mejorar la accesibilidad de sus instalaciones a personas con minusvalía. A partir de esta información determina la concentración del reparto mediante el cálculo del índice de Gini y la curva de Lorenz. Interpreta los resultados.

X_i	n_i
6.0	40
6.5	18
7.0	11
7.5	8
8.0	6
8.5	3

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

A continuación las afirmaciones correctas:

1. a, c
2. c, d
3. a
4. d
5. b
6. a, b, d
7. b, c
8. b, d

EJERCICIO 2

La solución del ejercicio se encuentra en el fichero *sol_ejercicio_3_2.xls*.

EJERCICIO 3

La solución del ejercicio se encuentra en el fichero *sol_ejercicio_3_3.xls*.

EJERCICIO 4

$(x-\bar{x})^3$	$(x-\bar{x})^3 f_i$	$(x-\bar{x})^4 f_i$	Z_i	$Z_i f_i$	$Z_i^2 f_i$
P-134.61	-10.10	51.74	-1.08	-0.08	0.09
-18.09	-9.04	23.74	-0.55	-0.28	0.15
-0.00	-0.00	0.00	-0.03	-0.01	0.00
962.97	168.52	1,664.13	2.08	0.36	0.76
810.27	149.38	1,739.61	0.42	0.00	1.00

Resultados:

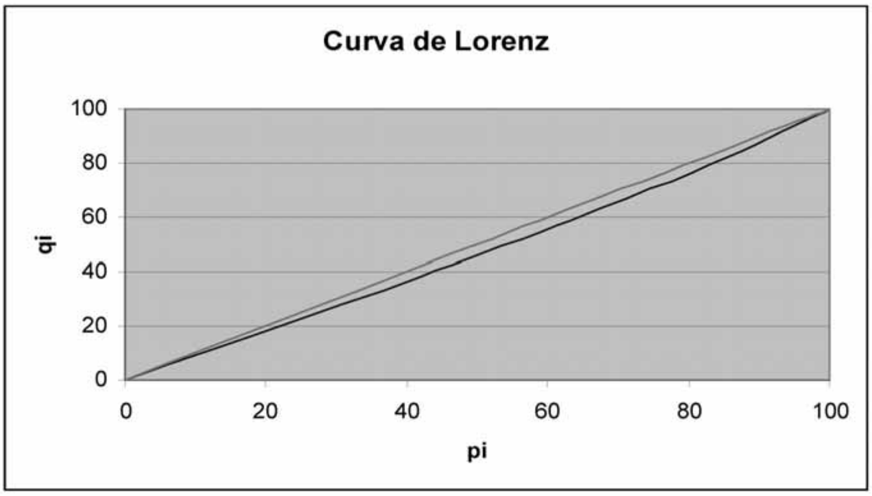
Varianza	22.48	C. Asimetría	1.40
D. típica	4.74	Curtosis	3.44
C. V. de P.	0.31		

EJERCICIO 5

X_i	n_i	N_i	$X_i n_i$	u_i	p_i	q_i	$p_i - q_i$
6.0	40	40	240	240	46.51	42.29	4.22
6.5	18	58	117	357	67.44	62.91	4.53
7.0	11	69	77	434	80.23	76.48	3.76
7.5	8	77	60	494	89.53	87.05	2.49
8.0	6	83	48	542	96.51	95.51	1.01
8.5	3	86	25.5	567.5	100	100	0
Total	86		567.5		380.23		16.00

$$I_G = \frac{\sum_{i=1}^{J-1}(p_i - q_i)}{\sum_{i=1}^{J-1} p_i} = \frac{16}{380.23} = 0.04$$

Un valor próximo a cero como 0.04 implica la existencia de poca concentración.



Módulo 3

En consonancia con el valor obtenido en el índice de Gini, la curva de Lorenz está muy próxima a la bisectriz, lo que implica mínima concentración, recordemos que cuando la curva de Lorenz se confunde con la bisectriz, el índice de Gini es igual a cero.

GLOSARIO DE TÉRMINOS

Coeficiente de asimetría de Fisher: sirve para determinar el grado de asimetría de la distribución de una variable.

Coeficiente de curtosis: sirve para determinar el grado de apuntamiento de la distribución de una variable.

Coeficiente de variación de Pearson: es el ratio entre la desviación típica de una variable y su media aritmética. Es una medida de dispersión.

Curva de Lorenz: sirve para determinar, mediante la elaboración de un gráfico, el grado de concentración de la distribución de una variable.

Desviación típica: es la raíz cuadrada de la varianza.

Distribución leptocúrtica: se dice de toda distribución con un apuntamiento mayor que el correspondiente a una distribución normal. Para este tipo de distribuciones el coeficiente de apuntamiento es mayor que tres ($K>3$).

Distribución mesocúrtica: se dice de toda distribución con un apuntamiento igual al correspondiente a una distribución normal. Para este tipo de distribuciones el coeficiente de apuntamiento es igual a tres ($K=3$).

Distribución planticúrtica: se dice de toda distribución con un apuntamiento inferior al correspondiente a una distribución normal. Para este tipo de distribuciones el coeficiente de apuntamiento es menor que tres ($K<3$).

Estandarización de una variable: ver tipificación de una variable en este mismo glosario.

Índice de Gini: índice que nos permite determinar el grado de concentración de la distribución de una variable.

Medidas de concentración: medias que nos permiten determinar el grado de reparto o equidistribución de la distribución de una variable.

Observación atípica: observación de una determinada variable que toma un valor muy alejado de lo que es habitual para ella.

Recorrido: diferencia entre el valor más grande y más pequeño de una variable.

Tipificación de una variable: proceso que consiste en restar a los valores de una variable su media y dividir este resultado por su correspondiente desviación típica.

Varianza: promedio del cuadrado de las diferencias entre los valores de una variable y su correspondiente media aritmética.

Módulo 4

Descripción bivariante. Regresión lineal simple

PRESENTACIÓN DEL MÓDULO

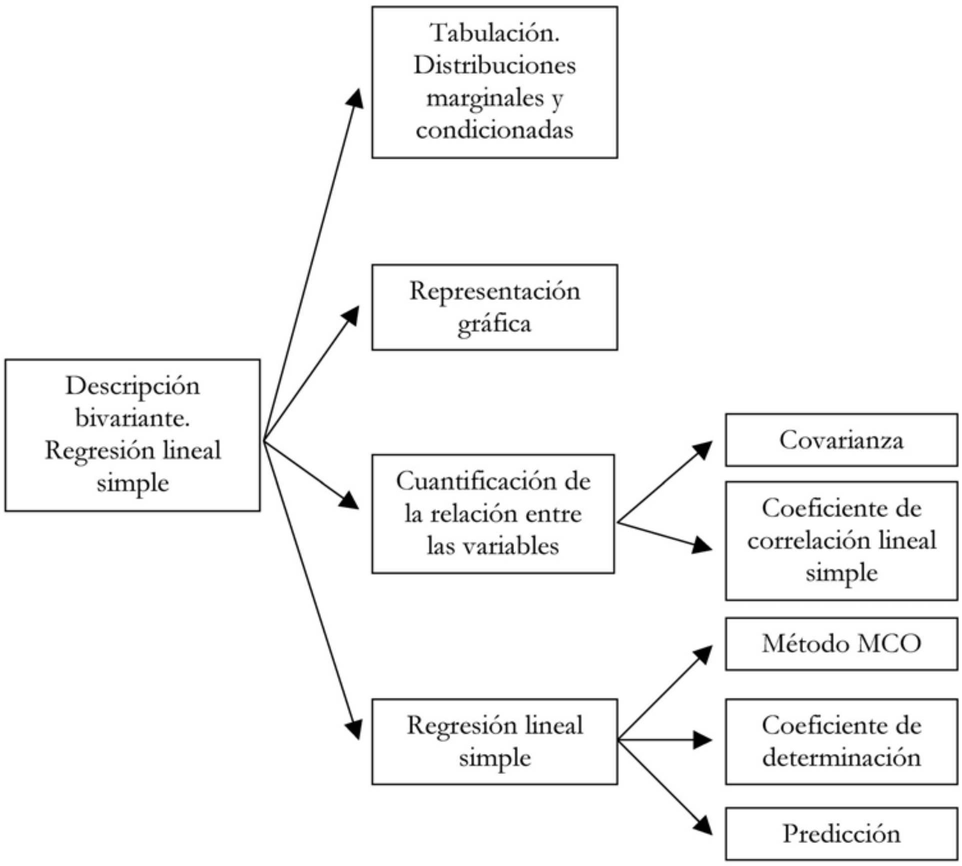
En los módulos 2 y 3 se ha estudiado la manera de describir las características más importantes de una variable. Sin embargo, en ocasiones, el estudio de más de una variable de forma conjunta puede ser nuestro foco de interés. En el ámbito de las Ciencias Sociales muchas variables están relacionadas entre sí, como por ejemplo, el precio de un determinado producto y su demanda, la inversión y los tipos de interés, la renta de las familias y sus gastos, etc. De esta manera, sería interesante contar con instrumentos que nos permitirán determinar efectivamente la existencia de estas relaciones, y por otro lado, conocer el grado o intensidad de ésta. En esta lección, éste será el objetivo principal. Así, se verá la manera de representar gráficamente dos variables para determinar la existencia de una relación, se estudiarán determinados estadísticos, como la covarianza y el coeficiente de correlación lineal, que nos permitirán establecer la relación entre variables y cuantificar el grado y signo de esta relación, y finalmente se verá una introducción al tema de la regresión lineal simple. Mediante esta técnica estadística seremos capaces, una vez determinada la relación entre dos variables, obtener estimaciones de una de ella a partir de valores de la otra. Esta posibilidad será especialmente interesante para realizar predicciones.

OBJETIVOS DEL MÓDULO

Fundamentalmente, existen cinco objetivos en este módulo que han de alcanzarse para que sea superado. Concretamente, estos objetivos son:

1. Tabular los valores de dos variables en una tabla de doble entrada y agruparlos por frecuencias absolutas.
2. Representar gráficamente una relación entre dos variables y saber interpretar el resultado.
3. Cuantificar y saber interpretar la relación entre dos variables mediante el cálculo de la covarianza y el coeficiente de correlación lineal simple.
4. Saber distinguir una variable *causa* de una variable *efecto*, y a partir de aquí ser capaz de calcular la recta de regresión que las relacionan.
5. Saber estimar y utilizar una recta de regresión.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. Introducción. Concepto de causalidad

En los módulos anteriores el foco de atención se centraba en una única variable, sobre la cual se obtenían diferentes estadísticos, lo que posibilitaba obtener una descripción de la misma. En esta lección, sin embargo, el foco de atención es el estudio descriptivo conjunto de dos variables. Así, se verán estadísticos que nos ayudarán a determinar la existencia o no de relación entre ellas, aspecto muy importante dentro del ámbito de la economía en general y de la empresa en particular.

Un concepto más preciso aún que el de relación entre variables es el de causalidad. Se dice que existe causalidad entre dos variables cuando el comportamiento de una de ellas *causa* al menos en parte el comportamiento de la otra. El estudio de la causalidad entre variables por parte de la Estadística, y más concretamente en la Econometría, es demasiado extenso para ser tratado en este manual, sin embargo, es importante tener claro algunos aspectos al respecto para, de esta manera, no cometer errores a la hora de utilizar algunos de los estadísticos que se verán a lo largo de este módulo. La utilización de los mismos sólo tendrán sentido cuando de antemano se presuponga la existencia de una relación causal entre las variables objeto de estudio. En ocasiones puede que no se conozca la dirección de la causalidad[1], pero en cualquier caso, hemos de tener clara su existencia para que los resultados que se obtengan sean interpretables, en caso contrario, los resultados de estos estadísticos no tendrán significado alguno, sobre todo cuando se trabaje con series temporales. En nuestro caso, esta relación de causalidad vendrá justificada por la propia teoría económica o la lógica de las cosas[2].

2. Tabulación. Distribuciones marginales y condicionadas

De la misma manera que en la lección anterior se vio como se tabulaban las observaciones de una determinada variable, en ésta se verá esta tabulación para el caso de dos variables relacionadas entre sí. Nos ayudaremos con un ejemplo. En la tabla 1 se presentan datos mensuales del porcentaje de ocupación y el coste, en euros, de la ropa de pisos de un determinado hotel. Tal y como se puede observar el tamaño muestral es de 48 meses, para cada mes se tiene el dato del *porcentaje de ocupación* y el *coste de la ropa de pisos*. Como se recordará, agrupar los datos por frecuencia absoluta tiene la ventaja de poder trabajar con una tabla más cómoda de tratar. Sin embargo, en este ejemplo, debido a que las frecuencias no se repiten no tendría sentido realizar esta opción. En esta situación, la única manera de poder trabajar con una tabla de menor dimensión sería agrupando mediante la utilización de intervalos[3]. Viendo los datos de la tabla 1, se pueden crear intervalos de amplitud 50 y 10 para el coste de la ropa de pisos y porcentaje de ocupación, respectivamente, empezando y terminando en 200 y 550 para la primera variable y 40 y 100 para la segunda. Con estos criterios se ha confeccionado la tabla 2. A este tipo de tablas se les denomina de doble entrada o de correlación[4].

1 Saber la dirección de la causalidad implica saber cuál es la variable *causa* y cuál la variable *efecto*. Por ejemplo, en la relación entre la renta y el consumo de las familias, la renta sería la variable *causa* y el consumo la variable *efecto*. El comportamiento de la variable renta *causa* parte del comportamiento de la variable *consumo*. Otros ejemplos serían las relaciones entre la *temperatura ambiente* (variable causa) y el *número de personas que acuden a la playa* (variable efecto), o el *consumo de tabaco* (variable causa) y padecer *cáncer de pulmón* (variable efecto). Sin embargo, no siempre es fácil determinar la dirección de la causalidad. Incluso, en ocasiones, esta relación puede ser bidireccional.

2 La utilización indiscriminada de algunos de los estadísticos que en esta lección se verán, sin una justificación razonada o fundamentada en alguna teoría que ligue las variables objeto de estudio, puede llevar a considerar la existencia de relaciones, donde sólo existen relaciones espurias. En la disciplina de la Econometría, y principalmente en el ámbito de las series temporales, se estudian procedimientos estadísticos para determinar la existencia de causalidad, y de esta manera distinguirlas de relaciones meramente espurias. Un autor destacado en este ámbito es Clive W.J. Granger premio Nóbel en Economía en el año 2003 por sus aportaciones en este sentido. Lo complejo de esta materia para un curso de Estadística básica aconseja que la relación de causalidad se establezca de la manera descrita en el texto.

3 Recordemos que tal y como se vio en la lección 2 esta opción supone pérdida de información.

4 Cuando la tabla de doble entrada se utiliza para variables cualitativas se le denomina de contingencia.

Tabla 1. Coste Ropa de Pisos [(Y_i) euros]. Porcentaje de Ocupación (X_i). Datos mensuales

Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i
296	48	503	86	350	60	418	74
518	97	405	75	484	84	398	68
415	72	316	61	512	89	441	83
457	78	337	60	443	77	388	69
482	89	372	59	493	86	357	67
474	82	271	45	430	77	433	76
445	79	246	44	496	91	524	92
362	65	447	73	501	87	473	76
302	57	401	70	402	74	420	72
317	53	381	68	441	80	319	58
360	62	460	88	336	54	403	69
520	87	533	97	491	94	419	81

Tabla 2. Tabla de doble entrada o de correlación

Módulo 4

	[40, 50)	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	Total
[200, 250)	1	0	0	0	0	0	1
[250, 300)	2	0	0	0	0	0	2
[300, 350)	0	4	2	0	0	0	6
[350, 400)	0	1	7	0	0	0	8
[400, 450)	0	0	1	11	3	0	15
[450, 500)	0	0	0	2	5	2	9
[500, 550)	0	0	0	0	4	3	7
Total	3	5	10	13	12	5	48

Como se puede observar en la tabla 2, los valores de las variables aparecen en la parte izquierda (coste de la ropa de pisos) y en la parte superior (porcentaje de ocupación). En el resto de la tabla aparecen las frecuencias absolutas conjuntas. Así, por ejemplo, el número de meses en que el porcentaje de ocupación estuvo entre el 50% y el 60% y el coste de la ropa de pisos entre 300 y 350 euros fueron 4 ¿Cuántos meses estuvo el porcentaje de ocupación entre el 60% y 70% y el coste de la ropa de pisos entre 350 euros y 400 euros? Además, en la última fila y columna se presenta la suma de cada una de ellas. Estas dos relaciones representan las frecuencias absolutas para cada una de las variables por separado, serían las mismas que se obtendrían en el caso de trabajar con cada una de las variables por separado, tal y como se hacía en los módulos 2 y 3. En este contexto, a estas distribuciones se les denominan *distribuciones marginales*. Así, la última fila corresponde a la distribución marginal del porcentaje de ocupación, mientras que la última columna a la distribución marginal del coste de la ropa de pisos. La suma de los valores de la última columna coincide con la suma de la última fila, siendo ésta igual al número de observaciones (N).

Tabla 3. Tabla de doble entrada de correlación

	45	55	65	75	85	95	Total
225	1	0	0	0	0	0	1
275	2	0	0	0	0	0	2
325	0	4	2	0	0	0	6
375	0	1	7	0	0	0	8
425	0	0	1	11	3	0	15
475	0	0	0	2	5	2	9
525	0	0	0	0	4	3	7
Total	3	5	10	13	12	5	48

Al resto de las filas y columnas por separado se les denomina *distribuciones condicionales*. Así, por ejemplo, la primera columna de frecuencias absolutas conjuntas representa a la distribución del coste de la ropa de pisos condicionada a que el porcentaje de ocupación esté entre un 40 y 50 por ciento. De la misma manera, la primera fila de frecuencias absolutas conjuntas corresponde con la distribución del porcentaje de ocupación condicionada a que el coste de la ropa de los pisos se encuentre entre 230 y 250 euros, y así sucesivamente para todas las filas y columnas, excepto las últimas.

Tabla 4. Agrupación por frecuencias Absolutas

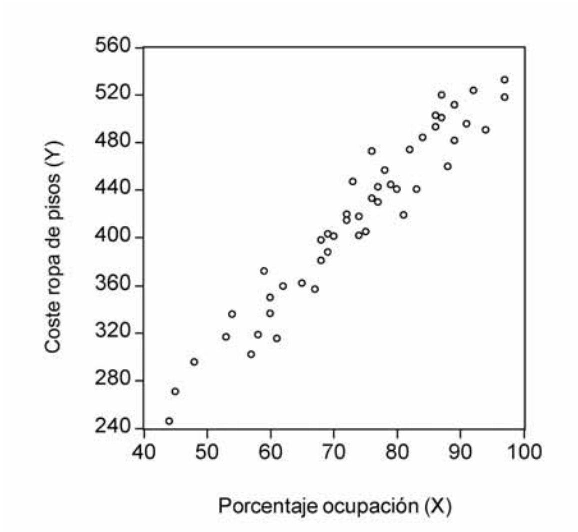
Y_i	X_i	n_i
225	45	1
275	45	2
325	55	4
325	65	2
375	55	1
375	65	7
425	65	1
425	75	11
425	85	3
475	75	2
475	85	5
475	95	2
525	85	4
525	95	3
Total		48

Ya se vio en el módulo 2 que cuando se agrupan los valores de las variables en intervalos era necesario calcular la marca de clase de cada uno de ellos para poder obtener los diferentes estadísticos. De la misma manera se debe proceder en el caso de trabajar con dos variables. Recordemos que la marca de clase es el valor que divide en dos partes iguales el intervalo, tal y como se definía en la expresión (2.5). En la tabla 3 se tienen los mismos resultados que en la tabla 2 pero en este caso utilizando la marca de clase. En este caso sí resulta interesante agrupar por frecuencias absolutas a diferencia de lo que ocurría con la tabla 1 donde las frecuencias no se repetían. Esta agrupación se presenta en la tabla 4. Con esta tabla resulta más cómodo el cálculo de los diferentes estadísticos.

3. Representación gráfica

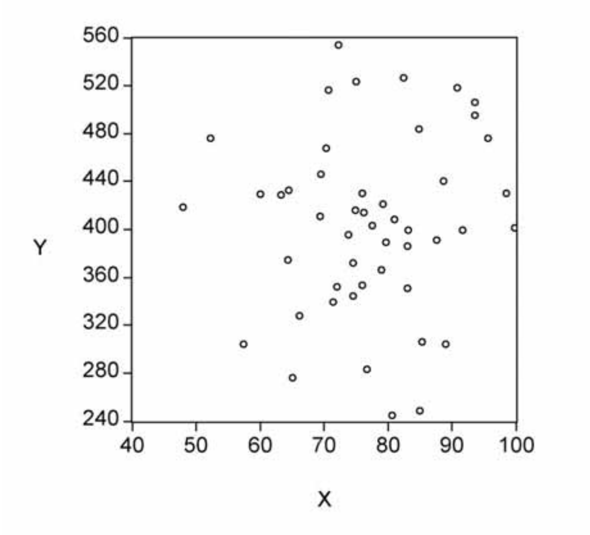
Una de las presentaciones gráficas más utilizada a la hora de relacionar variables es el gráfico de dispersión, también denominado nube de puntos. Este gráfico consiste en dibujar los puntos de cada una de las realizaciones de dos variable en el eje (X,Y). Así, para el ejemplo del porcentaje de ocupación y coste de la ropa de los pisos, cuyos datos se encuentran en la tabla 1, el gráfico de dispersión quedaría tal y como aparece en la figura 1.

Figura 1. Gráfico de dispersión o nube de puntos



El gráfico de dispersión permite en muchas ocasiones determinar visualmente la existencia de una relación entre las variables. Tal y como se observa en la figura 1, a medida que se incrementa el porcentaje de ocupación se incrementa el coste de la ropa de los pisos, tal y como cabría esperar inicialmente. En el caso de que no exista relación entre ambas variables, el gráfico de dispersión aparecería con los puntos esparcidos de forma aleatoria, sin dibujar ningún comportamiento claramente ascendente o descendente, tal y como se refleja en la figura 2.

Figura 2. Gráfico de dispersión o nube de puntos



4. Cuantificación de la relación entre variables

La covarianza

Ya hemos comentado con anterioridad el significado que le damos al hecho de que dos variables estén relacionadas. Desde un punto de vista funcional, esta relación puede adoptar diferentes formas. La más sencilla, y por otro lado la más habitual, es la lineal, supuesto que se mantendrá a lo largo de todo el módulo. La covarianza es un estadístico que se utiliza para medir la relación lineal entre variables. Su expresión es la siguiente:

Módulo 4

$$Cov(X,Y) = \frac{\sum_{i=1}^{J}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)n_i}{N} \tag{4.1}$$

como expresiones alternativas se tienen[5]:

$$Cov(X,Y) = \sum_{i=1}^{J}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)f_i \quad Cov(X,Y) = \frac{\sum_{i=1}^{J} X_i Y_i n_i}{N} - \overline{X}\overline{Y}$$

$$Cov(X,Y) = \sum_{i=1}^{J} X_i Y_i f_i - \overline{X}\overline{Y} \tag{4.2}$$

5 En este contexto de dos variables, J representa el número de pares de valores que toman las variables Y_i y X_i distintos entre sí.

Cuando el valor de la covarianza es menor que cero significa que la relación entre las variables es negativa, y por tanto, cuando una variable se incrementa (disminuye) la otra tiende a disminuir (incrementarse). Por el contrario, cuando el valor de la covarianza es mayor que cero la relación es positiva, y en este caso cuando una de las variables se incrementa (disminuye) la otra tiende a incrementarse (disminuir). En el caso de que la covarianza sea igual a cero no existe relación lineal entre ambas variables. En este caso se dice que las variables son linealmente independientes. En la tabla 5 se presentan los cálculos necesarios para la obtención de la covarianza en el ejemplo del porcentaje de ocupación y el coste de la ropa de pisos.

Tabla 5. Cálculo de la covarianza para el ejemplo de la tabla 4

	Y_i	X_i	n_i	$X_i n_i$	$Y_i n_i$	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})n_i$
	225	45	1	45	225	-28.54	-192.71	5500.22
	275	45	2	90	550	-28.54	-142.71	8146.27
	325	55	4	220	1300	-18.54	-92.71	6875.87
	325	65	2	130	650	-8.54	-92.71	1583.77
	375	55	1	55	375	-18.54	-42.71	791.88
	375	65	7	455	2625	-8.54	-42.71	2553.60
	425	65	1	65	425	-8.54	7.29	-62.28
	425	75	11	825	4675	1.46	7.29	116.97
	425	85	3	255	1275	11.46	7.29	250.65
	475	75	2	150	950	1.46	57.29	167.10
	475	85	5	425	2375	11.46	57.29	3282.34
	475	95	2	190	950	21.46	57.29	2458.77
	525	85	4	340	2100	11.46	107.29	4917.53
	525	95	3	285	1575	21.46	107.29	6906.90
Total			48	3530	20050			43489.58

$$\bar{X} = \frac{\sum_{i=1}^{J} X_i n_i}{N} = \frac{3530}{48} = 73.54 ; \qquad \bar{Y} = \frac{\sum_{i=1}^{J} Y_i n_i}{N} = \frac{20050}{48} = 417.71$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{J} (X_i - \bar{X})(Y_i - \bar{Y}) n_i}{N} = \frac{43489.58}{48} = 906.03$$

En este caso, el valor de la covarianza es mayor que cero lo que significa que la relación entre ambas variables es positiva; el incremento (disminución) de una ellas supone normalmente un incremento (disminución) en la otra, aspecto que ya se había comprobado al observar el gráfico de dispersión de la figura 1.

Hay dos cuestiones fundamentales a la hora de valorar la relación entre dos variables. En primer lugar, determinar si efectivamente existe tal relación y en segundo lugar cuantificar el grado de esta relación en el caso de que exista, ya que éste puede ir desde muy débil a muy fuerte. En cuanto a la primera cuestión, ya se comentó que esta relación existe siempre y cuando el valor de la covarianza sea distinto de cero. Sin embargo, ya conocemos por otros casos similares que se han estudiado, que la obtención de un valor exacto para un determinado estadístico es muy difícil debido a la aleatoriedad que conlleva toda variable estadística. Por esta razón, aún cuando el valor de la covarianza no sea exactamente igual a cero, valores muy próximos puede implicar la no existencia de relación entre las variables. En cuanto a la segunda cuestión, la cuantificación del grado de relación entre las variables, es muy difícil de establecerla teniendo en cuenta que la covarianza no está acotada en un determinado rango. De esta manera, un mismo valor de la covarianza para dos casos distintos puede significar cosas totalmente distintas. Por esta razón, en muchas ocasiones se utiliza como medida de relación entre dos variables el coeficiente de correlación lineal simple.

El coeficiente de correlación lineal simple

Este coeficiente se deriva a partir de la covarianza, pero tiene como ventaja frente a ésta, que sus valores están acotados en el intervalo [-1,1]. Su expresión es la siguiente:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{Dt(X)\,Dt(Y)} \tag{4.3}$$

sus valores se pueden interpretar de la siguiente manera:

- $\rho_{X,Y} = 0 \Rightarrow$ No hay relación lineal entre las variables.
- $\rho_{X,Y} = 1 \Rightarrow$ Existe una relación lineal *exacta* positiva entre las variables.
- $\rho_{X,Y} = -1 \Rightarrow$ Existe una relación lineal *exacta* negativa entre las variables.
- $0<\rho_{X,Y}<1 \Rightarrow$ Existe una relación lineal positiva entre las variables.
- $-1<\rho_{X,Y}<0 \Rightarrow$ Existe una relación negativa entre las variables.

Módulo 4

Gracias al acotamiento que presenta el coeficiente de correlación lineal, es más sencillo que para el caso de la covarianza, tener una idea del grado de relación entre las variables objeto de estudio. De esta manera, valores próximos a 1 ó -1 implican una relación alta. En el primer caso en sentido positivo y en el segundo caso en sentido negativo. Sin embargo, valores próximos a cero implica la ausencia de relación. Como criterio de referencia se pueden dar los siguientes rangos de valores para calificar el grado de la relación entre variables[6]:

$0.75 \le \lvert\rho_{X,Y}\rvert < 1$	Relación alta.
$0.25 \le \lvert\rho_{X,Y}\rvert < 0.75$	Relación mediana.
$0 < \lvert\rho_{X,Y}\rvert < 0.25$	Relación baja o inexistente.

6 Hemos de tener en cuenta que estos criterios son subjetivos y responden a la necesidad de tener criterios comunes a la hora de calificar las relaciones entre las variables. Otros criterios alternativos puede igualmente ser aceptados, siempre y cuando no caigan en extremos absurdos. Además, para determinar la exactitud del coeficiente de correlación lineal es igualmente importante tener en cuenta el tamaño muestral que se ha utilizado en su cálculo.

donde $|\cdot|$ es el operador de valor absoluto.

En la tala 6 se presentan los cálculos necesarios para obtener el coeficiente de correlación lineal simple entre el coste de la ropa de pisos y el porcentaje de ocupación del ejemplo que se está utilizando. Tal y como se puede observar en esta tabla, el valor del coeficiente de correlación lineal simple es 0.92, lo que implica que la relación lineal entre ambas variables es alta.

Tabla 6. Cálculo del coeficiente de correlación lineal simple

$\left(X_i-\overline{X}\right)^2 n_i$	$\left(Y_i-\overline{Y}\right)^2 n_i$
814.63	37136.50
1629.25	40731.34
1375.17	34379.34
145.92	17189.67
343.79	1824.00
510.72	12768.01
72.96	53.17
23.39	584.85
393.88	159.51
4.25	6564.67
656.47	16411.68
920.92	6564.67
525.17	46046.01
1381.38	34534.51
8797.92	254947.92

$$Dt(X)=\sqrt{\frac{\sum_{i=1}^{J}\left(X_i-\overline{X}\right)^2 n_i}{N}}=13.54$$

$$Dt(Y)=\sqrt{\frac{\sum_{i=1}^{J}\left(Y_i-\overline{Y}\right)^2 n_i}{N}}=72.88$$

$$\rho_{X,Y}=\frac{Cov(X,Y)}{Dt(Y)\,Dt(X)}=\frac{906.03}{72.88\times 13.54}=0.92$$

Módulo 4

5. Regresión lineal simple. El método de mínimos cuadrados ordinarios

Regresión lineal simple

El objetivo de la regresión lineal simple es determinar la relación de dependencia lineal que existe entre una variable Y_i, que se denomina dependiente[7], y otra X_i, que se denomina independiente[8].

7 En este momento cobra especial interés al concepto de causalidad que se introdujo al principio de este módulo. La variable Y_i depende y, por tanto, es causada por X_i y no al revés.

8 A la variable Y_i también se le denomina variable endógena, mientras que a la variable X_i también se le denomina variable explicativa o exógena.

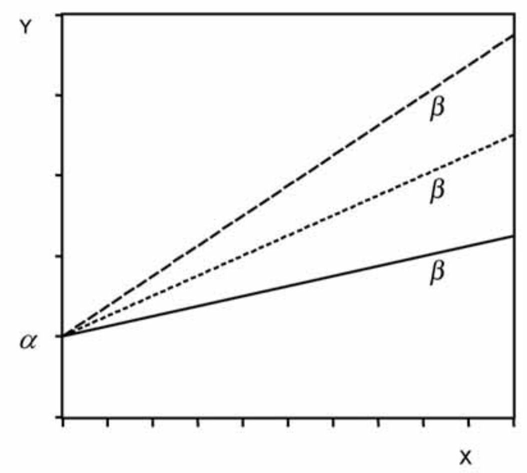
Este objetivo se lleva a cabo a partir de las observaciones que se tenga de ambas variables (Y_i,X_i).

La recta es la forma funcional lineal de dependencia que liga a dos variables tal que:

$$Y_i = \alpha + \beta X_i \tag{4.4}$$

donde α y β son los parámetros que la definen. Concretamente α es la ordenada en el origen de la recta mientras que β es su pendiente. En la figura 3 se observan tres rectas con la misma ordenada en el origen (α) pero con tres pendientes (β) diferentes. Mediante técnicas de regresión, y a partir de las observaciones que se tengan de las variables Y_i y X_i, se estiman estos parámetros. La cuestión a plantear sería: ¿cuál es la mejor ordenada en el origen y pendiente que hace que la recta se ajusta adecuadamente a las observaciones que se tienen de las dos variables? Un método que se utiliza para determinar el valor de α y β es el de Mínimos Cuadrados Ordinarios que se estudiará en el siguiente apartado.

Figura 3. Ejemplo de tres rectas con diferentes pendientes y la misma ordenada en el origen



Un aspecto importante a tener en cuenta cuando se intenta relacionar dos variables a través de la relación (4.4) es el siguiente: En la figura 1 se observa que la nube de puntos indica la existencia de una relación positiva entre ambas variables [coste de la ropa de pisos (Y_i) y porcentaje de ocupación (X_i)], sin embargo, esta relación no se produce exactamente a través de una recta, ya que en ese caso todos los puntos deberían estar consecutivamente uno detrás de otro, formando una recta, y obviamente esto no es lo que se refleja en la figura 1. O dicho de forma más concreta: En dos meses en los que el porcentaje de ocupación fuera el mismo, el coste de la ropa de pisos no tiene que ser exactamente el mismo. En la tabla 1 se puede observar que en dos meses en el que el porcentaje de ocupación fue del 77%, sin embargo, el coste de la ropa de pisos no fue el mismo; en un caso fue de 388 euros y en otro de 433 euros. Por esta razón, para que la igualdad que presenta (4.4) sea exacta para los valores de Y_i, hay que añadirle un término de error e_i que recoja esta diferencia, con lo que realmente:

$$Y_i = \alpha + \beta X_i + e_i \tag{4.5}$$

Para no confundir el valor de Y_i de la recta en (4.4) de la que aparece en (4.5), a la de (4.4) se le denomina Y_i estimada que se representa por $\hat{Y}_i$, con lo que:

$$\hat{Y}_i = \alpha + \beta X_i \tag{4.6}$$

De (4.5) y (4.6) se desprende fácilmente que:

$$Y_i = \hat{Y}_i + e_i \qquad \text{o que} \qquad e_i = Y_i - \hat{Y}_i \tag{4.7}$$

El método de Mínimos Cuadrados Ordinarios (MCO)

En la figura 4 se pueden observar tres rectas que pretenden ajustar los datos del ejemplo de la tabla 4. La recta B parece ser la que mejor se ajusta a los datos. La recta A tiene una pendiente más elevada que la que parece indicar la nube de puntos. Por otro lado, la recta C, aún cuando tiene la misma pendiente que la recta B, su desplazamiento hacia el lado derecho de la nube de puntos hace que su ordenada en el origen no sea la más adecuada. En la figura 5 se representa la misma nube de puntos pero en este caso la elección entre la recta A y B es más difícil, ambas parecen ajustarse más o menos bien a la nube de puntos. En estos casos sería interesante tener un método que nos permitiera elegir entre ambas rectas o incluso que rechazara ambas en favor de una alternativa que las mejorara. El método MCO es uno de los más utilizados en estos casos.

Módulo 4

Figura 4. Diferentes rectas que intentan ajustarse a la nube de puntos

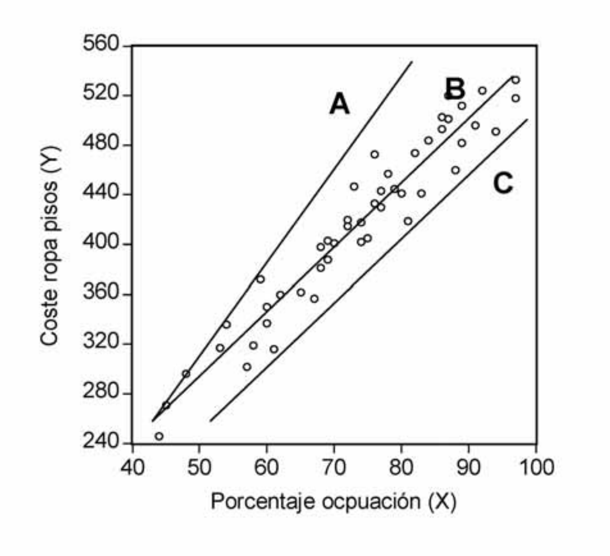
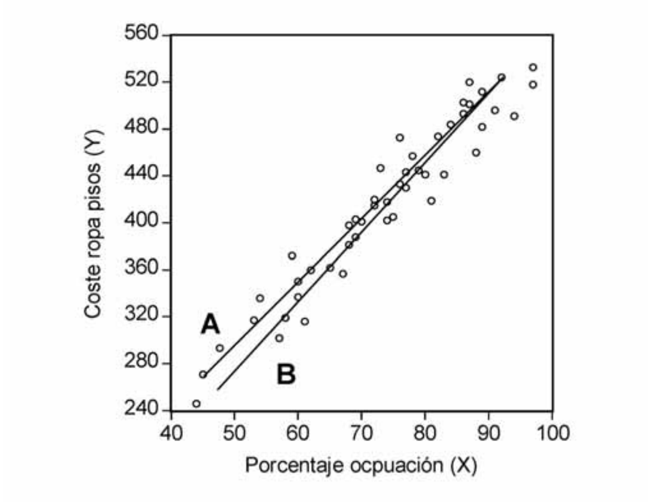


Figura 5. Diferentes rectas que intentan ajustarse a la nube de puntos



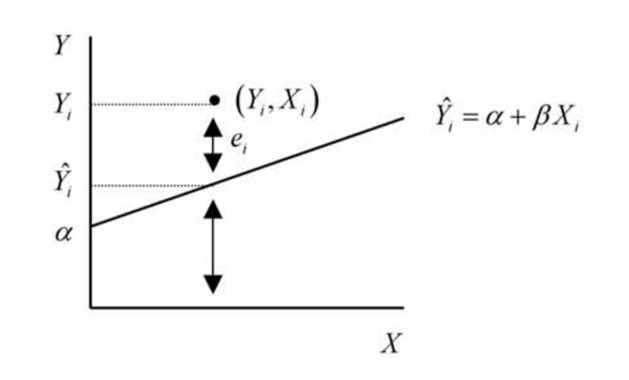
Gráficamente el error e_i de la euación (4.5), que representa la diferencia entre Y_i e $\hat{Y}_i$, tal y como se vio en (4.7), está recogido en la figura 6. El método MCO calcula unos valores para α y β de tal manera que los e_i sean lo más pequeños posibles para todas las observaciones. Concretamente minimiza el sumatorio del cuadrado de e_i. La razón de minimizar el cuadrado y no simplemente su suma se debe a que de esta manera se evita que los errores negativos resten en el sumatorio. Usando este criterio se obtiene que el valor de β es igual a[9]:

$$\beta = \frac{Cov(\mathrm{X,Y})}{Var(\mathrm{X})} \tag{4.8}$$

mientras que el valor de α es igual a:

$$\alpha = \overline{Y} - \beta \overline{X} \tag{4.9}$$

Figura 6. Representación del error (e_i)



9 La expresión que se minimiza es $\sum_{i=1}^{N} e_i^2$ o más concretamente su expresión equivalente $\sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2$. La minimización de esta expresión se realiza calculando sus derivadas respecto a α y β respectivamente, igualando a cero y despejando ambos parámetros de las expresiones resultantes. El desarrollo pormenorizado de este procedimiento se evita debido a que va más allá del objetivo planteado para este curso.

Tal y como se desprende de la expresión (4.8), el signo de β indica el tipo de relación entre la variable Y_i y X_i. Cuando el signo es positivo la relación será positiva, cuando la variable X_i se incrementa (disminuye), la variable Y_i tiende a incrementarse (disminuir). Por el contrario, cuando el signo es negativo, la relación será negativa con lo que un incremento (disminución) en la variable X_i tenderá a producir en la variable Y_i una disminución (incremento). Hay que tener en cuenta que el signo de la expresión (4.8) depende tan sólo del correspondiente signo de la covarianza entre ambas variables.

Siguiendo con el ejemplo del coste de la ropa de pisos, el valor de la pendiente será:

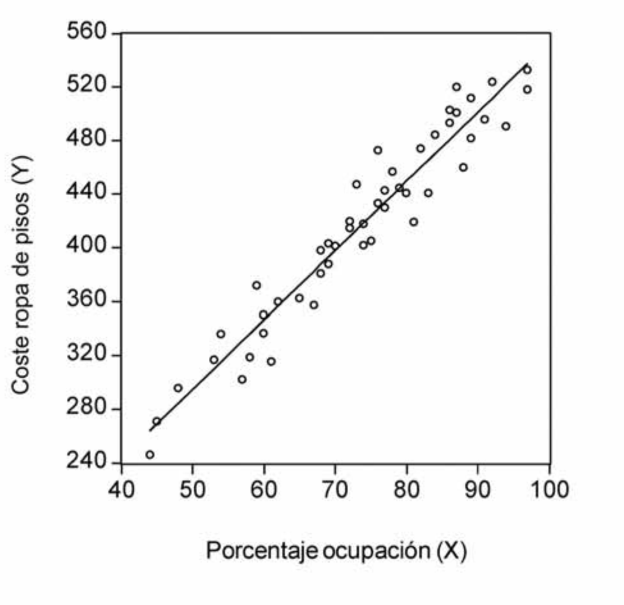
$$\beta = \frac{Cov(\mathrm{X,Y})}{Var(\mathrm{X})} = \frac{906.03}{13.54^2} = 4.94$$

y el de la ordenada en el origen: $\alpha = \overline{Y} - \beta\overline{X}$ = 417.71 - 4.94 x 73.54 = 54.18, lo que implica la siguiente recta de regresión:

$$\hat{Y}_i = 54.18 + 4.94X_i$$

En la figura 7 se puede observar la recta de regresión calculada atravesando la nube de puntos que representan las observaciones de los valores de X_i y Y_i. La recta calculada minimiza la suma del cuadrado de los errores, con lo que no es posible obtener otra recta que obtenga unos cuadrados de los errores más pequeños.

Figura 7. Recta de regresión



El coeficiente de determinación

Tal y como se ha puesto de manifiesto, el método MCO minimiza el sumatorio del cuadrado de los errores, lo que garantiza un buen ajuste a la nube de puntos. Sin embargo, aún así, puede ocurrir que la recta calculada implique la existencia de errores, que aunque sean los más pequeños posibles, sean demasiado grandes, con lo que el ajuste no será bueno. Por esta razón, es importante contar con alguna medida que permita determinar lo bien o lo mal que la recta se ajusta a la nube de puntos. El coeficiente de determinación es una de estas medidas, que se denota como R^2, y se obtiene a partir de cualquiera de las siguientes expresiones:

$$(a)\, R^2 = \rho_{X,Y}^2; \quad (b)\, R^2 = \frac{Var(\hat{Y})}{Var(Y)}; \quad (c)\, R^2 = 1 - \frac{Var(e)}{Var(Y)} \tag{4.10}$$

donde $Var(e) = \frac{\sum_{i=1}^{J} e_i^2 n_i}{N}$, dado que la media de e_i es cero. Por otro lado, es conveniente tener en cuenta, a la hora de calcular la varianza de $\hat{Y}_i$, que la media de Y_i es igual a la media de $\hat{Y}_i$.

Los posibles valores del coeficiente de determinación se encuentran en el intervalo [0,1]. Un valor próximo a uno significa que la recta de regresión se ajusta bastante bien a la nube de puntos, por el contrario valores próximos a cero, implican un mal ajuste. En los extremos, cuando $R^2 = 1$ el ajuste es perfecto, la nube de puntos forma una recta con los puntos alineados uno detrás de otro. Por el contrario, cuando $R^2 = 0$ el ajuste es muy malo y no existe ninguna relación lineal entre las variables. Evidentemente, estos casos extremos difícilmente se encontrarán en la realidad, lo habitual es encontrar valores entre ellos. Se pueden fijar algunos valores umbrales para tenerlos como referencia, siendo siempre conscientes del grado de subjetividad que supone la determinación de los mismos. Cuando $R^2 \geq 0,8$ vamos a considerar que la regresión presenta un buen ajuste. Valores inferiores supondrá una valoración menor del ajuste. Para el ejemplo del coste de la ropa de pisos, el coeficiente de determinación es: $R^2 = 0.92^2 = 0.85$, que se ha obtenido utilizando la expresión (a) de (4.10). Este resultado hace que se pueda calificar de bueno el ajuste que se ha conseguido con la recta de regresión estimada.

Predicción

Una de las utilidades más importantes de la regresión es la predicción. Una vez se han estimado los valores de α y β, es posible tener estimaciones de Y_i ($\hat{Y}_i$) a partir de (4.6). Así, por ejemplo, en el caso del coste de la ropa de pisos, si conocemos con anticipación cuál va a ser el porcentaje de ocupación del establecimiento turístico, se podrá realizar una predicción de cuál será el coste de la ropa de pisos, lo que sin duda ayudará a la hora de realizar los presupuestos del negocio. Sin embargo, siempre que se utilice la regresión para la predicción hemos de tener presente la fiabilidad del resultado obtenido. Esta fiabilidad puede ser indicada por el coeficiente de determinación. Cuando este coeficiente es calificado de “bueno”, tendremos cierta garantía de que el error que se producirá entre el valor predicho y el real no será normalmente muy grande. Por el contrario, cuando el coeficiente de determinación no es muy elevado, hemos de tener especial cuidado a la hora de utilizar la predicción en cualquier toma de decisiones. Otras consideraciones que se deben tener en cuenta a la hora de otorgar fiabilidad a las predicciones es el tamaño muestral o número de observaciones que se han utilizado en la estimación de α y β, cuanto mayor sea mejor. Por otra lado, la fiabilidad también viene determinada por lo alejado que esté el valor de X_i de su media correspondiente, cuanto mayor sea el alejamiento peor será la fiabilidad. Así, por ejemplo, supongamos que un determinado hotel tiene una media de ocupación del 75%, y desea predecir el coste de la ropa de pisos del próximo mes a partir del porcentaje de ocupación que se espera ese mismo mes. En la medida que el porcentaje de ocupación del próximo mes se aleja de la media del 75%, la predicción tenderá a ser menos fiable, por el contrario, entre más se acerque el porcentaje de ocupación del próximo mes al 75% más fiable será la predicción.

ACTIVIDADES

ACTIVIDAD 1

Recopila información sobre dos variables del ámbito turístico que consideres que se encuentran relacionadas, y a partir de dicha información realiza las siguientes cuestiones:

a) Construye una tabla agrupando los datos por frecuencias absolutas.

b) Dibuja una nube de puntos. ¿Qué conclusiones sacas a la vista del dibujo?

c) ¿Qué relación esperas que exista entre ambas variables, positiva o negativa? ¿Cuál es la variable independiente y cuál la variable dependiente?

d) Calcula la covarianza y el coeficiente de correlación lineal simple. Interpreta el resultado.

e) Calcula la recta de regresión que relacionaría ambas variables.

f) Calcula el coeficiente de correlación de la recta calculada. Interpreta el resultado.

g) Predice los valores de la variable dependiente a partir de los valores que tengas de la independiente.

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: teoría y ejercicios resueltos*. Málaga: Agora.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

1. Relación entre variables:
   a. En el ámbito de la economía y de la empresa en particular es muy difícil encontrar variables que estén relacionadas entre sí.
   b. Una vez se cuenten con las medidas estadísticas que veremos en este módulo, seremos capaces de medir la relación entre variables, independientemente de lo que diga la Teoría Económica o la propia lógica de las relaciones.
   c. Se dice que existe causalidad entre dos variables cuando una de ellas afecta al comportamiento de la otra.
   d. Se dice que existe causalidad entre dos variables cuando cada una de ellas afecta al comportamiento de la otra.

2. Tabulación y distribución de variables:
   a. Cuando una tabla de doble entrada se utiliza para variables cualitativas se le denomina de contingencia.
   b. Cuando se desea trabajar con una sola variable, independientemente de los valores que tome otra variable, debemos utilizar la distribución marginal de la variable de interés.
   c. Cuando se desea trabajar con una sola variable, condicionada a los valores que toma otra variable, debemos utilizar todas las frecuencias absolutas conjuntas.
   d. En una tabla de doble entrada, la suma de las frecuencias marginales de cada una de las variables deben coincidir.

3. La covarianza:
   a. La covarianza es una medida que mide la relación lineal y no lineal entre variables.
   b. La covarianza es independiente de las unidades de medida de las variables que intervienen para su cálculo.
   c. Para conocer el valor de la covarianza es necesario conocer la media de las variables que intervienen en su cálculo.
   d. El valor de la covarianza nos permite saber si la relación entre las variables es positiva o negativa.

4. El coeficiente de correlación lineal simple:
   a. El valor del coeficiente de correlación lineal simple siempre es positivo.
   b. El valor del coeficiente de correlación lineal simple está acotado entre los valores -1 y 1.
   c. El coeficiente de correlación lineal simple es independiente de las unidades de medida de las variables que intervienen para su cálculo.
   d. Un valor del coeficiente de correlación lineal simple de -0.90 indica una relación alta entre las variables.

5. Regresión lineal simple (I):
   a. La regresión lineal simple relaciona mediante una recta a dos variables.
   b. A la hora de calcular una recta de regresión es importante saber cuál es la variable *causa* y cuál la variable *efecto*.
   c. El objetivo de un ajuste por mínimos cuadrados ordinario de una regresión lineal simple, es obtener aquella recta que minimiza el sumatorio de los errores (e_i).
   d. El signo de la pendiente de la recta de regresión depende del signo de la covarianza entre la variable dependiente e independiente.

6. Regresión lineal simple (II):
   a. A medida que aumenta la varianza de la Y estimada ($\hat{Y}_i$) mejora el ajuste medido a través del coeficiente de determinación (R^2).
   b. La media de la variable dependiente es igual a la media de la variable dependiente estimada ($\hat{Y}_i$) en una regresión lineal simple de mínimos cuadrados ordinarios.
   c. El sumatorio de la variable dependiente estimada en una regresión lineal simple por mínimos cuadrados ordinarios es igual a cero.
   d. El sumatorio de los errores de la regresión lineal simple no es igual a cero, pero su media sí.

7. Regresión lineal simple (III):
   a. El coeficiente de determinación no puede ser negativo.
   b. Para poder realizar una predicción de la variable Y_i, a partir de una recta de regresión, es necesario conocer previamente el valor de la variable X_i.
   c. Una medida que nos puede ayudar para determinar la fiabilidad de una predicción, realizada con una regresión lineal simple, es el coeficiente de determinación.
   d. Una predicción es más fiable en la medida en que la variable X_i se acerque a su media aritmética.

Módulo 4

EJERCICIO 2

A continuación se presentan el precio y la demanda de unidades de *botellitas* de whisky de los minibares de un determinado hotel en los últimos 10 años.

Unidades	Precio (euros)
13943	1.05
13476	1.10
13419	1.45
13645	1.50
13557	1.50
13638	1.70
13253	1.80
13246	2.00
12578	2.10

A partir de esta información responder a las siguientes cuestiones:

a) ¿Qué relación esperas que exista entre ambas variables, positiva o negativa? ¿Cuál es la variable independiente y cuál la variable dependiente?

b) Calcula la covarianza y el coeficiente de correlación entre ambas variables. Interpreta el resultado.

c) Calcula la recta de regresión entre ambas variables.

d) Obtén la estimación de la demanda de *botellitas* de whisky y sus errores correspondientes. Además, calcula el coeficiente de determinación de la regresión.

e) ¿Qué efecto tiene sobre la demanda de *botellitas* de whisky el incremento de su precio en 0.50 euros? ¿Aconsejarías esta subida a la dirección del hotel, sabiendo que actualmente el precio es de 2.50 euros?

f) Dibuja la nube de puntos y sobre ésta la recta de regresión.

EJERCICIO 3

Se pretende estudiar el comportamiento diario del PH de una determinada piscina al final de cada día. Se supone que el PH esta relacionado con la temperatura media, en grados centígrados, que se ha producido a lo largo de ese mismo día. Para este propósito se cuenta con los siguientes datos diarios.

PH	6.63	6.46	6.69	7.04	6.89	6.54	6.68
Temperatura	28	31	24	20	21	26	25

a) ¿Qué relación existe entre el PH y la temperatura?. ¿Positiva o negativa?.

b) Para las temperaturas que aparecen en el cuadro de los datos, predice el PH correspondiente y obtén el error de dicha predicción. Además, calcula el coeficiente de determinación. ¿Qué conclusiones sacas a la vista del resultado?

c) El PH que debe tener la piscina al final del día ha de estar entre los valores 6.40 y 7.60. En caso contrario, se debe tratar el agua para que se recuperen valores normales. Si la temperatura media del día ha sido de 35 ºC, ¿qué predicción harías al final del día?, ¿se tendrá que tratar el agua de la piscina o no?

Justifica todas las respuestas mediante la obtención de alguna medida estadística adecuada al caso.

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

1. c, d
2. a, b, d
3. c, d
4. b, c, d
5. a, b, d
6. a, b
7. a, b, c, d

EJERCICIO 2

La solución del ejercicio se encuentra en el fichero *sol_ejercicio_4_2.pdf*.

EJERCICIO 3

Módulo 4

$PH(Y_i)$	$Temp.(X_i)$	$(X_i-\bar{X})(Y_i-\bar{Y})$	$(X_i-\bar{X})^2$	$(Y_i-\bar{Y})^2$	$\hat{Y}_i$	e_i	e_i^2
6.63	28	-0.22286	9	0.00552	**6.55906**	**0.0709**	0.00503
6.46	31	-1.46571	36	0.05968	**6.41383**	**0.0462**	0.00213
6.69	24	0.01429	1	0.00020	**6.75270**	**-0.0627**	0.00393
7.04	20	-1.67857	25	0.11270	**6.94633**	**0.0937**	0.00877
6.89	21	-0.74286	16	0.03449	**6.89792**	**-0.0079**	0.00006
6.54	26	-0.16429	1	0.02699	**6.65588**	**-0.1159**	0.01343
6.68	25	0.00000	0	0.00059	**6.70429**	**-0.0243**	0.00059
46.93	175	-4.26	88	0.24017			0.03395

N= 7

Media (Y) =	6.70429	Beta =	-0.04841
Media (X) =	25	Alpha =	7.91451
Cov (X,Y) =	**-0.6086**	Var(e)=	0.00485
Var (X) =	12.57143	R^2 =	**0.85865**
Var (Y) =	0.03431		
		$\hat{Y}_i(X_i = 35) =$	**6.22019**

GLOSARIO DE TÉRMINOS

Coeficiente de correlación lineal simple: coeficiente que cuantifica el grado de relación lineal entre dos variables.

Coeficiente de determinación: coeficiente que mide la bondad del ajuste de una regresión.

Covarianza: medida que determina el grado de relación lineal entre dos variables. A diferencia del coeficiente de correlación lineal simple, está afectada por las unidades de medida de las variables.

Distribución condicionada: distribución de una variable condicionada a los valores de otra.

Distribución marginal: distribución de una variable independientemente de los valores de otra variable.

Gráfico de dispersión: gráfico consistente en dibujar los puntos de cada una de las realizaciones de dos variable en el eje (X,Y).

Regresión lineal simple: procedimiento mediante el cual se ajusta una recta, con una única variable explicativa, a partir de las realizaciones de dos variables.

Método de mínimos cuadrados ordinarios: método mediante el cual se estiman los parámetros de una determinada regresión minimizando el sumatorio del cuadrado de los errores.

Relación espuria: relación aparente entre dos variables que sin embargo es falsa.

Tabla de correlaciones: tabla que recoge las distribuciones conjuntas de dos variables cuantitativas.

Tabla de contingencia: tabla que recoge las distribuciones conjuntas de dos variables cualitativas.

Tabla de doble entrada: ver tabla de correlaciones.

Módulo 5

Series temporales

PRESENTACIÓN DEL MÓDULO

Tal y como se vio en el módulo 1, las series temporales son el tipo de datos que representa una variable que se manifiesta a lo largo del tiempo, en contraposición, por ejemplo, con los datos de corte transversal, donde las observaciones están referenciadas a un mismo instante del tiempo. Una de las características más destacable de las series temporales es que sus observaciones están ordenadas en función de otra variable que es el tiempo.

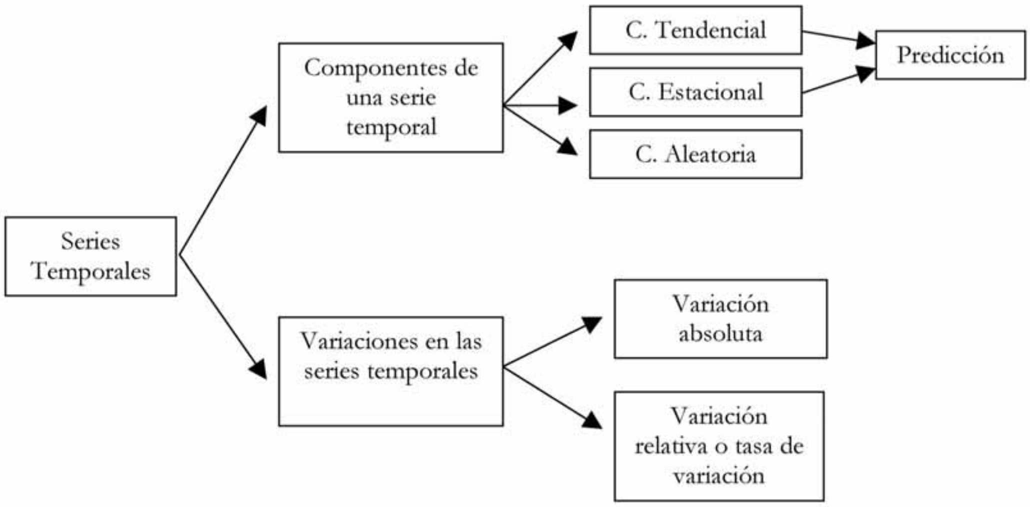
Muchas de las técnicas estadísticas que se han visto a lo largo del curso pueden ser aplicadas tanto a datos de corte transversal como a series temporales. Sin embargo, las que se verán en este módulo están dirigidas a ser utilizadas únicamente en series temporales. Fundamentalmente, estas técnicas intentan describir el comportamiento que estas variables mantienen a lo largo del tiempo, detectando sus patrones de comportamiento, utilidad muy útil si se pretende realizar predicciones de estas series.

OBJETIVOS DEL MÓDULO

La superación de este módulo supone alcanzar los siguientes objetivos:

1. Representar gráficamente una serie temporal e identificar en el mismo sus componentes.
2. Conocer las características que determinan las componentes de una serie temporal.
3. Ser capaz de obtener las diferentes componentes de una serie temporal.
4. Predecir la evaluación futura de una serie temporal.
5. Calcular y utilizar tasas de variación de una serie temporal.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. INTRODUCCIÓN

Módulo 5

Los procedimientos vistos en los módulos 2, 3 y 4 se pueden aplicar tanto a datos de corte transversal como a series temporales. Por el contrario, las técnicas que se verán en este módulo sólo son aplicables a series temporales. Tal y como ya se indicó en el módulo 1, una serie temporal se caracteriza por representar la evolución de una determinada variable a lo largo del tiempo, lo que implica la existencia de un orden preestablecido de carácter temporal. De esta manera, si se está estudiando la evolución temporal de la entrada anual de turistas en España, siempre el dato correspondiente al año 1995 irá antes que el correspondiente al año 1996. La intervención de una segunda variable, (en este caso el tiempo), que predetermina el orden no existe en los datos de corte transversal.

La periodicidad de los datos es otra cuestión de especial interés cuando se trabaja con series temporales. Ésta puede ser diaria, semanal, mensual, trimestral, anual, bianual, etc. Así, por ejemplo, si se dispone de datos mensuales o de periodicidad mensual, para cada año se tendrá 12 observaciones que multiplicadas por el número de años de los que se dispone se obtendrá el número total de observaciones. En la tabla 1 se presentan las estadísticas de la entrada de turistas en España para datos de periodicidad mensual, desde enero de 1995 a diciembre de 2004, con lo que se dispone de 10 años completos, con un número de observaciones igual a: 10 x 12 = 120. Si, por ejemplo, los datos de la tabla 1 en vez de periodicidad mensual fueran trimestral ¿cuál sería el número total de observaciones de la muestra? En este caso se tendrían: 10 x 4 = 40 observaciones.

En este módulo se estudiarán los diferentes componentes que pueden formar parte de los valores de una determinada serie temporal, principalmente con el objetivo de poder predecir la serie en el futuro. Si se es capaz de conocer con cierta exactitud cuáles son los patrones de comportamiento de la serie en el pasado, extrapolándolos al futuro se puede obtener una predicción de la misma. Evidentemente, la posibilidad de llevar a cabo esta acción depende de que la serie tenga elementos importantes de su comportamiento pasado susceptibles de ser sistematizados y, por tanto, reproducibles en el futuro. No siempre esto es posible debido a que la evolución temporal de muchas series son en gran parte aleatoria y, por tanto, carentes de un comportamiento sistemático de importancia que permitan extrapolar sus valores futuros. Ejemplo de este tipo de series se encuentran en los mercados financieros (índices bursátiles, tipos de interés, etc.). Sin embargo, existen otro tipo de series con unas pautas de comportamiento más predecibles, éstas serán las que preferiblemente nos interesarán en nuestro análisis.

Tabla 1. Entrada total de turistas en España

Años	ENE JUL	FEB AGO	MAR SEP	ABR OCT	MAY NOV	JUN DIC
1995	1397.41	1493.76	1978	2802.91	3261.81	3554.69
	5376.01	5043.68	3642.68	2860.26	1665.42	1842.95
1996	1455.41	1631.08	2265.31	2730.61	3371.58	3623.4
	5193.44	5409.62	3702.97	3070.35	1836.25	1930.99
1997	1590.06	1765.81	2606.06	2691.99	3882.9	3689.1
	5653.39	6051.17	4125.36	3320.08	2061.4	2115.39
1998	1768.23	1980.71	2508.11	3366.77	4376.67	4039.72
	5972.8	6363.47	4481.17	3844.27	2335.78	2358.38
1999	1966.65	2166.03	2905.49	3411.65	4566.17	4362.04
	6678.02	6643.77	4762.28	4429.67	2519.4	2364.7
2000	2054.97	2301.79	3071.03	4302.77	4192.24	4726.61
	6472.4	6174.2	5252.93	4264.54	2601.79	2482.65
2001	2303.3	2466.14	3179.44	4340.9	4502.59	5042.15
	6753.49	6859.61	5324.44	4137.15	2565.27	2619.07
2002	2263.36	2656.4	3795.35	3814.24	4647.01	5110.79
	6862.32	7824.51	5287.8	4473.84	2800.06	2791.08
2003	2436.93	2698.24	3344.84	4259.45	4755.32	5103.45
	6921.26	7289.87	5138.1	4403.22	2692.59	2786.33
2004	2544.55	2948.89	3348.99	4251.11	4973.19	4942.22
	6964.66	7228.03	5428.68	4789.1	2939.29	3145.27

Unidad: miles personas
Fuente: Instituto de Estudios Turísticos: FRONTUR

2. Representación gráfica y componentes de una serie temporal

La representación gráfica de una serie temporal se realiza en un eje de ordenadas bidimensional entre los valores de la variable de interés (eje Y) y los del tiempo (eje X). En la figura 1 se encuentra la representación gráfica de la serie temporal de la entrada de turistas en España de la tabla 1. Inicialmente se pueden destacar dos comportamientos en la evaluación de la serie: (1) Existe un patrón de comportamiento muy parecido dentro de cada uno de los años. En los meses de verano, principalmente julio y agosto, hay una elevada entrada de turistas, que se refleja en la figura mediante "picos", mientras que para los meses de invierno existen unas caídas en la entrada de turistas muy acusada, tal y como reflejan los "valles" de la figura 1. (2). Otra característica importante que se destaca en la figura 1, es la evolución al alza de la entrada de turistas. Si se toma como referencia un determinado mes y se comprueba la evolución del mismo mes a lo largo de los años se comprueba como a medida que pasa el tiempo la entrada de turistas se suele incrementar. Estos dos comportamientos deberían tenerse en cuenta la hora de realizar cualquier tipo de predicción sobre esta serie. Así, por ejemplo, si se deseara predecir la serie para el mes de agosto de 2005, lo más lógico sería dar una cifra parecida al mes de agosto del año 2004, a la que se incrementaría alguna cantidad, ya que se ha observado que normalmente cada mes de agosto supera al anterior en entrada de turistas. Para poder llevar a cabo este proceso con algo más de rigor es necesario identificar los componentes de la serie temporal. El valor de una serie temporal se puede descomponer en los siguientes elementos[1,2].

$$Y_t = T_t + S_t + C_t + A_t \tag{5.1}$$

donde T_t representa la componente tendencial en el momento t, S_t la componente estacional en el momento t, C_t la componente cíclica en el momento t, y A_t la componente aleatoria en el momento t. La definición de cada uno de estos componentes se presenta a continuación:

- **Componente tendencial:** evolución de la serie a largo plazo. Así, en la figura 1, se puede observar que la evolución de la entrada de turistas a largo plazo es creciente. O dicho de otra manera, la entrada de turistas tiene una tendencia creciente. Es habitual encontrar en las series económicas este tipo de tendencias (positivas y negativas en ocasiones) para el periodo de tiempo estudiado. Sin embargo, una misma serie pude contar a lo largo del tiempo con más de una tendencia diferenciada. Incluso una serie podría no tener tendencia.
- **Componente estacional:** son movimientos regulares que se producen en la serie en periodos inferiores al año. Por esta razón, para que una serie tenga componente estacional, es necesario que su periodicidad sea inferior al año. O sea, datos mensuales, trimestrales, etc. Nunca una serie de periodicidad anual o bianual, por ejemplo, tendrá componente estacional. En la figura 1 se puede apreciar claramente la existencia de una componente estacional. Se observa que en los

1 Hasta este momento se ha utilizado el subíndice i para identificar los diferentes valores que puede tomar la variable. Cuando se quiere recalcar que los datos con los que se están trabajando corresponden a una serie temporal es habitual utilizar el subíndice t.

2 Al esquema de combinación de los componentes de una serie temporal que aparece en la ecuación (5.1) se le denomina aditivo. Además de este esquema, existen otros como los multiplicativos, algunos de los cuales mediante transformaciones de las variables se pueden convertir en aditivos.

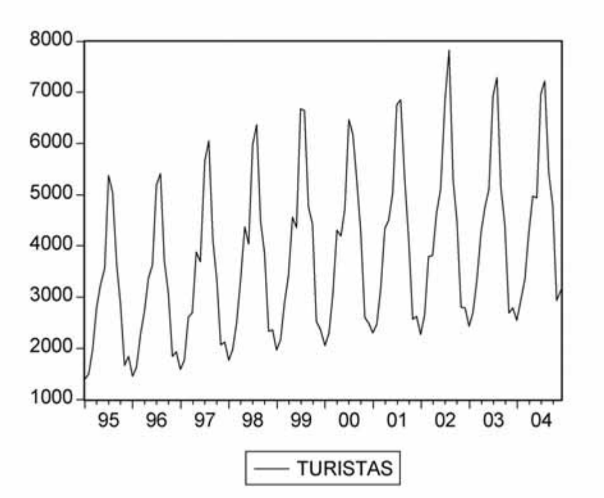
meses de verano de todos los años la entrada de turistas se incrementa, llegando a su valor máximo en los meses de agosto. Por el contrario, se observa que en los meses de invierno de todos los años la entrada de turistas se reduce sensiblemente. De la misma manera que la serie que se representa en la figura 1 tiene un marcado componente estacional, en otros casos esta componente puede ser menos acusada o incluso no existir.

- **Componente cíclica:** son movimientos regulares que se producen en la serie en periodos superiores al año. En la práctica, es la componente más difícil de determinar en el caso de que exista. Una serie puede tener más de una componente cíclica. Debido a que se produce en periodos superiores al año, es necesario tener series que abarquen periodos muy grandes para poder detectar su existencia. Desgraciadamente, no siempre es posible disponer de series económicas para periodos tan largos. Normalmente esta componente está relacionada con los ciclos económicos, esto hace que sea muy dificultoso determinar la amplitud del ciclo (años que dura el ciclo) ya que los ciclos económicos no siempre duran lo mismo. La dificultad que entraña su cálculo, como el hecho de que no siempre las series económicas presentan comportamientos cíclicos, hace que no se considere esta componente en el resto de los desarrollos del módulo, con lo que la expresión (5.1) se convierte en

$$Y_t = T_t + S_t + A_t \tag{5.2}$$

- **Componente aleatoria:**[3] son movimientos que no tienen un comportamiento regular definido y, por tanto, no se identifican con ninguna de las componentes anteriores. Normalmente está asociada a acontecimientos que producen movimientos en la serie que no son habituales a lo largo de la evolución de la misma. Por ejemplo, un incremento de las temperaturas en Europa un determinado verano puede incrementar más de lo habitual la asistencia de turistas a las costas españolas. O una determinada huelga en los aeropuertos reducir el número de visitantes a una determinada ciudad. Estos acontecimientos quedarán reflejados en la componente aleatoria al no tener una constancia en el devenir de la evolución de los acontecimientos.

Figura 1. Entrada de turistas en España desde enero 1995 a diciembre 2004



3 A esta componente también se le conoce con el nombre de componente residual o errática.

3. Cálculo de la tendencia. Método analítico

El cálculo de la tendencia mediante el método analítico consiste en estimar la siguiente regresión lineal simple:

$$T_t = \alpha + \beta t \qquad t = 1, 2,..., T \tag{5.3}$$

donde T representa el número de observaciones[4]. De esta manera, cada valor T_t, obtenido a través de la ecuación (5.3), representaría el valor de la tendencia en el momento t. Si adaptamos la ecuación (4.8) y (4.9) para la obtención de α y β en (5.3), se tendrá que:

$$\beta = \frac{Cov(Y,t)}{Var(t)} \tag{5.4}$$

mientras que:

$$\alpha = \overline{Y} - \beta \, \overline{t} \tag{5.5}$$

siendo $\overline{t}$ la media de la variable t. Así, por ejemplo, para el caso de la entrada de turistas en España, cuyos datos se encuentran en la tabla 5.1, a partir del desarrollo de la tabla 5.2, se pueden obtener los valores de α y β, de esta manera se tendrá que:

Módulo 5

$$\beta = \frac{Cov(Y,t)}{Var(t)} = \frac{\dfrac{\sum_{i=1}^{T}\left(Y_i - \overline{Y}\right)\left(t - \overline{t}\right)}{T}}{\dfrac{\sum_{i=1}^{T}\left(t - \overline{t}\right)^2}{T}} = \frac{\dfrac{2311686.586}{120}}{\dfrac{143990}{120}} = 16.05449$$

y

$$\alpha = \overline{Y} - \beta \, \overline{t} = 3804.309 - 16.05449 \times 60.5 = 2833.012$$

con lo que los valores de la tendencia de la entrada de turistas en España se calcularían mediante la siguiente expresión:

$$T_t = 2833.012 + 16.05449 \cdot t \qquad t = 1,2,...,T$$

4 Recordemos que hasta este momento el número de observaciones siempre se ha denotado por N, sin embargo, cuando se trabaja en un contexto exclusivo de series temporales, es habitual designar al número total de observaciones mediante la letra T.

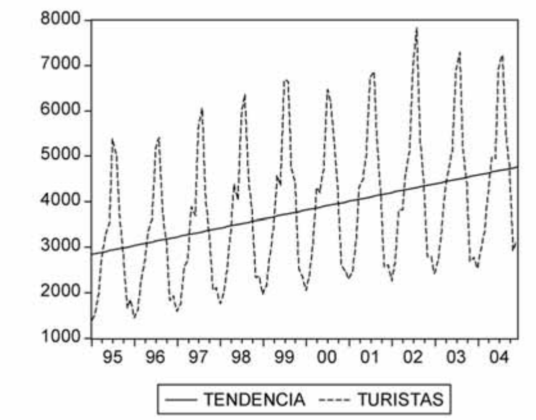
cuyos valores concretos se presentan en la penúltima columna (T_t) de la tabla 2. En la última columna de esta misma tabla se puede observar la evolución temporal de la serie sin tendencia, obtenida directamente como la resta entre la serie original (entrada de turistas) y la tendencia (T_t).

Tabla 2. Entrada de turistas (miles de personas) en España desde enero de 1995 a diciembre de 2004

(Y_t)	t	$(Y_t-\bar{Y})$	$(t-\bar{t})$	$(Y_t-\bar{Y})(t-\bar{t})$	$(t-\bar{t})^2$	T_t	(Y_t-T_t)
1397.410	1	-2406.899	-59.5	143210.488	3540.25	2849.067	-1451.657
1493.764	2	-2310.545	-58.5	135166.880	3422.25	2865.121	-1371.357
1978.003	3	-1826.306	-57.5	105012.592	3306.25	2881.176	-903.173
2802.909	4	-1001.400	-56.5	56579.097	3192.25	2897.230	-94.321
3261.808	5	-542.501	-55.5	30108.803	3080.25	2913.285	348.523
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
5428.684	117	1624.375	56.5	91777.190	3192.25	4711.388	717.296
4789.098	118	984.789	57.5	56625.370	3306.25	4727.442	61.656
2939.293	119	-865.016	58.5	-50603.433	3422.25	4743.497	-1804.204
3145.272	120	-659.037	59.5	-39212.699	3540.25	4759.551	-1614.279
456517.07	7260			2311686.586	143990		

El la figura 2 se puede observar la recta que representa la tendencia para cada instante t, en contraposición con la serie original. Tal y como ya se observaba en la serie original de la figura 1, la tendencia de la serie para el periodo estudiado es claramente positiva.

Figura 2. Evaluación de la entrada de turistas en España desde enero de 1995 a diciembre de 2004 y su tendencia



Veamos otro ejemplo, en la primera columna de la tabla 3 se encuentran los datos del número de toallas de la piscina de un determinado complejo turístico, que cuatrimestralmente se han tenido que reponer en los últimos tres años. A partir de esta información se desea conocer la tendencia que ha presentado la serie. Para llevar a cabo este cometido, mediante el método analítico, se obtienen el resto de las columnas que se presentan en la tabla 3. Concretamente, los valores de la tendencia (T_t) se obtienen a partir de la expresión (5.3) sustituyendo los valores de α y β por:

$$\beta = \frac{Cov(Y,t)}{Var(t)} = \frac{\frac{\sum_{i=1}^{T}(Y_i - \overline{Y})(t-\overline{t})}{T}}{\frac{\sum_{i=1}^{T}(t-\overline{t})^2}{T}} = \frac{\frac{377}{9}}{\frac{60}{9}} = 6.2833$$

y

$$\alpha = \overline{Y} - \beta\,\overline{t} = \frac{739}{9} - \frac{45}{9} \times\ 6.2833 = 50.69$$

En este caso la expresión de la tendencia será:

$$T_t = 50.69 + 6.2833 \cdot t$$

Módulo 5

En la penúltima columna de la tabla 3 se presentan los diferentes valores de la tendencia para cada uno de los instantes del tiempo de la muestra (t = 1,2,...T). Finalmente, en la última columna de esta misma tabla se encuentra la serie sin tendencia, que se obtiene restando a cada uno de los valores de la serie original su tendencia correspondiente.

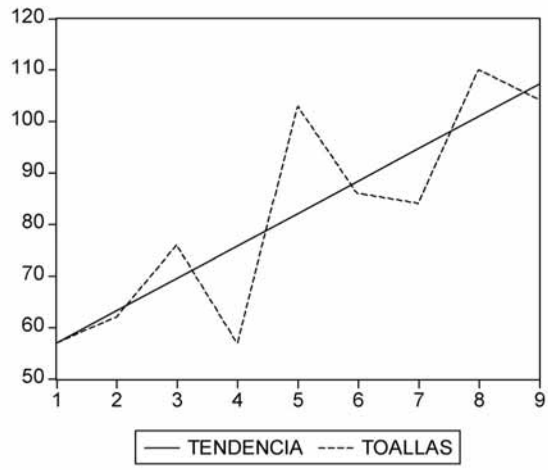
Tabla 3. Número de toallas de piscina repuestas cuatrimestralmente

Y_t	t	$(Y_t - \overline{Y})$	$(t-\overline{t})$	$(Y_t - \overline{Y})(t-\overline{t})$	$(t-\overline{t})^2$	T_t	$Y_t - T_t$
57	1	-25.11	-4	100.44	16	56.98	0.02
62	2	-20.11	-3	60.33	9	63.26	-1.26
76	3	-6.11	-2	12.22	4	69.54	6.46
57	4	-25.11	-1	25.11	1	75.83	-18.83
103	5	20.89	0	0.00	0	82.11	20.89
86	6	3.89	1	3.89	1	88.39	-2.39
84	7	1.89	2	3.78	4	94.68	-10.68
110	8	27.89	3	83.67	9	100.96	9.04
104	9	21.89	4	87.56	16	107.24	-3.24
739	45			377	60		

En la figura 3 se puede observar la evaluación de las toallas repuestas, además de su tendencia. Tal y como se puede comprobar, la tendencia de la serie es positiva.

Antes de terminar este apartado es importante tener en cuenta que, tal y como se ha indicado en el epígrafe anterior, una misma serie temporal puede tener más de una tendencia a lo largo del tiempo. Este resultado es evidente, por ejemplo, cuando una misma serie temporal presenta un periodo con tendencia positiva y otro con tendencia negativa. En estos casos, se tendrá que aplicar separadamente el método analítico para el cálculo de la tendencia en cada uno de los periodos, de tal manera que se estimarán dos regresiones, una para la tendencia positiva y otro para la tendencia negativa. Además, si el objetivo es la predicción, sólo será necesario trabajar con el último periodo que presenta la misma tendencia.

Figura 3. Toallas repuestas cuatrimestralmente y su tendencia



Módulo 5

4. Cálculo de la componente estacional. El método de las medias estacionales sin tendencia

Tal y como se definió la estacionalidad en el apartado 2, ésta sólo se pone de manifiesto en series temporales de periodicidad inferior al año, cuando al comparar las observaciones de los diferentes años entre sí, se comprueba la existencia de comportamientos regulares que se repiten en el mismo sentido cada año. En el ámbito de la economía, y de la empresa en particular, es habitual que las series temporales reflejen comportamientos estacionales.

Veamos con la ayuda del ejemplo de la reposición de toallas de piscina, cuyos datos se presentaron en la tabla 3, cómo obtener la componente estacional de una serie temporal por el método de las medias estacionales sin tendencia. En la tabla 4 se encuentran los valores sin tendencia de esta serie temporal, que se presentaron inicialmente en la última columna de la tabla 3 (Y_t - T_t). En este caso, sin embargo, la ordenación de los datos en la tabla 4 es algo diferente para facilitar los cálculos. Tal y como se puede observar en esta tabla, existe una agrupación por cuatrimestres, de tal manera que los datos correspondientes a los primeros cuatrimestres están en una columna, los correspondientes a los segundos cuatrimestres en otra y a los terceros cuatrimestres en la última. Una vez en este orden, se obtiene para cada uno de los cuatrimestres la suma total y la media, cuyo valor corresponde con la componente estacional. De esta manera, la suma de los primeros cuatrimestre es igual a -29.48, y su componente estacional o media igual a: -9.83, para los segundos cuatrimestres la suma es igual a 28.67 y la componente estacional o media igual a 9.56 y para los terceros cuatrimestres 0.82 es la suma total y 0.27 la componente estacional o media.

Tabla 4. Número de toallas de piscina repuestas cuatrimestralmente

Años	1er Cuatrimestre	2º Cuatrimestre	3er Cuatrimestre
1er Año	0.02	-1.26	6.46
2º Año	-18.83	20.89	-2.39
3er Año	-10.68	9.04	-3.24
Total	-29.48	28.67	0.82
Media (C. Estacional)	-9.83	9.56	0.27

En la tabla 5 se presentan los desarrollos necesarios para la obtención de la componente estacional para el caso de la entrada de turistas en España en el periodo 1995-2004, cuyos datos se presentaron en la tabla 1. Tal y como se puede comprobar, la serie sin tendencia que se obtuvo en la última columna de la tabla 2, se agrupa en la tabla 5 por meses, de tal manera que los meses de enero aparecen en una primera columna, los de febrero en la siguiente, y así sucesivamente hasta el mes de diciembre. En la última fila de la tabla 5 se encuentra la media para cada uno de los meses, que corresponden con cada uno de los componentes estacionales. En la tabla 6 se pormenoriza cada uno de estos cálculos.

La componente estacional corrige los valores que la tendencia determina para cada uno de los periodos de la serie temporal. Así, por ejemplo, para el caso de la entrada de turistas en España, sobre el valor esperado de la tendencia para el mes de agosto se incrementa el número de turistas que se espera entren en España debido a que este mes tiene una componente estacional claramente positiva. Lo contrario ocurriría, por ejemplo, en el mes de enero, donde la estacionalidad de este mes hace que cualquier valor esperado de la tendencia se vea reducido.

Tabla 5. Cálculo componente estacional. Entrada total de turistas en España

Años	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
1995	-1451.657	-1371.357	-903.173	-94.321	348.523	625.350	2430.614	2082.229	665.172	-133.297	-1344.190	-1182.715
1996	-1586.314	-1426.694	-808.522	-359.279	265.642	501.403	2055.397	2255.518	532.812	-115.857	-1366.011	-1287.327
1997	-1644.315	-1484.616	-660.420	-590.552	584.308	374.457	2322.693	2704.415	762.551	-58.789	-1333.518	-1295.582
1998	-1658.794	-1462.374	-951.031	-108.419	885.426	532.421	2449.445	2824.057	925.710	272.753	-1251.797	-1245.250
1999	-1653.032	-1469.709	-746.296	-256.197	882.274	662.085	2962.010	2911.707	1014.160	665.497	-1260.827	-1431.587
2000	-1757.369	-1526.601	-773.419	442.270	315.683	833.999	2563.739	2249.483	1312.156	307.715	-1371.089	-1506.284
2001	-1701.689	-1554.904	-857.658	287.751	433.384	956.889	2652.171	2742.238	1191.012	-12.336	-1600.265	-1562.518
2002	-1934.281	-1557.299	-434.403	-431.564	385.148	832.869	2568.354	3514.483	961.719	131.701	-1558.125	-1583.161
2003	-1953.368	-1708.110	-1077.571	-179.014	300.800	632.882	2434.631	2787.194	619.371	-131.570	-1858.252	-1780.566
2004	-2038.398	-1650.118	-1266.071	-380.005	326.017	278.995	2285.383	2532.695	717.296	61.656	-1804.204	-1614.279
Totales	-17379.218	-15211.781	-8478.565	-1669.329	4727.204	6231.351	24724.436	26604.020	8701.958	987.473	-14748.280	-14489.269
Medias (C. Estacional)	-1737.922	-1521.178	-847.857	-166.933	472.720	623.135	2472.444	2660.402	870.196	98.747	-1474.828	-1448.927

Módulo 5

Tabla 6. Componente estacional de la entrada de turistas en España (1995-2004)

Mes	Componente estacional	Mes	Componente estacional
Enero	$-17379.218/10 = -1737.922$	Julio	$24724.436/10 = 2472.444$
Febrero	$-15211.781/10 = -1521.178$	Agosto	$26604.020/10 = 2660.402$
Marzo	$-8478.565/10 = -847.857$	Septiembre	$8701.958/10 = 870.196$
Abril	$-1669.329/10 = -166.933$	Octubre	$987.473/10 = 98.747$
Mayo	$4727.204/10 = 472.720$	Noviembre	$-14748.280/10 = -1474.282$
Junio	$6231.351/10 = 623.135$	Diciembre	$-14489.269/10 = -1448.927$

5. Predicción

La predicción de las series temporales puede ser un objetivo de los responsables económicos y empresariales. De esta manera se consigue reducir incertidumbre en la toma de decisiones, lo que se traduce en una mejor gestión pública o empresarial. Evidentemente, esto será así en la medida en que las predicciones sean fiables, y por tanto, los errores que se cometen no sean excesivamente elevados. Por esta razón, es importante contar con instrumentos capaces de utilizar la información disponible de forma eficiente para de este modo realizar buenas predicciones. Un segundo elemento que determina la fiabilidad de la predicción, es el grado de aleatoriedad no predecible propia de la serie temporal. Aún cuando los instrumentos que se utilicen para la predicción de una serie sean los mejores posibles, si ésta tiene un elevado componente aleatorio no predecible, difícilmente se podrá determinar su evolución futura. Sin embargo, aquellas series temporales que presentan importantes comportamientos sistemáticos en el pasado serán más fácilmente predecibles, dado que una vez que éstos son identificados, servirán para que su extrapolación hacia el futuro permita disponer de predicciones con bastante robustez.

Existen dos cuestiones importantes en relación a la información que se utiliza a la hora de realizar una predicción. Por un lado, hay que determinar el tipo de información que se utilizará, que normalmente vendrá determinada por el tipo de instrumentos que se maneja. Ésta puede ser la evolución pasada de la propia serie temporal, la evolución pasada de otras series temporales o incluso información cualitativa proveniente de la opinión de expertos. Por otro lado hay que determinar el momento temporal en que la información estará disponible para poder llevar a cabo la predicción. Así, por ejemplo, si se pretende predecir el valor de Y_{t+1} en el momento t, con la información de la variable X_{t+1}, no conoceríamos el valor de esta última variable al encontrarnos en el momento t, de esta manera se tendrá que tener, a su vez, una predicción de la misma. Concretamente, y para las técnicas que se han utilizado en este módulo, la información que se utilizará para llevar a cabo la predicción será la evolución pasada de la misma serie que se pretende predecir.

Supongamos que para una determinada serie temporal Y_t se dispone de su evolución para $t = 1,2,...,T$, y se desea predecir el valor que tendrá esta variable en el momento $T + 1$, o sea Y_{T+1}.

Teniendo en cuenta la ecuación (5.2), en este caso:

$$Y_{T+1} = Y_{T+1} + S_{T+1} + A_{T+1} \tag{5.6}$$

De los tres componentes que forman Y_{T+1}, según la ecuación (5.6), sería posible determinar en el momento T, la tendencia y la estacionalidad para el momento $T + 1$. Solamente la componente aleatoria no podría ser calculada. De esta manera, parece lógico considerar como predicción de Y_{T+1} aquellos componentes que en el momento T son conocidos; la tendencia y la estacionalidad. Concretamente, y denotando la predicción de Y_{T+1} como $\hat{Y}_{T+1}$, se tendrá que:

$$\hat{Y}_{T+1} = T_{T+1} + S_{T+1} \tag{5.7}$$

siendo $T_{T+1} = \alpha + \beta(T + 1)$ tal y como se deduce de la ecuación (5.3) . Dado que el valor de la tendencia como el de la componente estacional se pueden obtener en el momento T para periodos más allá de $T + 1$, como $T + 2$, $T + 3$,..., la ecuación (5.7) se puede generalizar para predicciones hasta h periodos hacia delante de tal manera que:

$$\hat{Y}_{T+h} = T_{T+h} + S_{T+h} \qquad h = 1, 2, 3, \ldots \tag{5.8}$$

Es evidente que a medida que el horizonte de la predicción se incrementa, o lo que es lo mismo, a medida que h se incrementa, la fiabilidad de la predicción será menor, con lo que el error de predicción será mayor. Teniendo en cuenta las ecuaciones (5.6) y (5.8), el error de la predicción h periodos hacia delante (e_{T+h}) es igual a la componente aleatoria, con lo que:

Módulo 5

$$e_{T+h} = A_{T+h} = Y_{T+h} - \hat{Y}_{T+h} \qquad h = 1, 2, 3, \ldots \tag{5.9}$$

Tabla 7. Predicción de la reposición de toallas

t	Componente tendencial	Componente estacional	Predicción
10	113.53	-9.83	103.7
11	119.81	9.56	129.37
12	126.09	0.27	126.36

Supongamos que la gerencia que administra la reposición de las toallas de piscina, cuyos datos se presentaron en la primera columna de la tabla 3, desea realizar una predicción de las toallas que se tendrán que reponer a lo largo de los tres cuatrimestres del próximo año que no aparece en la tabla. Para este caso, la predicción se presenta en la tabla 7, donde la componente tendencial para cada uno de los cuatrimestres se ha obtenido a partir de la ecuación (5.3) en el momento $T + 1$, $T + 2$ y $T + 3$. Dado que $T = 9$,

$$T_{T+1} = \alpha + \beta\,(T + 1) = T_{10} = 50.69 + 6.28 \cdot 10 = 113.53$$

$$T_{T+2} = \alpha + \beta\,(T + 2) = T_{11} = 50.69 + 6.28 \cdot 11 = 119.81$$

$$T_{T+3} = \alpha + \beta\,(T + 3) = T_{12} = 50.69 + 6.28 \cdot 12 = 126.09$$

siendo los valores de α y β los que se habían calculado en el apartado 3. Finalmente, tal y como se puede apreciar en la última columna de la tabla 7, la predicción se obtiene utilizando la ecuación (5.8); sumando la componente tendencial y la componente estacional, ésta última obtenida en el apartado 3.

Si a la hora de realizar la predicción de una determinada serie temporal, ésta no presenta componente estacional, no es necesario realizar su cálculo ni incluirla como elemento de la predicción. Sin embargo, en el caso de la tendencia el procedimiento no es el mismo. La tendencia siempre ha de ser un elemento que interviene en la predicción de la serie, aún cuando la serie no tenga tendencia. Esto es así, debido a que la recta de regresión que se utiliza para la obtención de la tendencia, mediante el método analítico, aparte de la evolución creciente o decreciente de la serie a largo plazo que se refleja en la pendiente de la recta (β), recoge, igualmente, el nivel de la serie en la ordenada en el origen (α).

Las predicciones pueden ser extra-muestrales y muestrales. En las predicciones extra-muestrales se predicen valores que no se encuentran en la muestra que dispone el investigador, o que por lo menos no han sido utilizados por éste para obtener los parámetros que determinan la tendencia (α y β) y la componente estacional. Un ejemplo de predicción extra-muestral es el que hemos descrito en el ejemplo anterior, en este caso, además, el error que se ha cometido con la predicción no puede ser determinado hasta que transcurra el tiempo y se conozca el verdadero valor de la serie, con lo que el error se podría determinar en ese momento mediante la ecuación (5.9)[5]. Por el contrario, en la predicción muestral el objetivo es predecir valores que sí se han utilizado en la determinación de los parámetros de la tendencia y de la componente estacional. En este caso siempre será posible determinar el error de predicción o componente aleatoria. La utilidad de este tipo de predicciones va principalmente dirigida a valorar el peso que la componente aleatoria tiene en relación con la componente tendencial y estacional, ya que esto nos puede determinar la capacidad predictiva del modelo. Entre mayor es el peso de esta componente, peor se explican los movimientos de la serie, por el contrario un peso pequeño de esta componente en relación a las otras, implicará una capacidad importante de explicar la evolución temporal de la serie.

6. Variaciones de las series temporales

Independientemente del valor que tiene el estudio de la evolución global de las series temporales, y la identificación de patrones de comportamiento que se hayan producido en el pasado, es igualmente importante centrarse en la evolución concreta de la misma mediante las variaciones

5 En ocasiones la predicción extra-muestral se utiliza para determinar la validez de la capacidad predictiva del modelo. En este caso, la muestra se suele dividir en dos partes, una de las cuales se utiliza para estimar los parámetros que determinan la tendencia y la componente estacional y a partir de ellos se realiza la predicción de los valores de la segunda muestra. Esta forma de valorar la capacidad predictiva de un modelo se considera más adecuada que realizar esta evaluación a partir de predicciones muestrales.

que en ella se produce en cada instante del tiempo, tanto en relación a su pasado inmediato como a otros instantes más alejados que se tomen como referencia. El estudio de esta evolución de las series temporales se puede llevar a cabo mediante el cálculo de las variaciones absolutas y de las variaciones relativas o tasas de variación

Variación absoluta

La variación absoluta de una serie temporal se obtiene a partir de la resta de dos valores de la misma en dos momentos del tiempo diferentes. Así,

$$\Delta Y_t(p) = Y_t - Y_{t-p} \qquad (5.10)$$

representa la diferencia entre el valor de la serie temporal en el momento *t* y en el momento *t-p*. Una de las variaciones más interesantes es la que se produce respecto al momento inmediatamente anterior, en este caso *p=1*, con lo que:

$$\Delta Y_t(1) = Y_t - Y_{t-1} \qquad (5.11)$$

que por simplificación denotaremos como ΔY_t. En muchas ocasiones, cuando la serie temporal presenta un comportamiento estacional, es interesante, además, conocer la variación que se produce entre el valor actual y el correspondiente al año anterior, constituyendo una *variación absoluta interanual*. Así, por ejemplo, en el caso de que la periodicidad de la serie temporal fuera mensual *p=12*, o si fuera trimestral *p=4*, se tendrán las siguientes variaciones absolutas interanuales:

$\Delta Y_t(12) = Y_t - Y_{t-12}$ → para periodicidad mensual

$\Delta Y_t(4) = Y_t - Y_{t-4}$ → para periodicidad trimestral

El signo de la variación absoluta nos indica si se ha producido un incremento (signo positivo) o una disminución (signo negativo). Sin embargo, la variación absoluta tiene el problema de no indicarnos directamente la relevancia de la variación en relación a la magnitud de comparación. Una misma variación puede considerarse *grande* en una serie temporal y *pequeña* en otra, debido a las diferentes valores de referencia y unidades de medidas en ambas series. Por esta razón, en muchas ocasiones se utiliza la variación relativa o tasa de variación que se describe a continuación.

Tasa de variación

Una tasa de variación se obtiene mediante la siguiente expresión:

$$Ta_t(p) = \frac{Y_t - Y_{t-p}}{Y_{t-p}} \times 100 \qquad (5.12)$$

En muchas ocasiones la tasa de variación se presenta en porcentaje, multiplicándola por 100, tal y como aparece en (5.12). En este caso indica el porcentaje en que se incrementa (cuando la tasa es positiva) o disminuye (cuando la tasa es negativa) la serie en el momento *t* respecto al momento de referencia *t-p*. Tal y como se puede observar en (5.12), la tasa de variación se obtiene a partir de la variación absoluta, de esta manera,

$$Ta_t(p) = \frac{\Delta Y_t(p)}{Y_{t-p}} \times 100 \tag{5.13}$$

Al igual que en la variación absoluta, es muy utilizada la tasa de variación respecto al periodo inmediatamente anterior,

$$Ta_t(1) = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100 \tag{5.13}$$

que para simplificar la denotaremos por Ta_t. También, es muy importante la tasa de variación que se produce entre momentos equidistantes un año, para el caso de series temporales con componente estacional. A esta tasa de variación se le denomina *tasa de variación interanual.* Así, por ejemplo,

$$Ta_t(12) = \frac{Y_t - Y_{t-12}}{Y_{t-12}} \times 100$$ → para periodicidad mensual

$$Ta_t(4) = \frac{Y_t - Y_{t-4}}{Y_{t-4}} \times 100$$ → para periodicidad trimestral

Módulo 5

En la tabla 8 se presentan las variaciones absolutas y las tasas de la reposición de toallas de piscina cuyos datos se presentaron en la tabla 3. ΔY_t representa la variación absoluta para $p=1$, $\Delta Y_t(3)$ sería la variación absoluta interanual, Ta_t es la tasa de variación para $p=1$ y, finalmente, $Ta_t(3)$ es la tasa de variación interanual.

Tabla 8. Variaciones de la reposición de toallas de piscina

Y_t	ΔY_t	$\Delta Y_t(3)$	Ta_t	$Ta_t(3)$
57				
62	5		8.77%	
76	14		22.58%	
57	-19	0	-25.00%	0.00%
103	46	41	80.70%	66.13%
86	-17	10	-16.50%	13.16%
84	-2	27	-2.33%	47.37%
110	26	7	30.95%	6.80%
104	-6	18	-5.45%	20.93%

Módulo 5

ACTIVIDADES

ACTIVIDAD 1

Recopila una serie temporal de periodicidad trimestral que abarque un mínimo de 6 años. Inicialmente sólo trabajarás con los primeros 5 años, de tal manera que los datos del último año (el más reciente) los mantendrás aparte. Con esta serie realiza las siguientes actividades:

a) Representa gráficamente la serie temporal y comenta los resultados. ¿Existe evidencia de la existencia de tendencia y de componente estacional?

b) Calcula la tendencia de la serie temporal. Comenta los resultados.

c) Calcula la componente estacional de la serie temporal. Comenta los resultados.

d) Predice los valores de la serie temporal para el último año que has mantenido aparte. Calcula los errores que has cometido en tus predicciones comparando las predicciones con los valores reales de la serie temporal. ¿Qué conclusiones sacas?

e) Calcula las tasas de variaciones interanuales de la serie temporal (utiliza en este caso los 6 años completos).

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: teoría y ejercicios resueltos*. Málaga: Agora.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

Determina cuáles de las siguientes afirmaciones son verdaderas. Ten en cuenta que en cada una de las preguntas puede haber más de una afirmación verdadera.

1. Series temporales:
   a. A diferencia de lo que ocurre con datos de corte transversal, las series temporales tienen un orden predeterminado.
   b. Para un mismo periodo de tiempo, datos con periodicidad mensual producen un tamaño muestral menor que para el caso de datos con periodicidad trimestral.
   c. Para una serie temporal con datos cuatrimestrales, el número de observaciones en 10 años es de 40.
   d. La predicción de series temporales no es un objetivo científicamente serio.

2. Representación gráfica y componentes de una serie temporal:
   a. El tiempo debe ser representado en el eje X en un gráfico de una serie temporal.
   b. No se puede realizar la representación gráfica de cada una de los componentes de una serie temporal, ya que éstos son desconocidos.
   c. Cada uno de las componentes de una serie temporal debe tener a lo largo del tiempo el mismo peso relativo.
   d. El componente estacional sólo se puede presentar en una serie temporal cuando la periodicidad de ésta es anual.

3. Componentes de una serie temporal:
   a. Los cuatro componentes de una serie temporal son: tendencia, variaciones estacionales, variaciones cíclicas y variaciones aleatorias.
   b. Las variaciones cíclicas son variaciones que se producen con un período inferior al año.
   c. Las variaciones aleatorias son constantes a lo largo del tiempo.
   d. El incremento del consumo de helados en verano es una variación estacional.

4. Componentes de una serie temporal:
   a. La componente aleatoria no tiene ningún comportamiento regular definido.
   b. La componente estacional recoge las oscilaciones que se producen en períodos superiores al año.
   c. Una serie temporal puede contar con más de una tendencia, por ejemplo una positiva y otra negativa.
   d. Para un estudio de la componente cíclica es necesario el estudio de periodos largos de tiempo.

5. La tendencia:
   a. El inconveniente del método analítico para el cálculo de la tendencia es que se pierden datos.
   b. En el método analítico del cálculo de la tendencia se debe crear una variable que recoja la evolución del tiempo.
   c. La tendencia queda recogida mediante una recta en el método analítico.
   d. Cuando a una serie se le quita la tendencia, calculada mediante el método analítico, la suma de la serie sin tendencia es igual a cero.

6. La componente estacional:
   a. La suma de las componentes estacionales es igual a cero.
   b. La componente estacional de un determinado periodo puede cambiar de signo alternativamente de un año a otro.
   c. El número de componentes estacionales de una serie temporal de periodicidad anual es de 12.
   d. El número de componentes estacionales de una serie temporal de periodicidad bi-anual es de 24.

7. Los ingresos (en miles de euros) de un determinado hotel en los últimos años han sido los siguientes:

Año	Ingresos
2000	250
2001	255
2002	258
2003	262
2004	268
2005	270

   a. Una regresión lineal sería una solución para poder predecir los ingresos en 2006.
   b. Los ingresos predichos para 2006 son de 280000 euros.
   c. Cada año se espera un incremento de ingresos en torno a 4080 euros.
   d. Incluso sin necesidad de calcularlo, podemos decir que el coeficiente de correlación entre el paso del tiempo y los ingresos es positivo.
   e. Ninguna de las anteriores.

EJERCICIO 2

Los siguientes datos corresponden a una serie temporal de periodicidad trimestral. A partir de los mismos responde a las siguientes cuestiones:

Y_i	26	21	25	19	30	24	29	23	33	28	35	28

a. Obtén su componente tendencial mediante el método analítico.

b. Obtén su componente estacional.

c. Obtén su componente aleatorio.

d. Dibuja en un mismo gráfico la serie original y la tendencia.

e. Predice la serie para el próximo año.

f. Calcula la tasa de variación trimestral y la tasa de variación interanual.

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

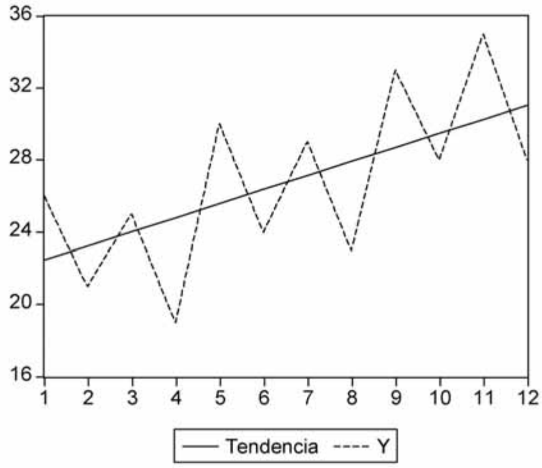
Afirmaciones verdaderas:

1. a
2. a, d
3. a, d
4. a, c, d
5. b, c, d
6. a
7. a, c, d,

EJERCICIO 2

Y_t	t	$(Y_t - \overline{Y})$	$(t - \overline{t})$	$(Y_t - \overline{Y})(t - \overline{t})$	$(t - \overline{t})^2$	T_t	$Y_t - T_t$	S_t	A_t
26	1	-0.75	-5.50	4.125	30.25	22.46	3.54	2.92	0.62
21	2	-5.75	-4.50	25.875	20.25	23.24	-2.24	-2.42	0.18
25	3	-1.75	-3.50	6.125	12.25	24.02	0.98	2.92	-1.94
19	4	-7.75	-2.50	19.375	6.25	24.80	-5.80	-3.42	-2.38
30	5	3.25	-1.50	-4.875	2.25	25.58	4.42	2.92	1.50
24	6	-2.75	-0.50	1.375	0.25	26.36	-2.36	-2.42	0.06
29	7	2.25	0.50	1.125	0.25	27.14	1.86	2.92	-1.06
23	8	-3.75	1.50	-5.625	2.25	27.92	-4.92	-3.42	-1.50
33	9	6.25	2.50	15.625	6.25	28.70	4.30	2.92	1.38
28	10	1.25	3.50	4.375	12.25	29.48	-1.48	-2.42	0.94
35	11	8.25	4.50	37.125	20.25	30.26	4.74	2.92	1.82
28	12	1.25	5.50	6.875	30.25	31.04	-3.04	-3.42	0.38
321	78			111.5	143				

t	S_t	Predicción
13.00	2.92	35.90
14.00	-2.42	30.57
15.00	2.92	35.90
16.00	-3.42	29.57



Tasas de variación

Ta_t	$Ta_t(4)$
--	--
-19.23%	--
19.05%	--
-24.00%	--
57.89%	15.38%
-20.00%	14.29%
20.83%	16.00%
-20.69%	21.05%
43.48%	10.00%
-15.15%	16.67%
25.00%	20.69%
-20.00%	21.74%

Módulo 5

GLOSARIO DE TÉRMINOS

Componente aleatorio: componente de una serie temporal que no tiene un comportamiento regular definido.

Componente cíclico: componente de una serie temporal de movimiento regular que se repite cada 2 o más años.

Componente estacional: componente de una serie temporal de movimiento regular que se repite cada año.

Tendencia: componente de una serie temporal que indica la evolución a largo plazo de la serie.

Tasa de variación: variación relativa (se suele dar en porcentajes) que se produce en una serie temporal entre dos momentos del tiempo.

Tasa de variación interanual: variación relativa (se suele dar en porcentajes) que se produce en una serie temporal entre momentos del tiempo distantes un año. De esta manera, sólo tiene sentido hablar de una tasa de variación interanual cuando los datos tienen una periodicidad inferior al año.

Módulo 6

Números índices

PRESENTACIÓN DEL MÓDULO

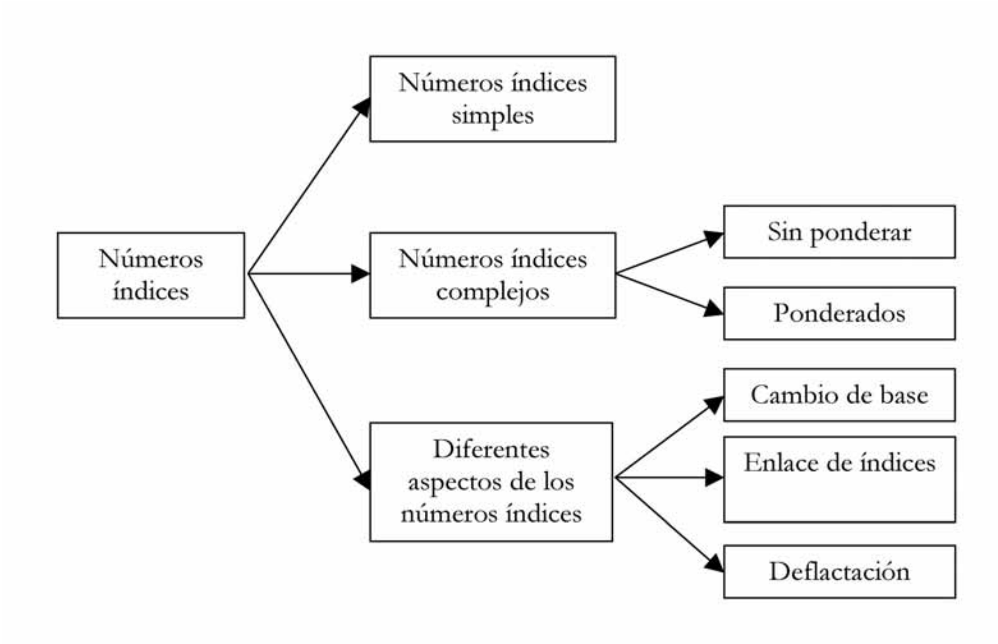
En este módulo se verán los números índices. Estos indicadores nos servirán para cuantificar la variación relativa de variables de interés, principalmente a lo largo del tiempo. En este sentido, estas medidas están relacionadas con las tasas de variación que se estudiaron en el módulo anterior, con la diferencia que en el caso de los números índices los valores de referencia de las variaciones relativas es fijo, mientras que para el caso de las tasas de variación, como ya se vio en su momento, eran variables. A lo largo del módulo se verá una de las utilidades más importantes de los números índices, la deflactación de series monetarias. Como es comúnmente conocido, el poder adquisitivo del dinero varía con el tiempo, de esta manera, cualquier serie temporal cuya unidad de medida sea monetaria, se verá influenciada por esta variabilidad. La construcción de números índices para los precios nos va a permitir eliminar de una serie temporal las variaciones producidas por el cambio del poder adquisitivo del dinero.

OBJETIVOS DEL MÓDULO

La superación de este módulo supone alcanzar los siguientes objetivos:

1. Saber distinguir entre los diferentes tipos de números índices y sus diferentes utilidades.
2. Ser capaz de construir e interpretar números índices simples y complejos.
3. Realizar cambios de base y enlaces de índices.
4. Deflactar series temporales.

ESQUEMA DE LOS CONTENIDOS



EXPOSICIÓN DE LOS CONTENIDOS

1. INTRODUCCIÓN. CLASIFICACIÓN DE LOS NÚMEROS ÍNDICES

Los números índices son un instrumento que permiten estudiar la evaluación temporal de variables estadísticas, normalmente de cantidades, precios y valores, respecto a un momento o periodo base. A lo largo de este módulo se verá que su obtención, así como algunas de sus características, está muy relacionada con la tasa de variación que ya se estudió en el módulo anterior, especialmente en lo relacionado a su carácter adimensional. A pesar de que la mayoría de las aplicaciones de los números índices se desarrollan en el contexto de las series temporales y, por tanto, éste será el supuesto que mantendremos a lo largo de este módulo, no se puede descartar su utilización en datos de corte transversal, fundamentalmente cuando el objetivo se centra en la comparación de determinadas variables cuyos valores se obtienen, por ejemplo, para diferentes áreas geográficas.

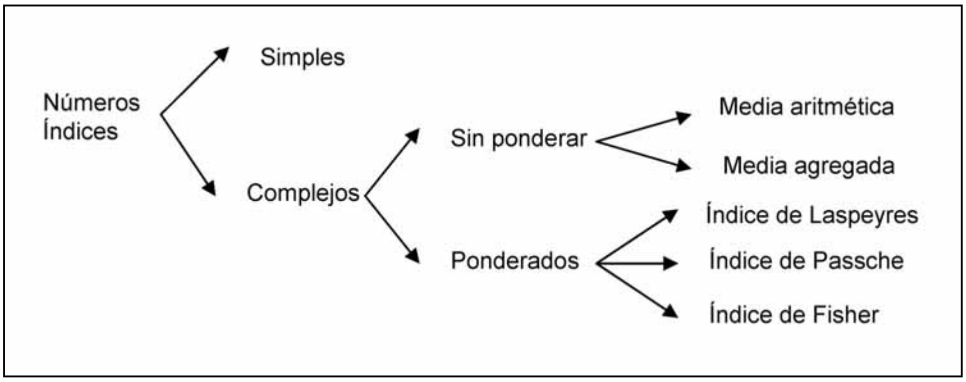
Una de las características más importantes de los números índices es su carácter adimensional. Ésta nos permitirá realizar comparaciones directas de la evolución de diferentes índices aún cuando las variables originales están expresadas en diferentes unidades de medida. Igualmente, esta propiedad nos permitirá construir índices complejos como combinación de índices simples. El carácter adimensional de los índices se deriva de que se obtienen como cociente de dos cantidades, de las cuales, la correspondiente al denominador actúa como valor de referencia o base. De esta manera, los números índices recogen la variación relativa entre dos momentos del tiempo. El periodo base que se elige como valor de referencia es arbitrario[1]. Mediante las oportunas operaciones una serie de números índices se puede transformar en otra con diferente periodo base.

1 En nuestro caso, y con el propósito de que los resultados de un mismo ejercicio no proporcionen diferentes resultados, vamos a considerar, salvo que el ejercicio en cuestión diga otra cosa, que el periodo base se corresponde con la primera observación temporal de la serie con la que trabajamos.

La utilización de los números índices en la economía es muy habitual. Entre ellos podemos destacar, por su conocimiento generalizado, el Índice de Precios al Consumo (IPC), que recoge la evolución del poder adquisitivo del dinero para los consumidores, y los índices bursátiles como el IBEX35, el Índice General de la Bolsa de Madrid o del Mercado Continuo, por citar algunos.

En la figura 1 se presenta una clasificación bastante habitual de los números índices. En primer lugar, los números índices pueden ser simples y complejos. Un número índice simple representa la evolución temporal de una única variable frente a uno complejo que representa la evolución de más de una variable. Así, por ejemplo, si se desea conocer la evolución de los precios de los *inputs* que utiliza una determinada actividad turística, se podría obtener un número índice simple de precios por cada uno de estos *inputs* utilizados. Sin embargo, si el número de *inputs* utilizados por la explotación es elevado, sería difícil tener una idea global de su evaluación, ya que algunos de los índices simples de precios podrían bajar, otros subir y algunos mantenerse más o menos constantes, por tanto, sería interesante poder contar con un único número índice de referencia que indicara, en conjunto, cuál ha sido la evolución de los precios. Para este fin, se utilizan los números índices complejos. A su vez, tal y como se refleja en la figura 1, estos índices se pueden dividir en: Sin ponderar y ponderados. En los índices complejos sin ponderar, los diferentes índices simples que intervienen en su confección tiene el mismo peso (ponderación). Por el contrario, en los índices complejos ponderados, los índices simples tienen diferentes pesos (ponderaciones). Siguiendo con nuestro ejemplo, en el deseo de conocer la evaluación de los precios de los *inputs* de una determinada actividad turística, en el caso de que éstos se consuman en valores parecidos, la utilización de un índice sin ponderación podría ser apropiado, sin embargo, si el valor del consumo de cada uno de los *inputs* difiere significativamente entre ellos, parece lógico que a la hora de calcular el índice de la evolución de los precios, aquellos *inputs* que se consumen en menor medida tengan un menor peso que aquellos que tienen un consumo más elevado. Entre los números índices complejos sin ponderar se encuentran el de la media aritmética y el de la media agregada. Para el caso de los números índices ponderados podemos destacar los índices de Laspeyres, Paasche y Fisher.

Figura 1. Clasificación de los números índices



2. Números índices simples

Los números índices simples se obtienen a partir de la evolución temporal de una única variable, los más utilizados son los de precios, cantidades y valores. Si p_t y q_t recogen la evolución de precios y de cantidades en el momento t, respectivamente, sus correspondientes números índices simples para ese mismo instante del tiempo, con periodo base en el momento 0, se obtienen de la siguiente manera:

$$I_0^t = \frac{p_t}{p_0} \times 100 \tag{6.1}$$

para el índice simple de precios, y,

$$I_0^t = \frac{q_t}{q_0} \times 100 \tag{6.2}$$

para el índice simple de cantidades. Es habitual presentar a los números índices en porcentajes. Por esta razón las expresiones de los índices han sido multiplicadas por 100.

La variable valor se obtiene como el producto del precio por la cantidad. De esta manera $v_t = p_t \cdot q_t$. En este caso el número índice simple del valor para el momento t con periodo base en el momento 0 será:

$$I_0^t = \frac{p_t\, q_t}{p_0\, q_0} \times 100 = \frac{v_t}{v_0} \times 100 \tag{6.3}$$

En la tabla 1 se presenta la evolución anual del precio por kilogramo de un *input* que se utiliza en la elaboración de algunos de los menús que elabora un restaurante, enclavado en un determinado complejo turístico. Además, a esta estadística se le añade las cantidades (en kilogramos) que se ha comprado de este *input* a lo largo de esos años.

Tabla 1. Evolución del precio y cantidades de un determinado *input*

Año	2001	2002	2003	2004
Precio	25	26	24	27
Cantidad	587	621	605	650

El precio viene expresado en euros y las cantidades en kilos.

Para poder obtener los números índices hay que elegir el periodo base. Tal y como se ha comentado, la determinación de este periodo es arbitrario. En nuestro caso, para seguir un mismo criterio, y siempre que no se diga lo contrario, vamos a elegir el primer periodo como base. Así, para el caso que nos ocupa, el periodo base será el año 2001. Teniendo en cuenta las ecuaciones (6.1) y (6.2), los número índices simples de precios y cantidades se obtendrán tal y como se presentan en la tabla 2.

Tabla 2. Índices simples de precios y cantidades

Año	Índice simple de precios	Índice simple de cantidades
2001	$I_{2001}^{2001}=\frac{p_{2001}}{p_{2001}}\times 100=\frac{25}{25}\times 100=100$	$I_{2001}^{2001}=\frac{q_{2001}}{q_{2001}}\times 100=\frac{587}{587}\times 100=100$
2002	$I_{2001}^{2002}=\frac{p_{2002}}{p_{2001}}\times 100=\frac{26}{25}\times 100=104$	$I_{2001}^{2002}=\frac{q_{2002}}{q_{2001}}\times 100=\frac{621}{587}\times 100=105.79$
2003	$I_{2001}^{2003}=\frac{p_{2003}}{p_{2001}}\times 100=\frac{24}{25}\times 100=96$	$I_{2001}^{2003}=\frac{q_{2003}}{q_{2001}}\times 100=\frac{605}{587}\times 100=103.07$
2004	$I_{2001}^{2004}=\frac{p_{2004}}{p_{2001}}\times 100=\frac{27}{25}\times 100=108$	$I_{2001}^{2004}=\frac{q_{2004}}{q_{2001}}\times 100=\frac{650}{587}\times 100=110.73$

Observando el índice de precios calculado en la tabla 2, se pueden destacar las siguientes consideraciones: (1) El número índice del periodo base toma siempre el valor 100. (2) Respecto al año base se ha producido un incremento de los precios en los años 2002 y 2004. Esto hecho se pone de manifiesto al comprobar que los índices de estos dos años son superiores al valor 100. En términos porcentuales los incrementos son del 4% para el año 2002 y del 8% para el año 2004. Este resultado se obtiene restando a cada uno de los índices el valor 100, o sea, (104-100) y (108-100), respectivamente. (3) Los precios sufren una disminución en el años 2003 respecto al año base. En este caso, el descenso se detecta al comprobar que el valor de su correspondiente índice es inferior a 100, concretamente 96. En términos porcentuales la disminución es del 4%, que se obtiene de la misma manera que se hizo en el caso de los incrementos de precios: (96-100). En relación a la cantidad de kilogramos del *input* adquirido por el restaurante, se observa que en todos los años se produce un incremento en relación al periodo base. De estos incrementos el menor se produce en el año 2003 con un 3.07% y el mayor en el año 2004 con un 10.73%.

El conocimiento de la evolución temporal del precio y la cantidad de un determinado *input* es de especial interés para cualquier explotación. En un determinado año puede producirse una disminución de los precios, pero si ésta viene acompaña de un incremento en el número de kilogramos comprados, el efecto no se verá reflejado en la adquisición total. Lo mismo puede ocurrir si se observa aisladamente la evolución del índice de las cantidades. En este caso, variaciones en las cantidades compradas pueden compensarse o agravarse en el caso de cambio en los precios.

En la tabla 3 se presenta la obtención del **índice de valor** para los datos de la tabla 1, éste se obtiene a partir la multiplicación entre precios y cantidades.

Tabla 3. Índice simple de valor

Año	Índice simple de valores
2001	$I_{2001}^{2001} = \frac{v_{2001}}{v_{2001}}100 = \frac{25 \times 587}{25 \times 587}100 = 100$
2002	$I_{2001}^{2002} = \frac{v_{2002}}{v_{2001}}100 = \frac{26 \times 621}{25 \times 587}100 = 110.02$
2003	$I_{2001}^{2003} = \frac{v_{2003}}{v_{2001}}100 = \frac{24 \times 605}{25 \times 587}100 = 98.94$
2004	$I_{2001}^{2004} = \frac{v_{2004}}{v_{2001}}100 = \frac{27 \times 650}{25 \times 587}100 = 119.59$

De la tabla 3 se pueden destacar dos aspectos que nos ayudarán a entender mejor su significado: (1) La reducción de precios que se puso de manifiesto en la tabla 2 para el año 2003, que se cifró en un 4%, realmente tiene un efecto sobre el total de la cantidad adquirida del 1.06% debido al incremento que se ha producido en las compras en ese año. (2) Tal y como se pudo observar, en el año 2004 se produjo el mayor incremento en precios y cantidades, concretamente del 8% y 10.73%, respectivamente. Sin embargo, el efecto que estos incrementos suponen en el total de las compras es mucho mayor de lo que pueden representar estos valores aisladamente. Concretamente, el incremento del valor de las compras es del 19.59%, prácticamente el doble de lo que indican los respectivos índices de precios y cantidades.

Módulo 6

3. Números índices complejos

En la mayoría de las ocasiones interesa estudiar la evolución conjunta de un fenómeno que viene constituido por un grupo de variables. En este caso, la utilización de números índices simples se puede volver engorrosa debido a que se tendrían tantos índices como variables se estuvieran estudiando, lo que dificultaría tener una descripción acertada de todo el fenómeno en su conjunto. Los números índices complejos intentan resolver este problema. Así, sería interesante contar con un único número índice en cada instante del tiempo en el cual intervengan todas las variables objeto de estudio.

Por tanto, un número índice complejo se puede definir como un índice en cuya confección intervienen más de una variable, las cuales forman parte de un mismo fenómeno objeto de estudio. Siguiendo con el ejemplo que se inició en la sección anterior, seguramente los administradores del restaurante no estarán solamente preocupados por la evaluación de precio, cantidad comprada o consumida y valor de un único *input*, sino que les interesará conocer la evolución de todos los *inputs* que intervienen en la producción de los menús. Sin embargo, si para cada uno de ellos se obtuvieran los índices simples, de precios, cantidades y de valor, posiblemente no se podrían manejar, y difícilmente de ellos se podría obtener una información clara de conjunto, salvo que a partir de los mismos se procediera a crear un único índice complejo.

Uno de los aspectos que se debe tener en cuenta a la hora de confeccionar un número índice complejo es la ponderación que se le debe dar a cada una de las variables que intervienen en el mismo. Volviendo a nuestro ejemplo, parece evidente que si queremos obtener un índice de precios complejo para los *inputs* que adquiere el restaurante, no deben tener el mismo peso los cambios en el precio de *inputs* que intervienen de forma marginal en los menús, que para el caso de un *input* que es muy demandado. Por esta razón, los número índices complejos de dividen en *sin ponderar* y *ponderados*. En los índices complejos sin ponderar todas las variables tienen el mismo peso. Por esta razón, estos índices deben utilizarse en el caso de que el peso entre ellas sea similar. Por el contrario, en los números índices complejos ponderados las variables que intervienen tienen un peso distinto. Por tanto, su utilización está justificada cuando existan diferencias claras entre las mismas. Un ejemplo de este tipo de índices es el IPC. Este índice intenta recoger la evolución de los precios a lo largo del tiempo. Se confecciona como un índice complejo ponderado. Parece evidente que las repercusiones que se producen como consecuencia de la variación de los precios no es igual en productos de consumo masivo entre la población que en aquellos donde su consumo es marginal. La repercusión que puede tener la variación del precio del pan o la carne de vaca, aún cuando esta variación sea pequeña, no es la misma que si varía el precio de un producto cuyo consumo sea marginal. De esta manera, parece lógico que cada uno de los productos que componen la *cesta de la compra* deben de estar ponderados en función de su consumo.

Números índices complejos sin ponderar

Tal y como se refleja en la figura 1, se tienen como índices complejos sin ponderar el de la media aritmética y el de la media agregada. El de la media aritmética se obtiene a partir de la media de los índices simples que intervienen en el índice complejo. De esta manera,

Módulo 6

$$\mathrm{Ima}_t^0 = \frac{\sum_{i=1}^{M} I_{it}^0}{M} \tag{6.4}$$

siendo M el número de índices simples o variables que intervienen en el índice complejo.

Por otro lado, el índice complejo de media agregada se obtiene sumando previamente los valores de las variables que intervienen en cada uno de los índices simples para a partir de esta agregación obtener el índice, concretamente,

$$\mathrm{Imag}_t^0 = \frac{\sum_{i=1}^{M} p_{it}}{\sum_{i=1}^{M} p_{i0}} \times 100 \tag{6.5}$$

para el caso de precios, en el caso de cantidades o valor simplemente se cambiaría la variable precio por la correspondiente.

En la tabla 1 se presentaron la evaluación del precio y las cantidades consumidas de un determinado *input*, utilizado en la elaboración de los menús de un determinado restaurante. Ahora en la tabla 4 se presentan, además del precio de este *input* (p2), el de otros dos *inputs* (p1 y p3), utilizados en la misma producción de menús. Igualmente, en esta misma tabla se encuentran los índices simples de precios de cada uno de los *inputs*, y en la última columna el valor del índice complejo de media aritmética, obtenidos a partir de la expresión (6.4).

Tabla 4. Cálculo del índice complejo sin ponderar de la media aritmética para el caso de tres *inputs*

Año	Precios			Índices simples de precios			Ima_t^0
	p_1	p_2	p_3	$I_0^t(1)$	$I_0^t(2)$	$I_0^t(3)$	
2001	14	25	5.50	100.00	100.00	100.00	100.00
2002	15	26	5.75	107.14	104.00	104.55	105.23
2003	14	24	5.25	100.00	96.00	95.45	97.15
2004	16	27	6.00	114.29	108.00	109.09	110.46

Para los mismos datos de la tabla 4, en la tabla 5 se presenta la obtención del índice complejo sin ponderar de la media agregada. Tal y como se puede observar, el resultado de ambos procedimientos, aunque son muy parecidos, no coinciden.

Tabla 5. Cálculo del índice complejo sin ponderar de la media agregada para el caso de tres *inputs*

Año	Precios			$p_1 + p_2 + p_3$	$Imag_t^0$
	p_1	p_2	p_3		
2001	14	25	5.50	44.50	100.00
2002	15	29	5.75	49.75	111.80
2003	14	31	5.25	50.25	112.92
2004	16	30	6.00	52.00	116.85

En los ejercicios de autocontrol se presenta para este mismo supuesto el cálculo de la media aritmética y media agregada para las cantidades y valor de los *inputs*.

Números índices complejos ponderados

Tal y como ha comentado anteriormente, cuando la importancia de los números índices simples o variables que intervienen en un índice complejo es desigual, es conveniente utilizar índices complejos ponderados, en lugar de los índices *sin ponderar* que acabamos de estudiar en el epígrafe anterior. De esta manera, en el ejemplo de los índices de precios, cuyos datos y cálculos se presentaron en las tablas 4 y 5, si cada uno de los tres *inputs* que intervienen en el mismo tienen una importancia muy distinta a la hora de intervenir en el proceso de transformación, por ejemplo, que uno sea realmente fundamental, mientras que los otros dos sólo tengan una intervención residual, es lógico pensar, que a la hora de construir el índice de precios, los *inputs* con mayor importancia en el proceso de transformación tengan una ponderación en el índice superior que en el caso de los *inputs* cuya intervención sea más reducida.

Antes de iniciar el estudio de los índices complejos ponderados es importante recordar la definición de media ponderada. La media ponderada, en el momento t, de una variable X_{it}, para M observaciones es:

$$mp(X_{it}) = \frac{\sum_{i=1}^{M} X_{it} W_{it}}{\sum_{i=1}^{M} W_{it}} = \sum_{i=1}^{M} X_{it} w_{it} \tag{6.6}$$

donde W_{it} representa la ponderación para el elemento i en el momento t. De (6.6) se desprende que:

$$w_{it} = \frac{W_{it}}{\sum_{i=1}^{M} W_{it}} \text{ con lo que } \sum_{i=1}^{M} w_i = 1 \tag{6.7}$$

Normalmente la ponderación que se utiliza en la elaboración de los números índices es el valor (precio por cantidades). De esta manera, para el caso de un número índice de precios, cada uno de los índices simples de los precios, que interviene en el índice complejo, pondera por el valor correspondiente a cada *input*. Supongamos, por ejemplo, que una determinada explotación utiliza tres *inputs* en la elaboración de sus productos, cuyos consumos han sido, en el momento t: 1200 euros para el *input* 1, 1500 euros para el *input* 2 y 300 euros para el *input* 3. En este caso, la ponderación para cada *input* será: $W_{1t} = 1200$, $W_{2t} = 1500$, y $W_{3t} = 300$, respectivamente, o lo que es lo mismo, teniendo en cuenta la expresión (6.7): $w_{1t} = 0.40$; $w_{2t} = 0.50$ y $w_{3t} = 0.10$.

La característica que define a cada uno de los índices complejos ponderados que ha continuación vamos a estudiar, es el momento de referencia que se utiliza a la hora de ponderar cada uno de los índices simples que intervienen en el índice complejo. Tal y como veremos, algunos índices ponderan considerando los precios y las cantidades en el periodo base. Otros, por el contrario, utilizan combinaciones de precios y cantidades en el momento base y en el momento t.

Índice de Laspeyres

El índice de Laspeyres, tanto para el índice de precios como para el de cantidades, utiliza la ponderación $w_{it} = \frac{p_{i0}\, q_{i0}}{\sum_{i=1}^{M} p_{i0}\, q_{i0}}$ en el momento t.

Teniendo en cuenta la representación de la media ponderada de la ecuación (6.6), se tiene que el índice de Laspeyres de precios toma la siguiente forma:

$$L_{p_t} = \sum_{i=1}^{M} I_{it} w_{it} = \frac{\sum_{i=1}^{M} \frac{p_{it}}{p_{i0}} p_{i0}\, q_{i0}}{\sum_{i=1}^{M} p_{i0}\, q_{i0}} = \frac{\sum_{i=1}^{M} p_{it}\, q_{i0}}{\sum_{i=1}^{M} p_{i0}\, q_{i0}} \tag{6.8}$$

mientras que el de cantidad es:

$$L_{q_t} = \sum_{i=1}^{M} I_{it} w_{it} = \frac{\sum_{i=1}^{M} \frac{q_{it}}{q_{i0}} p_{i0}\, q_{i0}}{\sum_{i=1}^{M} p_{i0}\, q_{i0}} = \frac{\sum_{i=1}^{M} p_{i0}\, q_{it}}{\sum_{i=1}^{M} p_{i0}\, q_{i0}} \tag{6.9}$$

En la tabla 6 se encuentran los datos del precio y las cantidades de los tres *inputs* que se dedican para la preparación de los menús del ejemplo que estamos siguiendo a lo largo del módulo.

Tabla 6. Evolución del precio y cantidades de tres *inputs*

Año	p_{1t}	q_{1t}	p_{2t}	q_{2t}	p_{3t}	q_{3t}
2001	14	26	25	587	5.50	128
2002	15	30	26	621	5.75	135
2003	14	29	24	605	5.25	136
2004	16	33	27	650	6.00	143

Tabla 7. Cálculo del índice de precios y cantidades de Laspeyres

Índice de precios de Laspeyres					
Año	$p_{1t}\, q_{10}$	$p_{2t}\, q_{20}$	$p_{3t}\, q_{30}$	$\sum_{i=1}^{3} p_{it}\, q_{i0}$	L_{p_t}
2001	364	14675	704	15743	100.00
2002	390	15262	736	16388	104.10
2003	364	14088	672	15124	96.07
2004	416	15849	768	17033	108.19
Índice de cantidades de Laspeyres					
Año	$p_{10}\, q_{1t}$	$p_{20}\, q_{2t}$	$p_{30}\, q_{3t}$	$\sum_{i=1}^{3} p_{i0}\, q_{it}$	L_{q_t}
2001	364	14675	704.0	15743.0	100.00
2002	420	15525	742.5	16687.5	106.00
2003	406	15125	748.0	16279.0	103.40
2004	462	16250	786.5	17498.5	111.15

En la tabla 7 se prestan todos los cálculos necesarios para la obtención de los índices de precios y cantidades de Laspeyres.

Índice de Paasche

En este caso la ponderación para el índice de precios queda definida tal que $w_{it} = \dfrac{p_{i0}\, q_{it}}{\sum_{i=1}^{M} p_{i0}\, q_{it}}$, con lo que el índice de precio de Paasche es igual a:

$$P_{p_t} = \sum_{i=1}^{M} I_{it} w_{it} = \frac{\sum_{i=1}^{M} \frac{p_{it}}{p_{i0}} p_{i0}\, q_{it}}{\sum_{i=1}^{M} p_{i0}\, q_{it}} = \frac{\sum_{i=1}^{M} p_{it}\, q_{it}}{\sum_{i=1}^{M} p_{i0}\, q_{it}} \tag{6.10}$$

mientras que la ponderación para el caso del índice de cantidades es igual a $w_{it} = \dfrac{p_{it}\, q_{i0}}{\sum_{i=1}^{M} p_{it}\, q_{i0}}$, con lo que el índice se obtiene como:

$$P_{q_t} = \sum_{i=1}^{M} I_{it} w_{it} = \frac{\sum_{i=1}^{M} \frac{q_{it}}{q_{i0}} p_{it}\, q_{i0}}{\sum_{i=1}^{M} p_{it}\, q_{i0}} = \frac{\sum_{i=1}^{M} p_{it}\, q_{it}}{\sum_{i=1}^{M} p_{it}\, q_{i0}} \tag{6.11}$$

Tabla 8. Cálculo del índice de precios y cantidades de Paasche

Índice de precios de Paasche								
$p_{1t}q_{1t}$	$p_{2t}q_{2t}$	$p_{3t}q_{3t}$	$p_{10}\, q_{1t}$	$p_{20}\, q_{2t}$	$p_{30}\, q_{3t}$	$\sum_{i=1}^{3} p_{it}\, q_{it}$	$\sum_{i=1}^{3} p_{i0}\, q_{it}$	P_p
364	14675	704.00	364	14675	704.0	15743.00	15743.0	100.00
450	16146	776.25	420	15525	742.5	17372.25	16687.5	104.10
406	14520	714.00	406	15125	748.0	15640.00	16279.0	96.07
528	17550	858.00	462	16250	786.5	18936.00	17498.5	108.21
Índice de cantidades de Paasche								
$p_{1t}q_{1t}$	$p_{2t}q_{2t}$	$p_{3t}q_{3t}$	$p_{1t}\, q_{10}$	$p_{2t}\, q_{20}$	$p_{3t}\, q_{30}$	$\sum_{i=1}^{3} p_{it}\, q_{it}$	$\sum_{i=1}^{3} p_{it}\, q_{i0}$	P_q
364	14675	704.00	364	14675	704	15743.00	15743	100.00
450	16146	776.25	390	15262	736	17372.25	16388	106.01
406	14520	714.00	364	14088	672	15640.00	15124	103.41
528	17550	858.00	416	15849	768	18936.00	17033	111.17

En la tabla 8 se encuentran los cálculos necesarios para la obtención del índice de precios y cantidades de Paasche.

Índice de Fisher

Se obtiene a partir del índice de Laspeyres y del índice de Paasche. En el caso del índice de precios, el índice de Fisher es igual a:

$$F_{p_t} = \sqrt{L_{p_t} \times P_{p_t}} \tag{6.12}$$

mientras que para el índice de cantidades:

$$F_{q_t} = \sqrt{L_{q_t} \times P_{q_t}} \tag{6.13}$$

Tabla 9. Cálculo del índice de precios y cantidades de Fisher

Año	L_{p_t}	P_{p_t}	Indice de precios de Fisher $\left(F_{p_t}\right)$	L_{q_t}	P_{q_t}	Indice de cantidades de Fisher $\left(F_{q_t}\right)$
2001	100.00	100.00	100.00	100.00	100.00	100.00
2002	104.10	104.10	104.10	106.00	106.01	106.00
2003	96.07	96.07	96.07	103.40	103.41	103.41
2004	108.19	108.21	108.20	111.15	111.17	111.16

En la tabla 9 se encuentran los cálculos necesarios para la obtención del índice de precios y cantidades del índice de Fisher.

4. Cambio de base y enlace de índices

Cambio de base

Para una determinada serie de números índices es posible cambiar el periodo base del momento 0 por otro en el momento 0'. Para llevar a cabo este cometido se puede utilizar la siguiente expresión:

$$I_{0'}^{t} = \frac{I_0^t}{I_0^{0'}} \times 100 \tag{6.14}$$

En la tabla 10 tenemos una misma serie de números índices trimestrales para dos periodos bases distintos, primer trimestre de 1998 y 2000, respectivamente.

Tabla 10. Cambio de base

Año/Trimes.	Índice (base 1998:01)	Índice (base 2000:01)
1998:01	100.00	91.05
1998:02	100.85	91.82
1998:03	102.94	93.72
1998:04	102.75	93.55
1999:01	105.37	95.94
1999:02	106.16	96.65
1999:03	106.93	97.36
1999:04	109.01	99.25
2000:01	109.83	100.00
2000:02	110.63	100.72
2000:03	113.12	102.99
2000:04	113.87	103.68

Teniendo en cuenta la expresión (6.14), podemos destacar, a título de ejemplo, algunos de los cálculos realizados para obtener los valores del índice para el periodo base del primer trimestre de 2000. Concretamente para el caso del segundo trimestre de 1998, tercer trimestre de 1999 y cuarto trimestre de 2000.

Módulo 6

Segundo trimestre 1998: $I_{00:01}^{98:02} = \frac{I_{98:01}^{98:02}}{I_{98:01}^{00:01}} \times 100 = \frac{100.85}{109.83} \times 100 = 91.82$

Tercer trimestre 1999: $I_{00:01}^{99:03} = \frac{I_{98:01}^{99:03}}{I_{98:01}^{00:01}} \times 100 = \frac{106.93}{109.83} \times 100 = 97.36$

Cuarto trimestre 2000: $I_{00:01}^{00:04} = \frac{I_{98:01}^{00:04}}{I_{98:01}^{00:01}} \times 100 = \frac{113.87}{109.83} \times 100 = 103.68$

Enlace de índices

Esta operación se realiza cuando para un mismo índice se tiene una ruptura como consecuencia de un cambio en el periodo base. Así, por ejemplo, para un índice complejo, el periodo base está unido al periodo en que se utilizan las ponderaciones. En el momento que se cambia el periodo base de las ponderaciones, se produce una ruptura en el índice. Sin embargo, en la práctica es necesario disponer de series continuas ya que muchos de estos índices suelen ser utilizados como referencia en pactos privados y públicos (p. ej. subidas salariales, incremento de alquileres, etc). O, en otras ocasiones, el estudio temporal de su evaluación implica la necesidad de trabajar con series continuas. En las dos primeras columnas de la tabla 11 se encuentra un caso de ruptura de un índice con base en el periodo 0 y en el periodo t'. El objetivo del enlace de índices es crear una única serie para el índice que vaya desde el momento 0 al momento T. Esto se puede realizar de dos maneras, rellenando las casillas vacías de la primera columna de la tabla 11 o las de la segunda columna de esta misma tabla. Lo más habitual es la segunda opción, o sea, acercar lo más posible al momento actual el periodo base. De esta manera, los valores de la segunda columna se obtendrán tal que:

$$I_{t'}^{t} = \frac{I_t^0}{I_0^{t'}} \times 100 \tag{6.15}$$

En la última columna de la tabla 11 se encuentran todos los valores del índice con base en el periodo t'.

Tabla 11. Índice con diferentes periodo base

Con base en el momento cero	Con base en el momento t'	Con base en el momento t' (enlazado)
$I_0^0 = 100$		$I_{t'}^0 = \frac{I_0^0}{I_0^{t'}} \times 100$
I_0^1		$I_{t'}^1 = \frac{I_1^0}{I_0^{t'}} \times 100$
I_0^2		$I_{t'}^2 = \frac{I_2^0}{I_0^{t'}} \times 100$
$\vdots$	$\vdots$	$\vdots$
$I_0^{t'}$	$I_{t'}^{t'} = 100$	$I_{t'}^{t'} = 100$
	$I_{t'}^{t'+1}$	$I_{t'}^{t'+1}$
	$I_{t'}^{t'+2}$	$I_{t'}^{t'+2}$
	$\vdots$	$\vdots$
	$I_{t'}^{T}$	$I_{t'}^{T}$

Volviendo con el ejemplo que representan los datos de la tabla 10, supongamos que el índice en cuestión es complejo y que se produjo un cambio de ponderaciones en el primer trimestre del año 2000, con lo que este periodo será la nueva base. Esta situación es la que aparece en la segunda y tercera columna de la tabla 12. Tal y como se puede apreciar en estas columnas, tanto para un periodo base (1998:01) como para el otro (2000:01) las series no son continuas. En la cuarta columna de esta misma tabla se presentan los cálculos necesarios, obtenidos a partir de la expresión (6.15), para enlazar los índices.

Tabla 12. Enlace de índices

Año/Trimes.	Índice (base 1998:01)	Índice (base 2000:01)	Índice (base 2000:01)
1998:01	100.00		$I_{2000:01}^{1998:01}=\frac{I_{1998:01}^{1998:01}}{I_{1998:01}^{2000:01}}\times 100=\frac{100}{109.83}\times 100=91.05$
1998:02	100.85		$I_{2000:01}^{1998:02}=\frac{I_{1998:01}^{1998:02}}{I_{1998:01}^{2000:01}}\times 100=\frac{100.85}{109.83}\times 100=91.82$
1998:03	102.94		$I_{2000:01}^{1998:03}=\frac{I_{1998:01}^{1998:03}}{I_{1998:01}^{2000:01}}\times 100=\frac{102.94}{109.83}\times 100=93.72$
1998:04	102.75		$I_{2000:01}^{1998:04}=\frac{I_{1998:01}^{1998:04}}{I_{1998:01}^{2000:01}}\times 100=\frac{102.75}{109.83}\times 100=93.55$
1999:01	105.37		$I_{2000:01}^{1999:01}=\frac{I_{1998:01}^{1999:01}}{I_{1998:01}^{2000:01}}\times 100=\frac{105.37}{109.83}\times 100=95.94$
1999:02	106.16		$I_{2000:01}^{1999:02}=\frac{I_{1998:01}^{1999:02}}{I_{1998:01}^{2000:01}}\times 100=\frac{106.16}{109.83}\times 100=96.65$
1999:03	106.93		$I_{2000:01}^{1999:03}=\frac{I_{1998:01}^{1999:03}}{I_{1998:01}^{2000:01}}\times 100=\frac{106.93}{109.83}\times 100=97.36$
1999:04	109.01		$I_{2000:01}^{1999:04}=\frac{I_{1998:01}^{1999:04}}{I_{1998:01}^{2000:01}}\times 100=\frac{109.01}{109.83}\times 100=99.25$
2000:01	109.83	100.00	100.00
2000:02		100.72	100.72
2000:03		102.99	102.99
2000:04		103.68	103.68

5. Deflactación de series temporales

Es muy habitual encontrarnos con series temporales que vienen referenciadas en unidades de medidas monetarias. Un inconveniente de la utilización de esta unidad de medida es el cambio que a lo largo del tiempo se produce del poder adquisitivo de la moneda[2]. Por esta razón, cuando se pretende realizar comparaciones temporales de estas series con su pasado, podemos llegar a conclusiones erróneas si no tenemos en cuenta esta circunstancia. Así, por ejemplo, la comparación del presente y del pasado del salario medio en España nos llevaría a que han existido incrementos mayores de los que realmente se han producido si no consideramos el poder adquisitivo de nuestra moneda a los largo del tiempo. Todos somos concientes de que el poder adquisitivo de 600 euros hace 20 años era mucho mayor que en la actualidad.

2 Cuando el poder adquisitivo de una moneda baja hablamos de inflación mientras que cuando sube hablamos de deflación.

Los números índices, y más concretamente, los números índices de precios son una herramienta importantísima a la hora de corregir los efectos de la variación del valor monetario. Una serie que no ha sido corregida como consecuencia de la variación del valor monetario se dice que está en términos constantes o nominales, mientras que si se muestra corregida, hablamos de que la serie se expresa en términos reales. Al procedimiento de corrección se le denomina deflactación de la serie y consiste en realizar la siguiente operación:

$$VR_t = \frac{VN_t}{I_o^t} \times 100 \tag{6.16}$$

siendo VR_t y VN_t el valor real y nominal en el momento t, respectivamente.

En la tabla 13 tenemos los ingresos nominales y reales mensuales por turismo en España para el año 2002 y el correspondiente IPC.

Tabla 13. Ingresos por turismo (Mill. de euros) e IPC en España para el año 2002

Año:Mes	Ingresos nominales	IPC (base:2001)	Ingresos reales
Enero	2093.78	101.262	2067.686
Febrero	2185.23	101.350	2156.122
Marzo	2358.31	102.188	2307.815
Abril	2435.00	103.575	2350.953
Mayo	3190.00	103.948	3068.842
Junio	3197.00	103.953	3075.428
Julio	4356.00	103.231	4219.663
Agosto	4043.57	103.527	3905.812
Septiembre	3139.78	103.914	3021.518
Octubre	3541.25	104.943	3374.451
Noviembre	2993.96	105.106	2848.515
Diciembre	2009.56	105.455	1905.609

Fuente: Banco de España

Módulo 6

Teniendo en cuenta (6.16) los ingresos reales se han obtenido de la siguiente manera:

$$VR_{02:01} = \frac{VN_{02:01}}{I_0^{02:01}} \times 100 = \frac{2093.78}{101.262} \times 100 = 2067.686$$

$$\vdots$$

$$VR_{02:12} = \frac{VN_{02:12}}{I_0^{02:12}} \times 100 = \frac{2009.56}{105.455} \times 100 = 1905.609$$

En este caso se ha utilizado el IPC para la deflactación. Si se considera necesario se puede utilizar cualquier otro índice que mida la variación del valor del dinero. Un índice sectorial también podría haber sido posible.

El efecto que produce la deflactación se puede apreciar comparando las figuras 2 y 3, donde se presentan los ingresos por turismo en España desde enero de 1990 a noviembre de 2003. Aún cuando ambas series se incrementan a lo largo del tiempo, el crecimiento de la serie real es más atenuado que el correspondiente a la serie nominal. Además, debido a que el periodo base del IPC se encuentra hacia el final del periodo, el efecto de la deflactación se nota más al principio de las series. Incluso, en periodos de alta inflación la tendencia de la serie puede cambiar. De manera tal, que se tenga una tendencia creciente para la serie nominal y una serie decreciente para la serie real.

Figura 2. Valores nominales

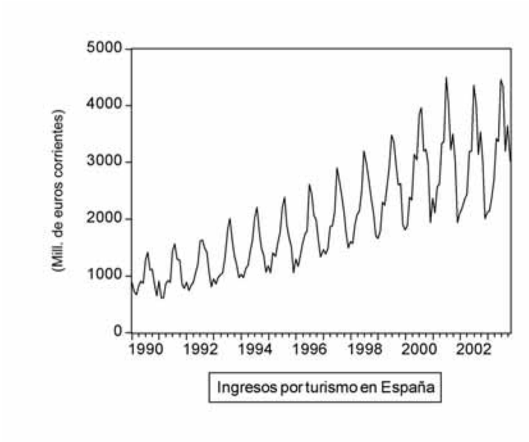
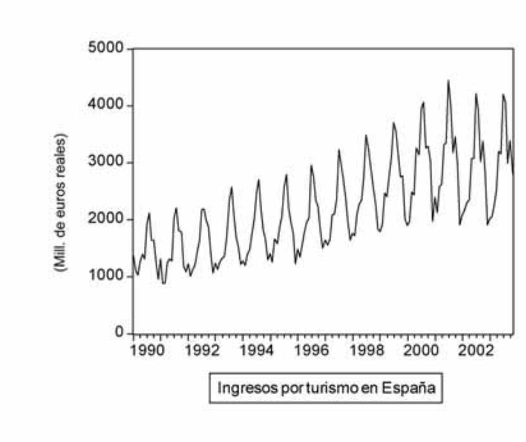


Figura 3. Valores reales



ACTIVIDADES

ACTIVIDAD 1

Recopila varias series temporales de precios y cantidades y construye los siguientes índices:

a) Índices simples para cada una de las series (precios, cantidades y valor).

b) Índices complejos sin ponderar de precios y cantidades, utilizando la media aritmética y la media agregada.

c) Índice de precio y cantidades de Laspeyres.

d) Índice de precio y cantidades de Paasche.

e) Índice de precios y cantidades de Fisher.

ACTIVIDAD 2

Recopila un índice de precios y cualquier serie temporal en unidades monetarias nominales que esté afectada por el índice de precios recopilado. Con esta información realiza las siguientes actividades:

a) Realiza un cambio de base en el índice de precios.

b) Deflacta la serie temporal y confronta gráficamente la serie temporal nominal y la real.

BIBLIOGRAFÍA

Bibliografía básica

Alegre Martín, J., Cladera Munar, M. y Juaneda Sampol, C. N. (2003). *Análisis cuantitativo de la actividad turística*. Madrid: Pirámide.

Martín Pliego, F. J. (1995). *Introducción a la estadística económica y empresarial*. Madrid: AC.

Bibliografía complementaria

Fernández Aguado, C. (1999). *Manual de estadística descriptiva aplicada al sector turístico*. Madrid: Síntesis.

Morales Fernández, A. y Lacomba Arias, B. (2000). *Estadística básica aplicada al sector turístico: teoría y ejercicios resueltos*. Málaga: Agora.

Ronquillo Melcio, A. (1997). *Estadística aplicada al sector turístico: técnicas cuantitativas y cualitativas de análisis turístico*. Madrid: Ramón Areces.

EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

Determina cuáles de las siguientes afirmaciones son verdaderas. Ten en cuenta que en cada una de las preguntas pueden haber más de una afirmación verdadera.

1. Números índices:
   a. Un número índice es una medida estadística que sirve para comparar una misma variable en dos situaciones.
   b. Un número índice es una medida estadística que se usa para ordenar en índices las magnitudes económicas de un país.
   c. El índice de precios al consumo no es propiamente un número índice.
   d. La ponderación de un índice de precios al consumo depende de la composición de la *cesta de la compra* de los consumidores.

2. Números índices:
   a. Los números índices podemos dividirlos en simples y complejos.
   b. Un número índice complejo siempre es ponderado mientras que un número índice simple siempre es sin ponderar.
   c. Las ponderaciones de un número índice no son necesariamente constantes y pueden variar con el tiempo.
   d. En un número índice ponderado las ponderaciones positivas deben estar compensadas con las ponderaciones negativas.

3. Números índices:
   a. El número índice obtenido como media agragada es un índice complejo sin ponderar.
   b. Un índice simple de valor 150 siempre se interpreta como que la variable se ha incrementado en un 50% respecto al período anterior.
   c. Si tienes una empresa que produce el mismo valor de tres productos diferentes, y se quiere obtener un índice de precios complejo de los mismos a través del tiempo, se podría utilizar un número índice complejo sin ponderar.
   d. El número índice complejo sin ponderar calculado a través de la media simple aritmética, coincide con el número índice complejo sin ponderar calculado a través de la media agregada simple.

4. Números índices:
   a. El número índice de Laspeyres de precios y cantidad pondera por el valor en el periodo cero.
   b. El número índice de Laspeyres de cantidades pondera por las cantidades en el periodo cero.
   c. El número índice de precios de Paasche pondera por el valor obtenido como el producto del precio en el periodo cero y las cantidades en el periodo t.

d. El número índice de Fisher de precios se obtiene a partir de los valores correspondientes de los índices de Laspeyres y Paasche.

5. Cambio de base:

a. Una serie temporal de números índices de precios puede cambiar de periodo base siempre y cuando conozcamos la serie temporal de los precios que ha sido utilizados en su elaboración.

b. Para un determinado momento t, se puede obtener un número índice con cambio de base a partir de la siguiente expresión: $I^t_{0'} = \frac{I^t_0}{I^{0'}_0} \times 10$, siendo 0 el periodo base original y 0' el nuevo periodo base.

c. Si se desea realizar un cambio de periodo base para un determinado número índice temporal, el nuevo periodo base nunca puede ser anterior en el tiempo al que se pretende sustituir.

d. Cuando se produce un cambio de periodo base en un serie temporal de números índices, el nuevo periodo base tomará el valor 100, sólo en el caso de que se trate de un número índice simple o complejo sin ponderar. En el caso de un número índice complejo ponderado el nuevo periodo base no necesariamente tomará el valor 100.

6. Enlace de índices:

a. El enlace de índices se realiza para transformar índices entre distintos países o zonas geográficas.

b. La necesidad de disponer de índices continuos y sin rupturas provoca la utilización de enlace de índices.

c. Un enlace de índices se produce cuando se tiene un mismo índice para periodos de tiempo distintos y con periodo base distinto.

d. La necesidad de realizar un enlace de índices suele ser más habitual en índices complejos que en índices simples.

7. La siguiente tabla muestra una relación de salarios anuales nominales y del índice de precios al consumo del mismo año:

Año	Salario	IPC
1989	21000	100
1990	23000	107
1991	24000	114
1992	25500	120
1993	26000	122
1994	26000	125
1995	26460	126

a. El incremento nominal de salarios entre 1989 y 1995 ha sido del 26%.
b. El poder adquisitivo de los salarios no se ha visto incrementado en el período que va desde 1989 a 1995.
c. Entre 1989 y 1995 hay dos años donde el poder adquisitivo de los salarios disminuye respecto al año anterior.
d. El mayor incremento real de salarios entre años consecutivos se produce entre 1989 y 1990.
e. La mayor disminución de salarios reales se produce entre 1993 y 1994.

EJERCICIO 2

A partir de los datos de la tabla 6 que contienen las **cantidades** que consume un determinado restaurante, calcula los números índices de las cantidades y los números índices complejos sin ponderar de la media aritmética y de la media agregada.

EJERCICIO 3

A continuación tenemos los beneficios anuales de una determinada empresa desde 1997 a 2003.

Año	Beneficio	Indice (base:1997)	Indice (base: 2000)
1997	2548	100	
1998	4874	111	
1999	5254	119	
2000	6465	125	100
2001	7924		107
2002	8625		115
2003	9587		121

Se pretende poner estos beneficios en términos reales, para lo cual se dispone de un único índice de precios pero obtenido para dos años base distintos, tal y como refleja la tabla anterior. Con la información disponible, calcular los beneficios en términos reales desde 1997 a 2003, tomando como año base 2002.

SOLUCIONES A LOS EJERCICIOS DE AUTOCONTROL

EJERCICIO 1

Afirmaciones verdaderas:

1. a, d
2. a, c
3. a, c
4. a, c, d
5. b
6. b, c, d
7. a, d, e

Año	Salario	IPC	Salario Real	Variación Salario Real
1989	21000	100	21000.00	--
1990	23000	107	21495.33	495.33
1991	24000	114	21052.63	-442.70
1992	25500	120	21250.00	197.37
1993	26000	122	21311.48	61.48
1994	26000	125	20800.00	-511.48
1995	26460	126	21000.00	200.00

EJERCICIO 2

Índices simples y complejo sin ponderar de la media aritmética.

Año	Cantidades			Índices simples			Ima_t^0
	q_1	q_2	q_3	$I_0^t(1)$	$I_0^t(2)$	$I_0^t(3)$	
2001	26	587	128	100.00	100.00	100.00	100.00
2002	30	621	135	115.38	105.79	105.47	108.88
2003	29	605	136	111.54	103.07	106.25	106.95
2004	33	650	143	126.92	110.73	111.72	116.46

Índice complejo sin ponderar de la media agregada.

Año	Cantidades q_1	q_2	q_3	$q_1 + q_2 + q_3$	Imag_t^0
2001	26	587	128	741	100.00
2002	30	621	135	786	106.07
2003	29	605	136	770	103.91
2004	33	650	143	826	111.47

SOLUCIÓN EJERCICIO 3

Año	Beneficio	Indice(1997)	Indice(2000)	Indice(2000)	Indice(2002)	**Beneficio real**
1997	2548	100		80	69.57	3662.75
1998	4874	111		88.8	77.22	6312.05
1999	5254	119		95.2	82.78	6346.74
2000	6465	125	100	100	86.96	7434.75
2001	7924		107	107	93.04	8516.45
2002	8625		115	115	100.00	8625.00
2003	9587		121	121	105.22	9111.61

GLOSARIO DE TÉRMINOS

Cambio de base: procedimiento mediante el cual se cambia el periodo base de un número índice.

Números índices: instrumento que permiten estudiar la evaluación, especialmente temporal, de variables en relación a un periodo base o de referencia.

Números índices complejos: número índice donde intervienen más de una variable.

Números índices simples: número índice donde interviene una única variable.