INTRODUCCIÓN A LA ESTADÍSTICA PARA TURISMO

Alberto Muñoz Cabanes
Alfonso Herrero de Egaña y Espinosa de los Monteros
Azahara Muñoz Martínez

EDICIONES ACADÉMICAS

Alberto Muñoz Cabanes
Alfonso Herrero de Egaña y Espinosa de los Monteros
Azahara Muñoz Martínez

INTRODUCCIÓN A LA ESTADÍSTICA PARA TURISMO

UNED
EDICIONES ACADÉMICAS

ISBN: 978-84-92477-46-3
Depósito legal: M-3731-2011

Preimpresión: MonoComp, S. A.
Impresión: Edigrafos, S. A.

Impreso en España - *Printed in Spain*

Índice

Prólogo

El presente libro constituye el texto base de la asignatura de Introducción a la Estadística que se cursa en el primer curso del grado de Turismo de la Universidad Nacional de Educación a Distancia, cuyo objetivo es que el alumno aprenda los conceptos estadísticos básicos que le facilitarán la toma de decisiones en el campo de la empresa turística; asimismo, los contenidos de esta asignatura sirven como apoyo para el aprendizaje de otras materias que se encuentran en el programa del grado tales como el marketing o la economía.

Para poder seguir los contenidos desarrollados tan sólo es necesario conocer las principales operaciones matemáticas estudiadas en el Bachillerato. En todo caso, los autores han optado en todo momento por la simplicidad en la exposición de los conceptos a la vez que se ha intentado mantener el rigor matemático, incluyendo asimismo ejemplos que permiten una rápida comprensión de las ideas presentadas.

El libro se estructura en cinco capítulos que abarcan todo el temario de la asignatura. En el primero de ellos se introduce al alumno en la ciencia estadística, destacando la importancia que tiene su conocimiento para el sector turístico. Seguidamente se realiza un breve repaso a los orígenes históricos de la Estadística, finalizando con la exposición de los principales conceptos básicos utilizados en este campo e introduciendo someramente las diferentes etapas en las que se divide una investigación estadística.

En el segundo capítulo se realiza una revisión de las principales fuentes estadísticas del sector turístico español, incluyendo aquellas que estudian aspectos de la oferta y la demanda turística, así como indicadores coyunturales, estadísticas de síntesis y datos de empleo.

El estudio de las distribuciones de frecuencias unidimensionales se inicia en el tercer capítulo, mostrando la forma de construir e interpretar la información contenida en tablas de distribuciones de frecuencias, así como diferentes formas de representación gráfica de la información contenida en las tablas. En este capítulo también se presenta la notación que se utilizará en el resto de capítulos, por lo que conviene que el alumno se familiarice con ella antes de continuar con el estudio.

El capítulo cuarto presenta al alumno las principales medidas de posición, dispersión, forma y concentración, mediante las que es posible describir y sintetizar la información de una distribución de frecuencias unidimensional.

Finalmente en el quinto capítulo se da el salto al mundo bidimensional, pasando a estudiarse el comportamiento simultáneo de dos variables, con objeto de establecer relaciones de dependencia estadística entre ellas y estimar relaciones lineales para realizar predicciones.

Al término de cada capítulo se presenta una serie de ejercicios de autoevaluación de tipo test, con el propósito de que el alumno pueda afianzar el aprendizaje de los conocimientos recién adquiridos y cuyas soluciones se pueden encontrar al final del libro. También se ha incluido una selección de lecturas recomendadas mediante las cuales el alumno puede ampliar información sobre los contenidos tratados en cada capítulo, así como una lista de palabras clave cuya comprensión es fundamental para asegurar una correcta comprensión del tema.

CAPÍTULO 1

Introducción. Conceptos básicos

ESQUEMA

OBJETIVOS

Al finalizar el estudio de este capítulo, el alumno deberá ser capaz de:

1. Definir qué es la Estadística, señalar sus principales ramas y comprender la importancia que presenta para el sector turístico.
2. Conocer los orígenes históricos de la Estadística.
3. Utilizar adecuadamente los principales conceptos básicos utilizados en ciencia estadística.
4. Explicar las diferentes etapas en las que se divide una investigación estadística.

1.1. INTRODUCCIÓN

El término *estadística* proviene del latín *statisticum collegium* («consejo de Estado») y de su derivado italiano *statista* («hombre de Estado» o «político»). Asimismo el término alemán *statistik*, introducido por Achenwall en 1749, designaba originalmente el análisis de datos del Estado, es decir, la «ciencia del Estado», también denominada «aritmética política».

Como podemos deducir de su etimología, la estadística ha estado históricamente asociada en sus orígenes a la utilización de los datos por parte del gobierno y de los cuerpos administrativos. Los estados recababan datos, especialmente sobre renta y población, a efectos de recaudación de impuestos y mantenimiento del ejército. Esos datos se identificaban con el estado, razón por la cual terminaron conociéndose como estadísticas. En este sentido, la Estadística es tan antigua casi como el propio ser humano.

Sin embargo, ésta es una forma muy estrecha de entender y definir la Estadística. Actualmente podemos definir la ***estadística*** como aquella ciencia con base matemática que principalmente se ocupa de la recolección, análisis e interpretación de datos con objeto de detectar comportamientos regulares en fenómenos de tipo aleatorio y hacer más efectiva la toma de decisiones.

Debido a su versatilidad, la estadística es en la práctica una ciencia transversal a una amplia variedad de disciplinas, las cuales recurren a ella para resolver multitud de problemas. Así, la física, la mayoría de las ciencias sociales, las ciencias vinculadas a la salud, las áreas como el control de calidad y los negocios, y algunas instituciones gubernamentales suelen utilizar la ciencia estadística para comprender algunos de los fenómenos que constituyen su objeto de estudio.

La Estadística se divide a su vez en dos grandes ramas de estudio que son:

- La **Estadística Descriptiva**, la cual se encarga de la recolección, clasificación y descripción de datos muestrales o poblacionales, para su interpretación y análisis, que es de la que nos ocuparemos en este curso.
- La **Estadística Inferencial** o Inferencia Estadística, que se ocupa de la generación de los modelos y leyes a partir de datos procedentes de un determinado subconjunto de individuos o muestra, que pueden extrapolarse con un cierto nivel de fiabilidad a la totalidad de un colectivo o población a fin de realizar predicciones sobre la misma.

Estas dos ramas no son independientes; por el contrario, son complementarias y entre ambas se logra obtener la información suficiente para prever un posible escenario futuro, a fin de que quien tenga poder de decisión tome las medidas necesarias para transformar ese futuro o para mantener las condiciones existentes.

La Estadística constituye una herramienta esencial en muchos campos, resultando de especial interés para el gestor que trabaja en el sector del turismo, el cual, al igual que en cualquier otro sector de la economía, debe contar con toda la información necesaria para tomar decisiones de una forma efectiva. Gracias a la utilización de la ciencia estadística, el gestor de una empresa turística puede, entre otras cosas:

- Obtener información de la realidad que directa o indirectamente afecta a su trabajo, proporcionándole métodos y técnicas para la recogida de datos y su codificación o sistematización.
- Ordenar y reducir dicha información para hacer más fácil su comprensión mediante tablas y gráficos.
- Buscar repeticiones en determinados fenómenos con objeto de predecir su comportamiento en el futuro.

Nuestro objetivo a lo largo de este libro será el de proporcionar a los alumnos los conocimientos básicos necesarios para abordar este tipo de tareas.

1.2. HISTORIA

Con objeto de comprender mejor el estado actual de la Estadística, resulta de especial interés indagar en sus raíces históricas para obtener una visión de su naturaleza y de sus objetivos como disciplina científica.

Desde los comienzos de la civilización han existido formas sencillas de estadística, utilizándose representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para realizar un recuento del número de personas y animales, si bien es en el antiguo Egipto donde podemos encontrar los primeros registros estadísticos formales. De acuerdo con el historiador griego Heródoto, los faraones lograron recopilar hacia el año 3050 a.C. abundantes datos relacionados con la población y la riqueza del país con objeto de preparar la construcción de las pirámides. En el caso de las civilizaciones mesopotámicas se utilizaban tablillas de arcilla para recopilar datos tabulados sobre la producción agrícola y los géneros vendidos o intercambiados mediante trueque.

En Israel también encontramos referencias a trabajos estadísticos: en los libros bíblicos de Números y Crónicas aparecen dos censos de población así como datos sobre el bienestar material de las diversas tribus judías. Otras civilizaciones como la griega o la china también realizaban censos de población periódicamente con fines tributarios, sociales y militares.

Sin embargo, fueron sin duda los romanos quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas.

Por el contrario, en la Edad Media el número de operaciones estadísticas descendió notablemente, destacando tan sólo el *Capitulare de Villis*, un registro de los dominios y bienes privados realizado por Carlomagno en Francia, y el *Domesday Book* o Libro del Gran Catastro realizado por Guillermo el Conquistador, un documento que recogía la propiedad, extensión y valor de las tierras que constituyó el primer compendio estadístico de Inglaterra. Por su parte a finales del siglo XV Alonso de Quintanilla elabora por encargo de los Reyes Católicos el primer censo en España.

Entre los siglos XVI y XVII se produce una completa revolución en los métodos estadísticos, al calor de los nuevos descubrimientos científicos y el desarrollo del comercio. En particular, existía un especial interés por la Estadística Demográfica a fin de conocer si la población aumentaba, decrecía o permanecía estática. Por ello a comienzos del siglo XVI comienzan a registrarse los nacimientos, matrimonios y defunciones en Francia e Inglaterra.

En el siglo XVII se producen avances sustanciales, comenzándose a impartir en las universidades alemanas enseñanzas de Aritmética Política, asignatura que trataba sobre la descripción numérica de hechos de interés para la Administración Pública. Algunos autores destacados de la época son los ingleses Petty y Graunt. El primero puede ser considerado como el primero en proponer la creación de un servicio estadístico nacional, mientras que el segundo fue capaz de estimar tasas de mortalidad para la población londinense. Por otro lado, el interés mostrado por diversos matemá-

ticos como Pascal o Fermat por determinar las reglas que controlaban los juegos de azar, por entonces muy habituales en Europa, sentaba las bases de la teoría de la probabilidad.

El siglo XVIII supone el inicio del estudio de la estadística desde un punto de vista completamente matemático, con los trabajos de Bernoulli y de Moivre, la teoría de los errores de Cotes y Simpson y las reglas de combinatoria de Laplace. Asimismo en este siglo se elaboran los primeros censos en España, el de Ensenada en 1749 y el de Floridablanca en 1787, si bien los actuales censos de periodicidad decenal no comenzarán a elaborarse hasta 1860 por la Junta General de Estadística.

Posiblemente el siglo XIX sea la etapa en la que la Estadística recibe su mayor impulso gracias a los importantes avances realizados. Entre ellos destacan la teoría de los errores de observación de Laplace y Gauss, y la teoría de los mínimos cuadrados desarrollada por Gauss, Legendre y Adrain. A finales de este siglo Gaston presenta el método de Correlación, cuyo objeto era medir la influencia relativa de los factores sobre las variables. El trabajo de Gaston fue la base del coeficiente de correlación de Pearson y de algunos estudios sobre la medida de las relaciones realizados por Norton, Hooker y Yule dentro del campo de la Biometría. En el siglo XIX también comienza a aplicarse la teoría de la probabilidad en el campo de las ciencias sociales. En particular, Quételet introduce la noción del «hombre promedio» (*l'homme moyen*) como un medio para entender los fenómenos sociales complejos tales como la criminalidad, el número de matrimonios o los suicidios.

Finalmente en el siglo XX, la Estadística deja de ser considerada un área dentro de las Matemáticas para pasar a ser una ciencia con entidad propia que permite desarrollar herramientas para resolver problemas de diversa índole, ya estén relacionados con la salud pública (epidemiología, bioestadística, etc.) o con asuntos económicos y sociales (econometría, psicometría, etc.). En la actualidad el uso de la Estadística se ha extendido más allá de sus orígenes al servicio del Estado, siendo utilizada por las empresas para comprender mejor los datos que recibe y tomar decisiones en campos tan dispares como las ciencias naturales, las ciencias sociales, la medicina o las finanzas.

1.3. CONCEPTOS BÁSICOS

Pasamos a continuación a definir algunos conceptos básicos que son utilizados habitualmente en las investigaciones estadísticas.

1.3.1. Individuo, población y muestra

Llamamos ***población***, universo o colectivo al conjunto de elementos, individuos o entes sobre el cual van a recaer las observaciones o la realización del estudio. Por ejemplo, todos los visitantes que recibe anualmente el museo del Prado.

La población puede ser, según su tamaño, de dos tipos:

- Finita: aquella cuyos elementos pueden ser numerados o descritos completamente, como, por ejemplo, el censo electoral de una Comunidad Autónoma.
- Infinita: aquella en la que no es posible determinar el número concreto de elementos que la compone. Tal sería el caso del número de billetes vendidos por las agencias de viajes a nivel mundial

Las poblaciones están compuestas de unidades estadísticas denominadas ***individuos***. El individuo es un ente observable que no tiene por qué ser una persona, también puede ser un objeto, un ser vivo, o incluso algo abstracto.

Por su parte diremos que una ***muestra*** es un subconjunto de elementos que forman parte de la población y que se considera representativo de la misma, es decir, que los resultados extraídos a partir de ella pueden hacerse extensivos al resto de la población con cierto grado de fiabilidad. Un ejemplo de muestra sería la selección de 150 visitantes que hayan visitado este año el museo del Prado.

1.3.2. Parámetro, variable y atributo

Se denomina ***parámetro*** a un valor representativo de la población que el investigador desea estudiar. Tal sería el caso de la media de edad de los visitantes a un parque temático, la proporción de extranjeros que pernoctan en España por motivos turísticos o la desviación típica de la ocupación hotelera en Denia.

Por su parte, definimos ***variable*** como aquella característica poblacional susceptible de tomar valores numéricos, es decir, que admite unidades de medida. Ejemplos de variables serían el gasto de una empresa turística, el salario que se paga a los empleados de una agencia de viajes, la edad de los visitantes de un parque de atracciones, etc.

Podemos diferenciar dos tipos de variables:

- Variables Discretas: son aquellas que toman valores aislados (números naturales) y que no pueden tomar ningún valor intermedio entre dos consecutivos fijados. Por ejemplo, el número de estrellas de un hotel o el número de hijos de una familia.

- Variables Continuas: son aquellas que toman infinitos valores (números reales) en un intervalo dado, de forma que pueden tomar cualquier valor intermedio, al menos teóricamente, en su rango de variación. Por ejemplo, la distancia del hotel a la playa o el espesor de la nieve de una estación de esquí.

Tanto las variables discretas como las continuas pueden agruparse construyendo intervalos, entre cuyos valores extremos se ubicarán las diferentes observaciones registradas. Sin embargo, estrictamente hablando, sólo las variables continuas pueden ser objeto de categorización mediante intervalos.

Cuando las características de los individuos no son susceptibles de ser medidas numéricamente, entonces reciben el nombre de ***atributos***. Ejemplos de atributos serían la nacionalidad, el estado civil, el sexo, la profesión de un individuo, el tipo de transporte que elige en un viaje, etc.

Los atributos a su vez pueden clasificarse en:

- Nominales: cuando los datos se pueden agrupar en categorías, pero sin ninguna jerarquía entre sí. Por ejemplo, color de los ojos, profesión, marca de coche, etcétera.
- Ordinales: aquellos que poseen un orden, secuencia o progresión natural esperable. Por ejemplo, preguntas de encuesta sobre el grado de satisfacción de algún producto o servicio. Por ejemplo, mucho, poco, nada; bueno, regular, malo, etcétera.

En ocasiones, y con objeto de realizar algún tipo de análisis numérico con atributos, son transformados de manera ficticia en variables, asignando un número a cada una de sus modalidades. Por ejemplo, se puede convertir el atributo *sexo* en una variable asignando por ejemplo un 1 a la modalidad hombre y un 0 a la modalidad mujer.

1.4. LA INVESTIGACIÓN ESTADÍSTICA

De manera muy general podemos decir que las etapas de toda investigación estadística son las siguientes:

a) ***Definición de los objetivos perseguidos con la investigación***: se trata de la fase más importante, ya que en ella se definen los parámetros poblacionales que se pretenden investigar (por ejemplo, el gasto medio de los veraneantes en Gandía).

b) ***Recogida de datos***: para lo que existen básicamente dos formas:

- Mediante la ejecución de una *encuesta censal*, esto es, obteniendo información preguntando a todos los individuos que componen la población. No obstante, la realización de estudios censales es algo excepcional debido a su elevado coste y al largo periodo de ejecución.
- Por medio de la ejecución de una *encuesta muestral*, siendo ésta la alternativa generalmente utilizada en la investigación estadística al presentar enormes ventajas tales como un coste económico reducido, un corto periodo de ejecución en comparación con las encuestas censales y un mejor control de la calidad de los datos al tratarse de un volumen de información más reducido.

c) ***Descripción y estimación de parámetros poblacionales***: en el caso de la investigación censal, la investigación finaliza con la descripción de las características poblaciones a través de tablas y gráficos. Sin embargo, en el caso de la investigación muestral tan sólo tendremos estimaciones de los parámetros que deberemos validar estadísticamente. Es en último caso donde la inferencia estadística adquiere todo su significado, ya que a partir de los resultados obtenidos con los datos de la muestra, trataremos de generalizar las conclusiones obtenidas al total de la población.

1.5. EJERCICIOS DE AUTOEVALUACIÓN

1. ¿Cuáles de los siguientes fenómenos *no* son objeto del análisis estadístico?

   a) Viajes realizados a un destino turístico.
   b) Diagonales de un cuadrado.
   c) Lanzamiento de un dado.
   d) Número de meses que dura el verano.
   e) Extranjeros que visitan España cada año.
   f) Gasto en las vacaciones de Navidad de una familia.

2. Señale cuáles de las siguientes variables son discretas y cuáles continuas:

   a) Número de acciones vendidas cada día en la Bolsa.
   b) Temperaturas registradas cada hora en un observatorio.
   c) Período de duración de un automóvil.
   d) El diámetro de las tuercas producidas en una fábrica.
   e) Número de hijos de 50 familias.
   f) Censo anual de los españoles.

3. Indique si las siguientes características son variables o atributos:

   a) Preferencias políticas (Izquierda, Derecha o Centro).
   b) Marca de cerveza preferida.
   c) Velocidad (en Km/h).
   d) Peso (en kg).
   e) Estado civil.
   f) Nivel de estudios.
   g) Años de estudios completados.
   h) Tipo de enseñanza recibida.
   i) Número de empleados de una empresa.
   j) Temperatura de un paciente en grados Celsius.

4. Señale cuáles de los siguientes atributos son nominales y cuáles son ordinales:

   a) Nombres de personas.
   b) Días de la semana.
   c) Estado civil.
   d) Meses del año.
   e) Colores.
   f) Nivel de estudios.

5. ¿De qué población seleccionaría una muestra para estudiar el gasto de los españoles que visitan Punta Cana?

   a) Hoteles situados en Punta Cana.
   b) Restaurantes situados en República Dominicana.
   c) Españoles que han contratado un viaje a Punta Cana.
   d) Vuelos contratados a Punta Cana.

LECTURAS RECOMENDADAS

ALEGRE, J. *et al.* (2003). *Análisis Cuantitativo de la Actividad Turística*. Ed. Pirámide.
FERNÁNDEZ, C. (1993). *Manual de Estadística Descriptiva Aplicada al Sector Turístico*. Ed. Síntesis.
GONICK, L. y SMITH, W. (2002). *La Estadística en Cómic*. Ed. Zendrera.
GONZÁLEZ, J. M. (1994). *El Azar y la Historia*. Ed. Planeta.
RAO, C. R. (1994). *Estadística y Verdad (Aprovechando el Azar)*. Ed. PPU.
SÁNCHEZ, J. (1975). *Historia de la Estadística como Ciencia en España (1500-1900)*. Instituto Nacional de Estadística.
TANUR, J. M. *et al.* (1992). *La Estadística. Una Guía de lo Desconocido*. Alianza Editorial.

PALABRAS CLAVE

- Estadística
- Población
- Individuo
- Muestra
- Parámetro
- Variable
- Atributo

CAPÍTULO 2

Fuentes de información estadística de interés para el sector turístico

ESQUEMA

OBJETIVOS

Al finalizar el estudio de este capítulo, el alumno deberá ser capaz de:

1. Reflexionar sobre la importancia del turismo en la economía española.
2. Identificar y describir las principales fuentes estadísticas relacionadas con la oferta y la demanda del sector turístico español.
3. Conocer los principales indicadores coyunturales, las estadísticas de síntesis y las fuentes estadísticas de datos sobre el empleo en el sector del turismo en España.
4. Enumerar algunas de las estadísticas específicas del sector turístico elaboradas por las Comunidades Autónomas.

2.1. INTRODUCCIÓN

Realizar una presentación del conjunto de la información disponible para el análisis del turismo obliga a tener en cuenta algunos datos relativos a la importancia que la actividad turística representa en España. Se trata de un país cuya importancia turística reside, principalmente, en ser un país receptor de visitantes extranjeros.

España ocupa el segundo lugar en el ranking mundial tanto por ingresos turísticos como por número de turistas, siendo uno de los destinos preferidos por los europeos (cerca del 80% de nuestros visitantes), en particular por alemanes, británicos y franceses. El peso de la industria turística en el Producto Interior Bruto español supone más de un 10% y genera cerca de 2.000.000 de puestos de trabajo, lo que supone un 12% del empleo total generado en España.

Por lo que se refiere a la entrada de visitantes extranjeros, hay todo un conjunto de aspectos que merecen ser destacados:

- España recibió en 2008 un total de 99,1 millones de visitantes internacionales. El 59% de los mismos (57,3 millones) fueron turistas, es decir, pernoctaron en su destino al menos una noche, y el 41% restante fueron excursionistas.
- Los meses estivales (julio, agosto y septiembre) concentran la mayor parte de las llegadas del año (35%).
- Más del 60% de los turistas internacionales procedieron de tres mercados: Reino Unido, Alemania y Francia. Un 93,5% de los turistas tuvieron su origen en algún país europeo.
- En relación con las distintas vías de acceso, los resultados para 2009 recogen algunas pautas de comportamiento que merecen señalarse:
  - El 81,5% de las entradas de turistas se ha producido por aeropuertos, el 14,6% por carretera y el resto, por ferrocarril y puertos.
  - El 35% de los turistas vienen durante los meses de la temporada de verano (de junio a septiembre), confirmando la alta estacionalidad del turismo hacia España.

Además de este importante flujo de visitantes extranjeros, el sector turístico descansa también en buena medida en el hecho de que los españoles, en su gran mayoría, pasan sus vacaciones en España, algo que queda patente en los siguientes datos:

- De acuerdo con la estadística de Movimientos Turísticos de los Españoles (FAMILITUR), en 2008 se produjeron 168,8 millones de viajes. El 93% de estos viajes, es decir, 157,6 millones tuvieron como destino España.
- El número de pernoctaciones que realizaron los residentes en España en el año 2008 alcanzó los 780,9 millones de pernoctaciones.
- Con respecto a la estacionalidad, en el año 2008 los viajes de los residentes en España se concentraron esencialmente en la época estival (de junio a septiembre), así como en Semana Santa (marzo) y en el mes de mayo, debido al puente que se celebró en dicho mes. En el conjunto de estos meses se realizaron casi el 60% del total de viajes del año 2008.

- Los principales destinos internos de los residentes en España fueron los situados en la costa mediterránea de la península: Andalucía, Cataluña y Comunidad Valenciana. Le siguieron comunidades del interior como Castilla y León, Castilla-La Mancha y la Comunidad de Madrid.
- El medio de transporte utilizado en el 82,6% de los viajes de los residentes dentro de España fue el coche. Por su parte, el avión fue utilizado en el 5,2% de los desplazamientos.
- El motivo principal de los viajes de los residentes realizados dentro de España fue el ocio, recreo y vacaciones, presente en un 52% de los casos. Le siguen, por orden de importancia, la visita a familiares o amigos (23,6%) y los viajes de trabajo o negocios (16%).
- En el 33,6% de los viajes con destino España la vivienda de familiares o amigos fue el alojamiento elegido. En un 29,5% se utilizó la vivienda propia y en un 17,2% la opción preferida fueron los hoteles.
- La fidelidad a los destinos nacionales es muy elevada, ya que el 91,9% de los viajes se realizaron a lugares ya visitados.

Como se puede comprobar por este resumen sobre la importancia y comportamiento de los distintos colectivos de turistas, España dispone de abundantes estadísticas que recogen tanto aspectos relacionados con la demanda y actividad de los turistas (mediante encuestas y también incorporando registros administrativos de las unidades responsables del control de tráfico en distintas modalidades, tal y como veremos más adelante), como características del lado de la oferta tales como el número de establecimientos y empresas existentes en distintos subsectores turísticos.

El sistema global de estadísticas turísticas españolas está formado fundamentalmente por las estadísticas que producen el Instituto de Estudios Turísticos y el Instituto Nacional de Estadística, así como otros organismos de carácter nacional y regional.

A continuación pasamos a ver las principales estadísticas que elaboran estos organismos, las cuales conviene conocer, pues sin duda resultarán de utilidad a la hora de gestionar cualquier negocio relacionado con el sector turístico.

2.2. ESTADÍSTICAS SOBRE LA OFERTA DEL SECTOR TURÍSTICO

Las estadísticas relacionadas con la oferta del sector turístico español pueden dividirse en tres grupos: censos, estadísticas estructurales y estadísticas de ocupación de alojamientos turísticos. A continuación pasamos a ver en detalle cada uno de ellos.

2.2.1. Censos

Los censos que recogen información acerca de la oferta del sector turístico español son elaborados con carácter continuo por el Instituto de Turismo de España (Turespaña), organismo público creado en 1962 que, según lo establecido por el Real Decreto 561/2009, actualmente posee rango de Subdirección General, dependiendo directamente de la Presidencia de Turespaña, y cuyas funciones son la investigación de los factores que inciden sobre el turismo, así como la elaboración, recopilación y valoración de estadísticas, información y datos relativos al turismo. También es el responsable de la creación y difusión del conocimiento y la inteligencia turística y la coordinación de la información sobre el sector turístico generada por las distintas unidades administrativas dependientes de la Secretaría de Estado de Turismo y del propio Turespaña.

Para ello utiliza las bases de datos proporcionadas por las Comunidades Autónomas con un nivel de desagregación municipal. Los censos elaborados, cuyos resultados pueden consultarse en *http://www.iet.tourspain.es/*, son los siguientes:

a) Censo Continuo de Establecimientos Hoteleros

El Censo Continuo de Establecimientos Hoteleros recoge información sobre el nombre del hotel, ubicación, categoría, capacidad, precio, pertenencia o no a una cadena o holding de empresas y año de construcción y de la última remodelación. La difusión de los resultados la realiza anualmente Turespaña a través de la Guía Oficial de Hoteles y la Guía Profesional de Hoteles; asimismo el Instituto Nacional de Estadística (INE) incluye un resumen en el Anuario Estadístico de España.

b) Censo Continuo de Acampamentos Turísticos

El Censo Continuo de Acampamentos Turísticos recoge información acerca del nombre del camping, ubicación, categoría, capacidad, precio, y su pertenencia o no a una cadena o holding de empresas. La difusión de los resultados la realiza anualmente Turespaña a través de la Guía Oficial de Campings.

c) Censo Continuo de Apartamentos Turísticos Autorizados

El Censo Continuo de Apartamentos Turísticos Autorizados recoge información sobre el nombre, ubicación, categoría, precio, pertenencia o no a cadena o holding de

empresas, año de construcción o última remodelación. La encuesta se realiza de manera continua y su desagregación es municipal. La difusión de los resultados la realiza anualmente Turespaña a través de la publicación Hoteles, Camping, Apartamentos por Provincias.

d) Censo Continuo de Agencias de Viajes

El Censo Continuo de Agencias de Viajes recoge información relativa al nombre de la agencia, ubicación, puntos de venta, mayoristas, minoristas y mixtas, y pertenencia o no a una cadena o holding de empresas. La encuesta se realiza de manera continua y su desagregación es municipal. Si bien aún está por determinar su difusión externa, es posible acceder a la información a través del Centro de Documentación Turística de España de Turespaña.

Asimismo cabe destacar la información de carácter censal que publica el Instituto Nacional de Estadística relacionada con los Albergues y Ciudades de Vacaciones, en la que se recoge información relacionada con las pernoctaciones en albergues juveniles por país y año, albergues juveniles y plazas por Comunidades Autónomas y provincias y tipología, y el número de ciudades de vacaciones. El censo se elabora anualmente explotando la información procedente del Consorcio REAJ (Red Española de Albergues Juveniles), el Ministerio de Trabajo y Asuntos Sociales y la Secretaría de Estado de Comercio y Turismo. La difusión de los resultados se realiza a través del Anuario Estadístico de España y en Internet en *www.ine.es/inebase/*.

2.2.2. Encuestas estructurales

Las encuestas estructurales elaboradas por el INE permiten obtener una visión estructural de los aspectos más significativos (empleo, producción, inversión) que caracterizan a las empresas turísticas. Actualmente la encuesta estructural que recoge datos del turismo en España es la **Encuesta Anual de Servicios**[1], la cual se dirige a todas las empresas dedicadas al Comercio, Turismo, Transporte, Tecnologías de la Información, Actividades Inmobiliarias y Alquileres, Servicios Prestados a Empresas y Servicios Personales. La encuesta proporciona, por tanto, información anual sobre las características estructurales y económicas específicas de cada una de las actividades incluidas en el ámbito de estudio, tales como el tamaño de las empresas, datos contables (compras, gastos, operaciones de capital) o la estructura del empleo y la inversión.

La desagregación de esta encuesta es nacional, si bien para algunas variables alcanza el nivel autonómico. La difusión se realiza a través de Anuario Estadístico de España, pudiendo consultarse los resultados en Internet en *www.ine.es/inebase/*.

---

[1] En la década de los noventa se realizaron diferentes encuestas relacionadas directamente con el sector turístico español como la Encuesta sobre la Estructura de las Empresas Hoteleras, Encuesta sobre la Estructura de las Empresas de Restauración y la Encuesta sobre la Estructura de Empresas de Agencias de Viajes. Actualmente todas estas encuestas han sido englobadas dentro de la Encuesta Anual de Servicios.

2.2.3. Estadísticas de ocupación de alojamientos turísticos

Se trata de estadísticas elaboradas por el INE en colaboración con las Comunidades Autónomas cuyo objetivo es la obtención del número de viajeros que se alojan en los establecimientos turísticos, así como de las características de los mismos (número de empleados, número de establecimientos, precios pagados por los clientes). La difusión de los resultados se realiza al final de cada mes con referencia a los datos del mes anterior. Dentro de este apartado podemos distinguir las siguientes encuestas:

a) Encuesta de Ocupación Hotelera

La Encuesta de Ocupación Hotelera[2] proporciona información estadística referida a viajeros alojados en establecimientos hoteleros inscritos como tales en el correspondiente registro de las Consejerías de Turismo de cada Comunidad Autónoma y definidos como aquellos establecimientos que prestan servicios de alojamiento colectivo mediante precio con o sin otros servicios complementarios (hotel, apartahotel, motel, hostal, pensión, etc.).

Las principales variables analizadas son el número de pernoctaciones, grado de ocupación del establecimiento, estancia media del viajero, número de reservas, precio medio, y el personal ocupado. El nivel de desagregación de la encuesta es a nivel estatal, si bien para algunas variables alcanza el nivel provincial; asimismo en el caso de aquellos municipios en los que la concentración de la oferta turística es significativa, la información se detalla por puntos y por zonas de interes turístico. La información está disponible mensualmente a través de INEbase en *www.ine.es/inebase/*.

b) Encuesta de Ocupación en Acampamentos Turísticos

La Encuesta de Ocupación en Acampamentos Turísticos[3] proporciona información estadística referida a viajeros alojados en acampamentos turísticos (también denominados *campings*) inscritos como tales en el correspondiente registro de las Consejerías de Turismo de cada Comunidad Autónoma y definidos como aquellos espacios de terreno debidamente delimitados, dotados y acondicionados, destinados a facilitar a las personas, de modo habitual y mediante el pago de un precio estipulado, un lugar para hacer vida al aire libre durante tiempo limitado con fines vacacionales o turísticos y utilizando como residencia, albergues móviles, caravanas, tiendas de campaña u otros elementos similares fácilmente transportables.

Las principales variables analizadas son el volumen de pernoctaciones, viajeros según países de origen, estancia media, mercados emisores, tipología de estableci-

[2] La Encuesta de Ocupación Hotelera sustituye desde enero de 1999 a la antigua Encuesta de Movimiento de Viajeros en Establecimientos Hoteleros.

[3] La Encuesta de Ocupación en Acampamentos Turísticos sustituye desde enero de 1999 a la antigua Encuesta de Movimiento de Viajeros en Acampamentos.

miento, parcelas ocupadas y número de plazas. El nivel de desagregación de la encuesta es a nivel estatal, si bien para algunas variables alcanza el nivel provincial. La información está disponible mensualmente a través de INEbase en *www.ine.es/inebase/*.

c) Encuesta de Ocupación en Apartamentos Turísticos

La Encuesta de Ocupación en Apartamentos Turísticos proporciona información estadística. Proporciona información sobre la demanda y la oferta de los servicios de alojamiento que prestan los establecimientos de apartamentos turísticos y las empresas explotadoras de apartamentos turísticos en aquellas comunidades autónomas que sus normativas así lo contemplan (Cataluña y Comunidad Valenciana), inscritos como tales en las correspondientes Consejerías de Turismo de cada Comunidad Autónoma. A tal fin se considera apartamento turístico el inmueble, cuyo uso se cede en alquiler, de modo habitual para hospedaje ocasional, incluyéndose apartamentos propiamente dichos, chalets, villas, bungalows, etc.

Las principales variables analizadas son el número de apartamentos, empresas explotadoras, número de plazas estimadas, pernoctaciones, estancia media, grado de ocupación por plazas y personal ocupado. El nivel de desagregación de la encuesta es a nivel estatal, si bien para algunas variables alcanza el nivel provincial; asimismo en el caso de aquellos municipios en los que la concentración de la oferta turística es significativa, la información se detalla por puntos y por zonas de interés turístico. La información está disponible mensualmente a través de INEbase en *www.ine.es/inebase/*.

d) Encuesta de Ocupación en Alojamientos Rurales

La finalidad de la Encuesta de Ocupación en Alojamientos Rurales es conocer el comportamiento de variables que describen las características de los alojamientos de turismo rural, de acuerdo con las definiciones que sobre ellos figuran en las distintas normativas legales autonómicas. En general, se consideran alojamientos rurales, aquellos establecimientos o viviendas destinadas al alojamiento turístico mediante precio, con o sin otros servicios complementarios y que estén inscritos en el correspondiente Registro de Alojamientos Turísticos de cada Comunidad Autónoma. Estos establecimientos suelen presentar unas características determinadas:

- Están situados en un medio rural.
- Son edificaciones con una tipología arquitectónica propia de la zona o están situados en fincas que mantienen activas explotaciones agropecuarias (agroturismo).
- Ofrecen un número de plazas y habitaciones para el alojamiento de huéspedes limitado.

Las principales variables analizadas son el número de alojamientos de turismo rural, grado de ocupación del establecimiento, estancia media del viajero, pernocta-

ciones, precio medio y personal ocupado. El nivel de desagregación de la encuesta es a nivel estatal, si bien para algunas variables alcanza el nivel provincial. La información está disponible mensualmente a través de INEbase en *www.ine.es/inebase/*.

A continuación se presentan los datos disponibles a comienzos de 2010 para las encuestas señaladas anteriormente:

ENCUESTAS DE OCUPACIÓN (INE)

Ocupación hotelera	**Total**	**% Var. Interanual**
Viajeros en hoteles (residentes y no residentes)	3.783.809	1,4
Pernoctaciones (residentes y no residentes)	11.339.954	0,3
Grado de ocupación (%)	34	–1,6
Ocupación en acampamentos turísticos	**Total**	**% Var. Interanual**
Viajeros en acampamentos (residentes y no residentes)	92.381	–1,9
Pernoctaciones (residentes y no residentes)	831.040	3,0
Grado de ocupación (%)	34	–2,5
Ocupación en apartamentos turísticos	**Total**	**% Var. Interanual**
Viajeros en apartamentos (residentes y no residentes)	402.130	–0,1
Pernoctaciones (residentes y no residentes)	3.415.562	–9,5
Grado de ocupación (%)	31	–9,0
Ocupación en alojamientos de turismo rural	**Total**	**% Var. Interanual**
Viajeros en turismo rural (residentes y no residentes)	106.214	6,0
Pernoctaciones (residentes y no residentes)	298.055	–0,5
Grado de ocupación (%)	8	–6,4

Fuente: Instituto Nacional de Estadística. Encuestas de Ocupación de Alojamientos Turísticos. Enero 2010.

Asimismo dentro de este apartado cabe destacar la **Encuesta de Ocupación Turística (OCUPATUR)** elaborada por la Secretaría de Estado de Turismo, a través del Instituto de Estudios Turísticos, cuyo objetivo es determinar el grado de ocupación en Hoteles y Casas Rurales en épocas de mayor afluencia turística (viajes de temporada, puentes, etc.). No se trata, por tanto, de una encuesta continua, sino que se realiza de forma ocasional en los siguientes periodos:

- Semana Santa.
- Puente del 1 de Mayo.
- Puente del 12 de Octubre.
- Puente de Todos Los Santos.
- Puente de la Constitución.
- Vacaciones de Navidad.

2.3. ESTADÍSTICAS SOBRE LA DEMANDA DEL SECTOR TURÍSTICO

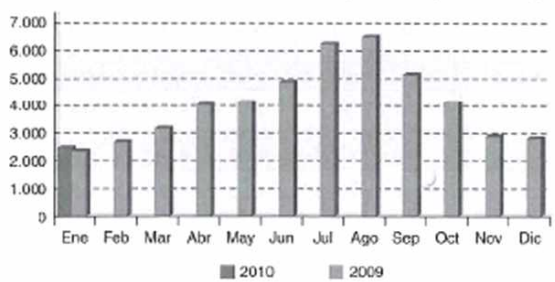
2.3.1. Encuesta de Gasto Turístico (EGATUR)

La Encuesta de Gasto Turístico en España (EGATUR) es realizada de forma continua por el Instituto de Estudios Turísticos con la colaboración del Ministerio de Industria, Turismo y Comercio, el Instituto Nacional de Estadística y el Banco de España. El objetivo de la encuesta es analizar mensualmente el gasto y el comportamiento turístico de los visitantes (turistas y excursionistas) no residentes en España que acceden al país por carretera o aeropuerto.

Las principales variables analizadas son la tipología de viajero, país de residencia, duración de la estancia, gasto medio por persona, medio de transporte, motivo del viaje, y los conceptos de gasto realizado. El nivel de desagregación de la encuesta es a nivel nacional y los resultados de la encuesta pueden consultarse a través de los diferentes informes mensuales, trimestrales y anuales disponibles en el Centro de Documentación Turística de España o a través de Internet en *www.iet.tourspain.es*.

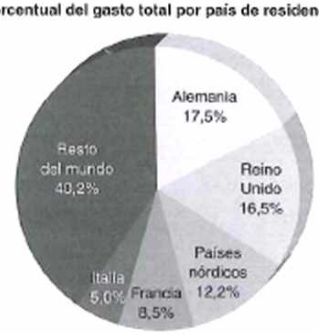
En los siguientes cuadros se presenta un resumen de algunos de los principales resultados de esta encuesta:

Gasto total de los turistas no residentes, según meses (millones de €)



Fuente: Instituto de Estudios Turísticos. EGATUR. Enero 2010.

Distribución porcentual del gasto total por país de residencia (Enero 2010)



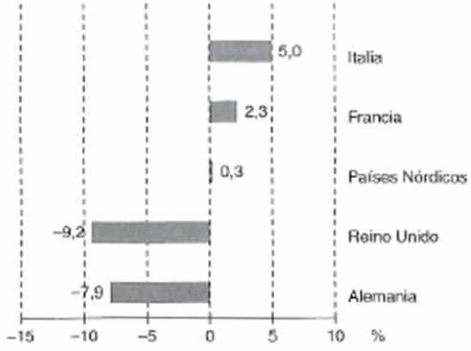
Fuente: Instituto de Estudios Turísticos. EGATUR. Enero 2010.

GASTO DE TURISTAS INTERNACIONALES. (ENERO 2010)

	Gasto total			Gasto medio en €			
	Mill. €	Var inter-anual	% vertical	Por turista	Var inter-anual	Diario	Var inter-anual
SEGÚN PAÍS DE RESIDENCIA							
Total	*2.397*	*4,1*	*100,0*	*947*	*3,2*	*94*	*3,1*
Alemania	420	–7,9	17,5	1.007	–1,7	82	–4,4
Reino Unido	398	–9,2	15,5	754	–1,3	73	3,7
Países Nórdicos	292	0,3	12,2	1.102	1,4	113	0,5
Francia	203	2,3	8,5	620	–2,3	92	5,3
Italia	120	5,0	5,0	711	–11,3	91	–11,1
Resto del mundo	965	19,9	40,2	1.169	11,5	110	5,0
SEGÚN DESTINO PRINCIPAL							
Total	*2.397*	*4,1*	*100,0*	*947*	*3,2*	*94*	*3,1*
Canarias	830	–1,3	34,6	1.043	–2,8	95	–4,0
Cataluña	447	14,6	18,6	795	14,2	104	12,0
Andalucía	355	5,9	14,8	1.026	–3,7	91	17,8
Madrid (C. de)	300	16,8	12,5	1.074	9,8	140	–7,9
C. Valenciana	158	–6,5	6,6	739	–3,6	59	4,8
Baleares (Illes)	97	–25,5	4,1	944	10,7	76	–11,7
Resto CCAA	211	16,7	8,8	911	8,6	89	–6,1
SEGÚN TIPO DE ALOJAMIENTO PRINCIPAL							
Total	*2.397*	*4,1*	*100,0*	*947*	*3,2*	*94*	*3,1*
Hotelero	1.518	1,3	63,3	976	3,4	129	0,1
No hotelero	879	9,5	36,7	901	3,4	64	8,4
SEGÚN FORMA DE ORGANIZACIÓN							
Total	*2.397*	*4,1*	*100,0*	*947*	*3,2*	*94*	*3,1*
Sin paquete tur.	1.592	7,6	66,4	893	3,8	87	4,4
Paquete turístico	805	–2,1	33,6	1.078	3,4	113	1,7

Fuente: Instituto de Estudios Turísticos. EGATUR. Enero 2010.

Porcentaje de variación interanual del gasto total de los turistas no residentes (enero 2010)



Fuente: Instituto de Estudios Turísticos. EGATUR. Enero 2010.

2.3.2. Encuesta sobre Movimientos Turísticos de los Españoles (FAMILITUR)

La Estadística de Movimientos Turísticos de los Españoles (FAMILITUR) es elaborada por el Instituto de Estudios Turísticos en colaboración con el Ministerio de Industria, Turismo y Comercio y constituye la principal fuente estadística para analizar el comportamiento turístico de los españoles. Fue implantada en 1996 a fin de cumplir con las exigencias de información de la Unión Europea, plasmadas en la Directiva 95/57/CE del Consejo sobre la recogida de información estadística en el ámbito del turismo.

La finalidad de la encuesta es la cuantificación y caracterización de los flujos de viajeros españoles entre las distintas Comunidades Autónomas y hacia el extranjero, siempre que impliquen al menos una pernoctación fuera del lugar de residencia e independientemente del motivo.

Las principales variables analizadas son el motivo del viaje, características del individuo entrevistado y la forma de organización del viaje.

El nivel de desagregación de la encuesta es a nivel autonómico y los resultados de la encuesta se publican con un desfase de tres meses con respecto al cuatrimestre de referencia.

La información de esta encuesta puede consultarse a través del informe anual disponible en el Centro de Documentación Turística de España y en Internet en la web *www.iet.tourspain.es*.

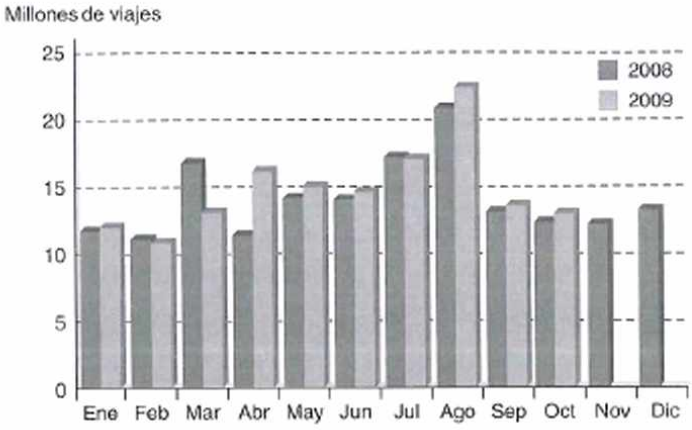
Se presentan en los siguientes cuadros algunos de los resultados de esta encuesta para los años 2008 y 2009.

MOVIMIENTOS TURÍSTICOS DE LOS ESPAÑOLES (FAMILITUR).
OCTUBRE 2009. DATOS PROVISIONALES

VIAJES DE LOS RESIDENTES EN ESPAÑA	Mensual			Acumulado		
	Viajes totales	% Variación inter-anual	Porcentajes verticales	Viajes totales	% Variación inter-anual	Porcentajes verticales
TOTAL	**13.079.289**	**4,0**	**100**	**148.185.512**	**1,5**	**100**
SEGÚN DESTINO PRINCIPAL						
Turismo emisor	908.588	–0,5	6,9	11.015.744	6,9	7,4
Turismo interno	12.170.701	4,3	93,1	137.169.764	1,0	92,6
Andalucía	1.854.969	–3,8	14,2	25.361.752	1,0	17,1
Cataluña	1.787.615	3,7	13,7	19.883.472	0,1	13,4
Castilla y León	1.405.843	3,5	10,7	14.425.659	2,2	9,7
Com. Valenciana	1.265.470	–2,3	9,7	16.216.962	6,8	10,9
Resto	5.856.805	9,2	44,8	61.281.923	–0,3	41,4
SEGÚN COMUNIDAD AUTÓNOMA DE ORIGEN						
Madrid (Comunidad de)	2.269.609	8,5	17,4	26.174.150	0,2	17,7
Cataluña	2.084.020	4,2	15,9	24.454.787	0,3	16,5
Andalucía	1.855.296	3,3	14,3	23.615.769	4,8	15,9
Comunidad Valenciana	1.477.420	–2,7	11,3	15.981.130	6,5	10,8
Resto	5.382.944	4,2	41,2	57.949.677	–0,1	39,1
SEGÚN MOTIVO PRINCIPAL						
Ocio, recreo, vacaciones	6.679.416	6,3	51,1	82.286.289	4,3	55,5
Visita a familiares o amigos	3.164.930	15,5	24,2	33.645.163	0,9	22,7
Trabajo/Negocios	1.969.937	–17,5	15,1	22.013.045	–3,9	14,9
Estudios	937.967	2,3	7,2	7.877.300	–8,6	5,3
Otros motivos	327.038	30,0	2,5	2.363.715	2,1	1,6
SEGÚN TIPO DE ALOJAMIENTO						
Alojamiento hotelero	2.655.905	–1,3	20,3	29.868.802	0,5	20,2
Alojamiento no hotelero	10.423.382	5,4	79,7	118.316.700	1,7	79,8
SEGÚN MEDIO DE TRANSPORTE						
Transporte por carretera	10.991.082	1,5	84,0	126.523.287	1,7	85,4
Avión	1.142.021	4,2	8,7	13.071.136	–4,1	8,8
Otros	946.185	43,0	7,2	8.572.527	7,0	5,8

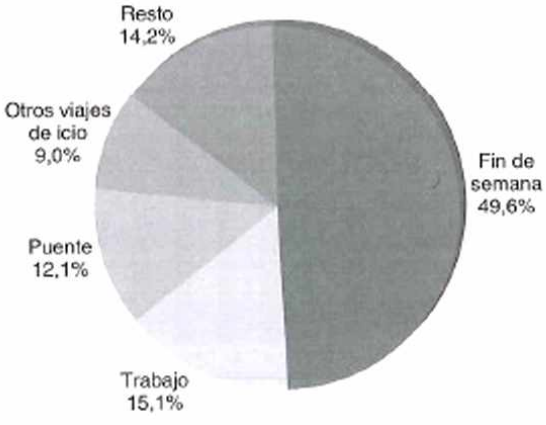
Fuente: Instituto de Estudios Turísticos. FAMILITUR. Octubre 2009.

Viajes de los residentes en España por meses, 2008 y 2009. Datos provisionales



Fuente: Instituto de Estudios Turísticos. FAMILITUR. Octubre 2009.

Viajes de los residentes en España según tipo de viaje realizado
Octubre 2009. Datos provisionales



Fuente: Instituto de Estudios Turísticos. FAMILITUR. Octubre 2009.

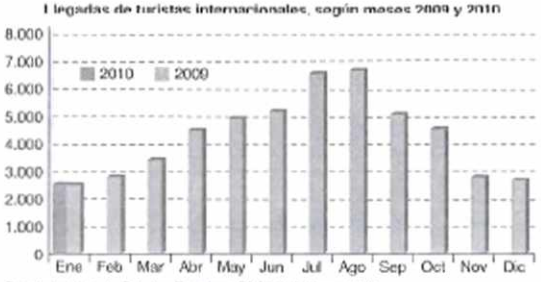
2.3.3. Estadística de Movimientos Turísticos en Fronteras (FRONTUR)

La Estadística de Movimientos Turísticos en Fronteras (FRONTUR) elaborada por el Instituto de Estudios Turísticos es el instrumento básico de observación de la Secretaría de Estado de Turismo y Comercio. Se trata de una estadística de carácter continuo con periodicidad mensual en la que también participan el Ministerio de Industria, Turismo y Comercio, la Dirección General de Tráfico, AENA, RENFE y Puertos del Estado.

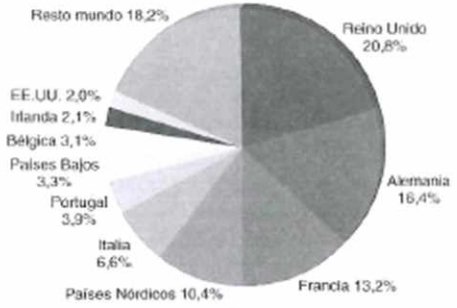
Frontur permite cuantificar los flujos turísticos en función de la vía de acceso (carretera, aeropuerto, puerto marítimo y ferrocarril), tanto de los residentes en España que van hacia el extranjero o regresan de él como de los extranjeros que entran, salen o transitan por España, a fin de conocer el comportamiento turístico de éstos en sus desplazamientos.

Las principales variables analizadas son el país de residencia, comunidad autónoma de destino, duración de la estancia, vías de acceso, motivos del viaje, tipo de alojamiento, formas de organización del viaje, etc. El nivel de desagregación de la encuesta es a nivel autonómico y los resultados de la encuesta se publican mensualmente con un desfase de 15 días tras la finalización del mes de referencia. La información de esta encuesta puede consultarse a través de los informes mensuales, de temporada de verano, anuales y de previsiones de vuelo disponibles en el Centro de Documentación Turística de España y en Internet en la web *www.iet.tourspain.es*.

Los resultados para esta encuesta en enero de 2010 se presentan a continuación de manera resumida en los siguientes cuadros.



Fuente: Instituto de Estudios Turísticos. FRONTUR. Enero 2010.



Fuente: Instituto de Estudios Turísticos. FRONTUR. Enero 2010.

LLEGADAS DE TURISTAS INTERNACIONALES. (ENERO 2010)

	Mensual		
	Total turistas	**Porcentajes verticales**	**Variación interanual**
TOTAL	*2.545.626*	*100*	*1,1*
Reino Unido	528.306	20,8	–8,1
Alemania	417.005	16,4	–6,3
Francia	336.014	13,2	5,5
Países Nórdicos	264.767	10,4	–1,1
Italia	169.241	6,6	19,1
Portugal	98.322	3,9	–21,7
Países Bajos	83.881	3,3	–18,1
Bélgica	77.993	3,1	5,9
Irlanda	54.512	2,1	–25,6
Suiza	55.838	2,2	16,5
Resto de Europa	200.916	7,9	34,1
Estados Unidos de América	51.684	2,0	2,9
Resto de América	79.931	3,1	14,8
Resto del mundo	127.217	5,0	63,9
TOTAL	*2.545.626*	*100*	*1,1*
Canarias	795.815	31,3	1,6
Cataluña	571.143	22,4	0,9
Andalucía	345.612	13,6	9,9
Madrid (Comunidad de)	282.719	11,1	7,1
Comunidad Valenciana	213.842	8,4	–3,0
Balears (Illes)	103.063	4,0	–32,7
Resto de CC.AA.	233.431	9,2	7,6
TOTAL	*2.545.626*	*100*	*1,1*
Aeropuertos	2.075.808	81,5	2,0
Carreteras	372.672	14,6	–4,7
Otros	97.146	3,8	6,5
TOTAL	*2.545.626*	*100*	*1,1*
Alojamiento hotelero	1.556.871	61,2	–1,9
Alojamiento no hotelero	974.872	38,3	7,3
Vivienda propia y de familiares o amigos	685.186	26,9	15,2
Vivienda alquilada	183.493	7,2	–3,6
Otros alojamientos	106.194	4,2	–14,1
Sin especificar	13.883	0,5	---
TOTAL	*2.545.626*	*100*	*1,1*
Sin paquete	1.724.462	67,7	2,9
Con paquete	812.685	31,9	–1,3

Fuente: Instituto de Estudios Turísticos. FRONTUR. Enero 2010.

2.3.4. Balanza de Pagos

La Balanza de Pagos es elaborada mensualmente desde 1999 por el Banco de España y registra las transacciones entre residentes españoles con el resto del mundo, con independencia de su nacionalidad. Para su elaboración, siguiendo las directrices y recomendaciones establecidas por el Quinto Manual del Fondo Monetario Internacional, se utiliza un sistema basado en el registro de las transacciones internacionales comunicadas por determinados agentes económicos que están obligados a informar directamente al propio Banco de España de todas aquellas operaciones que hayan realizado con unidades no residentes, tales como transferencias bancarias identificadas como viajes, compra y venta de divisas extranjeras en bancos y oficinas de cambio, movimientos de divisas entre bancos nacionales y extranjeros y pagos con tarjeta de crédito.

La información relevante relacionada con el sector turístico podemos encontrarla en la rúbrica de Turismo y Viajes dentro de la Balanza por Cuenta Corriente, en la que se incluyen los bienes y servicios adquiridos por residentes en España que se desplazan al extranjero y por residentes en otros países en España que se desplazan para fines de negocios o personales, incluidos los de salud y educación, con estancias inferiores a un año.

La difusión de los resultados se realiza a través de la monografía del mismo nombre publicada por el Banco de España. Asimismo es posible consultar la información de la Balanza de Pagos en Internet a través de la web del Banco de España (*www.bde.es*) y del Instituto Nacional de Estadística dentro de INEbase (*www.ine.es/inebase/*) con el nombre de «Ingresos y Pagos por Turismo».

2.4. INDICADORES COYUNTURALES

En este grupo de estadísticas se incluyen un conjunto de indicadores que permiten seguir la evolución temporal de aspectos coyunturales del sector, tales como los precios, los gastos y los ingresos. Entre ellas tenemos las siguientes:

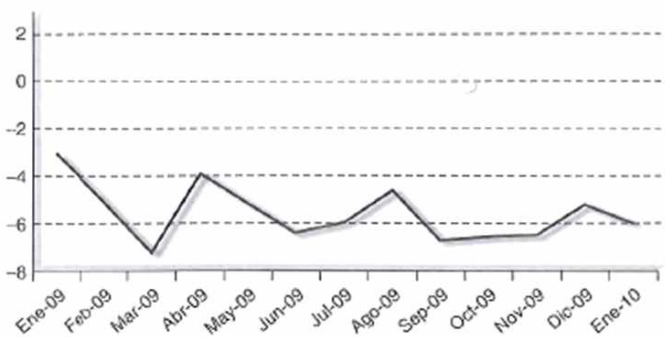
2.4.1. Índice de Precios Hoteleros

El Índice de Precios Hoteleros (IPH) es elaborado mensualmente por el Instituto Nacional de Estadística y permite medir la evolución del conjunto de precios aplicados por los empresarios a los distintos clientes que se alojan en los hoteles de España, es decir, desde el punto de vista de la oferta.

Las variables analizadas son los precios por habitación, por categoría hotelera y las tarifas de precios (normal, fin de semana, especiales a empresas, grupos y tour-operadores). El nivel de desagregación de la encuesta es autonómico y los resultados de la encuesta se publican al final de cada mes con referencia a datos del mes anterior. La información de esta encuesta puede consultarse a través del Boletín Mensual de Estadística y en Internet a través de INEbase (*www.ine.es/inebase/*).

A continuación se muestra la evolución reciente de este índice así como su desglose por categorías hoteleras:

IPH Tasa de variación interanual



Fuente: Instituto de Estudios Turísticos. IPH. Enero 2010.

IPH. DESGLOSE POR CATEGORÍAS. ENERO 2010. DATOS PROVISIONALES

	Total	Variación intermensual	Variación interanual
Total	**91,1**	**–3,9**	**6,0**
Hoteles: Estrellas oro			
Cinco	86	–6,3	–7,8
Cuatro	89	–4,9	–6,7
Tres	93,3	–2,1	–5,5
Dos	97,3	–0,6	–1,6
Una	97,1	97,1	–3,1
Hostales: Estrellas plata			
Tres y dos	95,4	–2,5	–2,3
Una	98,9	–0,2	–2,0

Fuente: Instituto Nacional de Estadística. IPH. Enero 2010.

2.4.2. Indicadores de Rentabilidad del Sector Hotelero

En el año 2010 el INE ha comenzado a publicar bajo el nombre de Indicadores de Rentabilidad del Sector Hotelero que analiza dos nuevas variables:

- ADR (*Average Daily Rate*): tarifa media facturada por el servicios de alojamiento en habitación doble con baño, no incluyendo el IVA ni otros servicios.
- RevPAR (*Revenue Per Available Room*): ingresos medios por habitación disponible.

Los datos utilizados para calcular estas variables se extraen de la información que los establecimientos hoteleros declaran en el cuestionario de la Encuesta de Ocupación Hotelera.

El objetivo de estos indicadores, que sustituyen al Índice de Ingresos Hoteleros, es reflejar los cambios acaecidos en el sector a lo largo de estos últimos años, recogiendo los nuevos canales de distribución propiciados por la generalización del uso de Internet, y adecuando el cuestionario a los términos y conceptos utilizados más frecuentemente por el sector, así como a los datos disponibles en los propios sistemas de gestión de las empresas.

El nivel de desagregación de la encuesta es autonómico y los resultados de la encuesta se publican al final de cada mes con referencia a datos del mes anterior. La información de esta encuesta puede consultarse a través del Boletín Mensual de Estadística y en Internet a través de INEbase (*www.ine.es/inebase/*).

En los siguientes cuadros se recogen el ADR y el RevPAR para las diferentes Comunidades Autónomas así como su desglose por categoría hotelera:

INDICADORES DE RENTABILIDAD DEL SECTOR HOTELERO. (ENERO 2010)
DATOS PROVISIONALES

ADR y RevPAR de comunidades autónomas y total nacional				
	ADR (en euros)	**Tasa de variación interanual**	**RevPar (en euros)**	**Tasa de variación interanual**
TOTAL	**67,2**	**–4,0**	**26,2**	**–4,5**
Andalucía	58,7	–2,0	17,3	–8,2
Aragón	63,2	–2,9	19,8	10,0
Asturias (Principado de)	58,8	1,9	11,6	–7,7
Balears (Illes)	57,0	–0,6	23,3	–8,6
Canarias	72,4	–5,4	50,8	–2,2
Cantabria	61,2	–4,9	15,7	20,6
Castilla y León	54,6	–4,0	12,0	–11,6
Castilla-La Mancha	58,9	1,8	12,7	–11,4
Cataluña	77,3	–6,8	27,2	–8,0
Comunidad Valenciana	47,8	–3,9	16,8	–8,0

ADR y RevPAR de comunidades autónomas y total nacional				
	ADR (en euros)	**Tasa de variación interanual**	**RevPar (en euros)**	**Tasa de variación interanual**
Extremadura	56,8	–5,4	11,9	–14,5
Galicia	49,5	0,4	11,0	–1,9
Madrid (Comunidad de)	82,2	–8,1	39,8	–5,6
Murcia	55,6	–11,5	16,5	–13,7
Navarra (Comunidad Foral de)	66,6	3,5	15,8	–11,7
País Vasco	66,3	–9,1	24,1	–9,5
La Rioja	58,4	–3,8	16,8	–9,2
Ceuta	68,8	–4,4	33,6	–16,4
Melilla	69,3	3,0	24,2	–9,0

ADR y RevPAR nacional y desglose por categorías				
	ADR (en euros)	**Tasa de variación interanual**	**RevPar (en euros)**	**Tasa de variación interanual**
TOTAL	**67,2**	**–4,0**	**26,2**	**–4,5**
Hoteles: Estrellas oro				
Cinco	138,9	–9,5	55,3	–1,6
Cuatro	73,1	–6,3	34,0	–5,9
Tres	54,7	–3,7	23,1	–6,3
Dos	50,0	0,5	15,2	–7,6
Una	47,3	2,2	11,0	–10,3
Hostales: Estrellas plata				
Tres y dos	41,4	–3,1	9,5	–10,6
Una	33,9	–1,0	8,2	8,4

Fuente: Instituto Nacional de Estadística. IRSH. Enero 2010.

2.4.3. Índice de Precios de Acampamentos Turísticos

El Índice de Precios de Acampamentos Turísticos mide la evolución mensual del conjunto de las principales tarifas de precios que los campings aplican a sus clientes. Proporciona información a nivel nacional y desglosado por tipo de tarifa y por categoría del establecimiento.

Para su obtención se utiliza la Encuesta de Ocupación en Acampamentos Turísticos con la información que se recoge, mensualmente, de todos los acampamentos tu-

rísticos del territorio español a los que se les envía un cuestionario. A partir de esta encuesta se obtiene información sobre la ocupación de los campings (viajeros entrados, pernoctaciones, grado de ocupación, etc.), su estructura (plazas, personal, etc.) y demás variables de interés, con una amplia desagregación geográfica y por categorías de los establecimientos (lujo y primera, segunda y tercera). En el cuestionario se les piden los precios aplicados a distintos tipos de clientes por la ocupación de una parcela, así como el porcentaje de aplicación de cada una de las tarifas.

2.4.4. Índice de Precios de Apartamentos Turísticos

El Índice de Precios de Apartamentos Turísticos mide la evolución mensual del conjunto de las principales tarifas de precios que los establecimientos de apartamentos turísticos aplican a sus clientes. Proporciona información a nivel nacional y desglosado por tipo de tarifa y por modalidad del apartamento.

Para su obtención se utiliza la Encuesta de Ocupación en Apartamentos Turísticos con la información que se recoge, mensualmente, de alrededor de 3.000 establecimientos de apartamentos turísticos en verano y 2.000 en invierno, a los que se les envía un cuestionario. A partir de esta encuesta se obtiene información sobre la ocupación de los apartamentos turísticos (viajeros entrados, pernoctaciones, grado de ocupación, etc.), su estructura (plazas, personal, etc.) y demás variables de interés, con una amplia desagregación geográfica. En el cuestionario se les piden los precios aplicados a distintos tipos de clientes por la ocupación de un apartamento (para distintas modalidades de apartamentos), así como el porcentaje de aplicación de cada una de las tarifas.

2.4.5. Índice de Precios de Alojamientos de Turismo Rural

El Índice de Precios en Alojamientos de Turismo Rural mide la evolución mensual de los precios que los empresarios de este tipo de establecimientos aplican a sus clientes. Proporciona información a nivel nacional desglosado por tipo de tarifa y modalidad de alquiler.

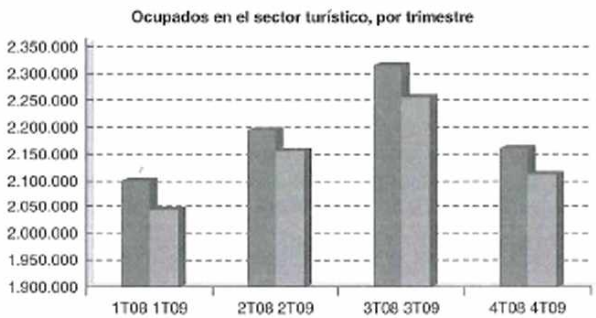
El Índice de Precios en Alojamientos de Turismo Rural es una medida estadística de la evolución mensual de los precios que los empresarios de este tipo de establecimientos aplican a sus clientes. Para su obtención se utiliza la Encuesta de Ocupación en Alojamientos de Turismo Rural con la información que se recoge, mensualmente, de alrededor de 5.300 establecimientos a los que se les envía un cuestionario. A partir de esta encuesta se obtiene información sobre la ocupación de este tipo de establecimiento (viajeros entrados, pernoctaciones, grado de ocupación, etc.), su estructura (plazas, personal, etc.) y demás variables de interés, con una amplia desagregación geográfica. En el cuestionario se les piden los precios aplicados a distintos tipos de clientes por la ocupación de una habitación doble y/o vivienda completa (según la modalidad de alquiler de la vivienda), así como el porcentaje de aplicación de cada una de las tarifas.

2.5. ESTADÍSTICAS SOBRE EL EMPLEO EN EL SECTOR TURÍSTICO

Si bien no existe una operación estadística específica para el empleo del sector turístico español, podemos encontrar algunas cifras de interés en las siguientes fuentes:

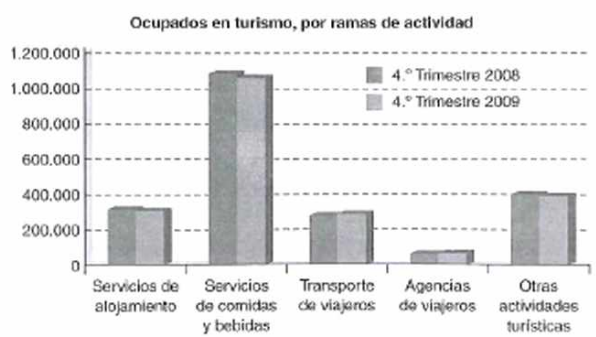
- La **Encuesta de Población Activa (EPA)** es elaborada con periodicidad trimestral por el Instituto Nacional de Estadística cuyo objetivo es informar sobre la actividad económica de los ocupados, parados, activos e inactivos. El Instituto de Estudios Turísticos elabora con la misma periodicidad una explotación específica de los ocupados en las actividades características de la industria turística y que puede encontrarse en la web *www.iet.tourspain.es*.

  Los últimos datos disponibles a finales de 2009 reflejan que actualmente hay cerca de 2,1 millones de empleados en el sector turístico español, representando el 11,3% del total de ocupados en España. Por su parte, el total de parados era de 357.309, lo que supone una tasa de paro del 14,5%, inferior a la del total nacional (18,8%), pero superior a la del sector servicios (9,6%).



Fuente: Instituto de Estudios Turísticos. Elaboración a partir de datos de la EPA.

Las actividades turísticas mayoritarias, por número de ocupados, son los servicios de comidas y bebidas (50,7% del total de ocupados en turismo), los servicios de alojamiento (14,4%) y el transporte de viajeros con el 13,5%.



Fuente: Instituto de Estudios Turísticos. Elaboración a partir de datos de la EPA.

- El Instituto de Estudios Turísticos también realiza una explotación similar de los datos de **Afiliación de Trabajadores al Sistema de Seguridad Social** que mensualmente publica el Ministerio de Trabajo e Inmigración.
- Otras fuentes estadísticas en las que es posible encontrar información relacionada con el empleo en el sector turístico español son:
  - La Encuesta de Coyuntura Laboral elaborada trimestralmente por el Ministerio de Trabajo e Inmigración, en la que se recoge información sobre el mercado de trabajo desde el punto de vista de las empresas.
  - La Encuesta Trimestral de Coste Laboral elaborada por el Instituto Nacional de Estadística, en la que se ofrecen niveles e indicadores sobre el coste laboral medio por trabajador y mes, el coste laboral medio por hora efectiva de trabajo y el tiempo trabajado y no trabajado.
  - La Estadística de Contratos Registrados publicada mensualmente por el Instituto Nacional de Empleo recoge estadísticas sobre paro registrado con desagregación de datos a nivel provincial, de edad y sexo y de los demandantes de empleo y sectores económicos.

2.6. ESTADÍSTICAS DE SÍNTESIS

2.6.1. Cuenta Satélite del Turismo en España

La Cuenta Satélite del Turismo de España es elaborada por el Instituto Nacional de Estadística en colaboración con el Instituto de Estudios Turísticos y el Banco de España. Está compuesta por un conjunto de cuentas y tablas que presenta los distintos parámetros económicos del turismo en España, obteniendo una representación sistemática y comparable de la actividad turística adaptándose a los conceptos y clasificaciones que figuran en el manual elaborado por la Organización Mundial del Turismo.

La Cuenta Satélite comprende tres tipos de elementos:

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

Hasta el momento se han publicado la serie contable desde 1995 a 2003 con base año 1995 y la serie contable 2000-2008 con base año 2000. El nivel de desagregación de la encuesta es nacional y los resultados pueden consultarse a través del monográfico anual «La Cuenta Satélite del Turismo en España» y a través de Internet en INEbase (*www.ine.es/inebase/*).

2.6.2. Ficha de Coyuntura Turística

El Instituto de Estudios Turísticos elabora mensualmente una recopilación de las principales cifras publicadas a lo largo del mes por los diferentes organismos con el objetivo de ofrecer una visión de conjunto del sector. Entre las cifras recopiladas se encuentran:

- Las cifras de viajeros de las encuestas FRONTUR, EGATUR y FAMILITUR.
- Las principales magnitudes de las diferentes Encuestas de Ocupación publicadas por el Instituto Nacional de Estadística.
- Los diferentes índices de precios (IPC, IPH) y de ingresos hoteleros.
- Datos relativos a la Balanza de Pagos elaborada por el Banco de España y al empleo en el sector a partir de la información suministrada por el Ministerio de Trabajo e Inmigración.

Los datos de la Ficha de Coyuntura Turística pueden consultarse en la página web del Instituto de Estudios Turísticos, *http://www.iet.tourspain.es/*.

2.7. ESTADÍSTICAS REALIZADAS POR LAS COMUNIDADES AUTÓNOMAS

Muchas de las Comunidades Autónomas españolas, particularmente aquellas en la que el turismo tiene un especial interés económico, realizan estudios periódicos u ocasionales sobre este sector. En las páginas siguientes se presenta un cuadro resumen con las más relevantes.

Comunidad Autónoma	Encuestas	Organismos que participan en su elaboración
Andalucía	• Balance del Año Turístico en Andalucía • Informe Mensual de Coyuntura Turística de Andalucía. • Encuesta de Coyuntura Turística de Andalucía. • Encuesta de Ocupación en Acampamentos Turísticos. Resultados de Andalucía. • Encuesta de Ocupación en Alojamientos Hoteleros. Resultados de Andalucía	Consejería de Turismo, Comercio y Deporte de la Junta de Andalucía. Instituto de Estadística de Andalucía.
Aragón	• Directorio de Alojamientos Turísticos en Aragón. • Anuario Estadístico de Turismo. • Guía de Servicios Turísticos de Aragón. • Movimiento Turístico. Hoteles, campings, turismo rural y apartamentos. • Información Estadística Económica Coyuntural.	Departamento de Economía, Hacienda y Empleo del Gobierno de Aragón. Instituto Aragonés de Estadística.
Baleares	• Pasajeros llegados. • Pasajeros llegados en cruceros. • Turistas internacionales llegados vía aérea. • Estancia media. • Pernoctaciones en establecimientos hoteleros. • Número de viajeros.	Consejería de Economía y Hacienda del Gobierno de las Illes Balears. Instituto de Estadística de las Illes Balears.
Canarias	• Encuesta sobre el Gasto Turístico. • Encuesta a usuarios de campos de golf. • Encuesta de Alojamiento en Establecimientos Hoteleros. • Encuesta de Alojamiento en Establecimientos Extrahoteleros. • Encuesta de Expectativas Hoteleras. • Recopilación de Estadísticas de Infraestructura Turística.	Instituto Canario de Estadística.

(Continúa)

Comunidad Autónoma	Encuestas	Organismos que participan en su elaboración
Castilla-La Mancha	• Boletines de Turismo. • Informe de Hábitos Turísticos en Castilla-La Mancha. • Indicadores de Actividad Turística. • Turismo Extranjero en Castilla-La Mancha. • Turismo Interno. • Turismo Potencial.	Sistema de Investigación Turística de Castilla-La Mancha (SITdCLM), iniciativa conjunta del Instituto de Promoción Turística y la Universidad de Castilla-La Mancha.
Cataluña	• Viajes de los catalanes. • Indicadores de actividad hotelera de Cataluña. • Indicadores de actividad en campings de Cataluña. • Indicadores de actividad en turismo rural. Informe trimestral. • Anuario Estadístico de Cataluña.	Institut d'Estadística de Catalunya (Idescat).
Comunidad Valenciana	• Oferta turística municipal y comarcal. • Enquesta turística – Històric. • Impactur. • Perfil del turista que visita la Comunitat Valenciana. • Perfil del turista alojado en oferta reglada.	Observatorio Turístico. Conselleria de Turisme de la Generalitat Valenciana.
Murcia	• Oferta y demanda turística en la Costa Cálida.	Centro Regional de Estadística de Murcia.
País Vasco	• Encuesta de Establecimientos Turísticos Receptores.	Instituto Vasco de Estadística.

2.8. EJERCICIOS DE AUTOEVALUACIÓN

1. La encuesta que nos permite conocer el grado de ocupación de hoteles y casas rurales en épocas de elevada afluencia turística es:

   a) Frontur.
   b) Egatur.
   c) Ocupatur.
   d) Familitur.

2. Para conocer el gasto realizado por turistas internacionales en España debemos consultar los resultados de:

   a) Frontur.
   b) Egatur.
   c) Ocupatur.
   d) Familitur.

3. Si queremos conocer los motivos por los que viajan los residentes en España debemos utilizar los resultados de:

   a) Frontur.
   b) Egatur.
   c) Ocupatur.
   d) Familitur.

4. Para saber el medio de transporte utilizado habitualmente por los turistas británicos consultaremos los resultados de:

   a) Frontur.
   b) Egatur.
   c) Ocupatur.
   d) Familitur.

5. El director de un hotel le encarga que busque información relativa a la evolución de los precios aplicados en el sector hotelero en la categoría de cinco estrellas. Para obtener esa información, Vd. consultaría:

   a) Los Indicadores de Rentabilidad del Sector Hotelero.
   b) El Índice de Precios al Consumo.
   c) El Índice de Precios Hoteleros.
   d) La Cuenta Satélite del Turismo.

6. El Censo Continuo de Establecimientos Hoteleros es elaborado por:

   a) Instituto Nacional de Estadística.
   b) Banco de España.
   c) Turespaña.
   d) Ministerio de Economía.

7. El organismo que difunde la información relativa a la Encuesta de Ocupación en Alojamientos Rurales es:

   a) Banco de España.
   b) Instituto Nacional de Estadística.
   c) Turespaña.
   d) Ministerio de Economía.

8. Uno de los objetivos de la encuesta Frontur es:

   a) Analizar el gasto de los visitantes no residentes en España que acceden al país por carretera o aeropuerto.
   b) Caracterizar los flujos de viajeros españoles entre las distintas Comunidades Autónomas y hacia el extranjero.
   c) Determinar el grado de ocupación de hoteles y casas rurales en Semana Santa.
   d) Caracterizar los flujos turísticos de los residentes en España que viajan hacia el extranjero o regresan de él, así como el de los extranjeros que entran, salen o transitan por España.

9. ¿En qué encuesta podemos encontrar datos relativos a las compras realizadas en España por turistas procedentes de Holanda?

   a) Frontur.
   b) Balanza de Pagos.
   c) Ficha de Coyuntura Turística.
   d) Ocupatur.

10. ¿A qué encuesta debe acudir para conocer los ingresos medios por habitación que obtienen los hoteles situados en Navarra?

    a) Índice de Precios Hoteleros.
    b) Indicadores de Rentabilidad del Sector Hotelero.
    c) Ocupatur.
    d) Balanza de Pagos.

LECTURAS RECOMENDADAS

Instituto de Estudios Turísticos (2008). *Balance de Resultados de Demanda Turística Internacional 2004-2007 desde la Óptica de los Mercados Emisores.*

Instituto de Estudios Turísticos (2008). *El Turismo Español en Cifras 2007.*

Instituto de Estudios Turísticos (2009). *España en Europa. El Comportamiento Turístico de los Residentes en la Unión Europea.*

PALABRAS CLAVE

- Censos
- Encuestas Estructurales
- Estadísticas de Ocupación
- EGATUR
- FAMILITUR
- FRONTUR
- Balanza de Pagos
- Indicadores Coyunturales
- Empleo

CAPÍTULO 3

Distribuciones de frecuencias unidimensionales

ESQUEMA

OBJETIVOS

Al finalizar el estudio de este capítulo, el alumno deberá ser capaz de:

1. Construir e interpretar tablas de distribuciones de frecuencias.
2. Conocer y utilizar con fluidez los conceptos y la notación desarrollada a lo largo del capítulo.
3. Determinar la mejor manera de agrupar la información en base a los datos de los que se dispongan.
4. Representar distribuciones de frecuencias mediante diferentes tipos de gráficos.

3.1. INTRODUCCIÓN

Como veíamos en el primer capítulo, habitualmente el propósito de la Estadística es el de sacar conclusiones de una población en estudio, examinando solamente una parte de ella denominada muestra. A este proceso se le denomina Inferencia Estadística, si bien normalmente suele venir precedido de otro en el que se aplican diferentes de Estadística Descriptiva, mediante las que información obtenida a partir de los datos es organizada a fin de condensarla y sintetizarla para extraer sus características más importantes. Generalmente este proceso, denominado ***tabulación***, consiste en ordenar de menor a mayor los valores de la variable analizada y agrupar todos aquellos valores contando el número de veces que se repiten.

Al conjunto resultante de k valores diferentes de la variable X, denotados por $x_1, x_2, \ldots, x_k$, ordenados de menor a mayor, acompañados de sus respectivas frecuencias absolutas $n_1, n_2, \ldots, n_k$, se le denomina ***distribución de frecuencias unidimensional.***

Para la construcción de una distribución de frecuencias es necesario definir previamente algunos conceptos. Es recomendable familiarizarse con ellos y con su notación, ya que serán utilizados a lo largo del resto de capítulos del libro:

- Frecuencia absoluta (n_i): se trata del número de repeticiones que se presenta una observación, es decir, del número de veces que aparece cada uno de los valores de una variable o cada una de las modalidades de un atributo.
- Frecuencia total (N): es el número total de datos considerados. Si se parte de la población, la frecuencia total será el tamaño de la población, N. Si se parte de las modalidades o valores de una muestra, la frecuencia total será el tamaño de la muestra, N. En todo caso, se verifica siempre que:

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

- Frecuencia relativa (f_i): es el cociente entre cada frecuencia absoluta y la frecuencia total. Es decir:

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

  La frecuencia relativa refleja la proporción en tanto por uno de individuos de cada modalidad, y nos da una idea de la importancia que una modalidad o un valor poseen respecto al total. La suma de todas las frecuencias relativas siempre debe ser igual a la unidad, tal que $\sum_{i=1}^{k} f_i = 1$.

- Frecuencia absoluta acumulada (N_i): es la suma de la frecuencia absoluta del dato con las frecuencias absolutas de todos los datos anteriores. La última frecuencia absoluta acumulada es igual a la frecuencia total tal que:

$$N_1 = n_1$$
$$N_1 = n_1 + n_2$$
$$\ldots$$
$$N_1 = n_1 + n_2 + \ldots + n_{k-1} + n_k = N$$

- Frecuencia relativa acumulada (F_i): es el cociente entre la frecuencia absoluta acumulada y la frecuencia total tal que $F_i = \frac{N_i}{N}$. Por tanto, al igual que la frecuencia relativa, se expresa en tantos por uno. También se puede definir como la suma de la frecuencia relativa del dato con las frecuencias relativas de todos los datos anteriores. La última frecuencia relativa acumulada es igual a la unidad.

La forma genérica de presentar la información relativa a una distribución de frecuencias tendrá un aspecto similar al de la siguiente tabla:

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
...	...	...	...	...
x_k	n_k	N_k	f_k	$F_k = 1$

Ejemplo 3.1. Se disponen de los siguientes datos relativos a los ingresos medios por habitación de una muestra de hoteles: 90 hoteles han facturado 30 euros de media por habitación; 35 hoteles han ingresado 150 euros de media; 95 han obtenido 80 euros; 105, una media de 40 euros por habitación y 75 han facturado 50 euros de media por cada habitación. Con estos datos vamos a construir la tabla con la distribución de frecuencias.

En primer lugar debemos identificar la característica X que en este caso es el ingreso medio por habitación. Los posibles valores que puede adoptar son, ordenados de menor a mayor, 30, 40, 50, 80 y 150 euros.

x_i	n_i	N_i	f_i	F_i
30	90	90	0,225	0,225
40	105	195	0,2625	0,4875
50	75	270	0,1875	0,675
80	95	365	0,2375	0,9125
150	35	400	0,0875	1
Suma total	400	–	1	–

A continuación se muestran los cálculos realizados para obtener las frecuencias absolutas acumuladas, frecuencias relativas y frecuencias relativas acumuladas:

Frecuencias absolutas acumuladas

$$N_1 = 90;$$
$$N_2 = 90 + 105 = 195;$$
$$N_3 = 90 + 105 + 75 = 270$$
$$N_4 = 90 + 105 + 75 + 95 = 365;$$
$$N_5 = 90 + 105 + 75 + 95 + 35 = 400$$

Frecuencias relativas

$$f_1 = \frac{n_1}{N} = \frac{90}{400} = 0{,}225;\ f_2 = \frac{105}{400} = 0{,}2625;\ f_3 = \frac{75}{400} = 0{,}1875$$

$$f_4 = \frac{95}{400} = 0{,}2375;\ f_5 = \frac{105}{400} = 0{,}0875$$

Frecuencias relativas acumuladas

$$F_1 = \frac{N_1}{N} = \frac{90}{400} = 0{,}225;\ F_2 = \frac{195}{400} = 0{,}4875;\ F_3 = \frac{270}{400} = 0{,}675$$

$$F_4 = \frac{365}{400} = 0{,}9125;\ F_5 = \frac{400}{400} = 1$$

3.2. TIPOS DE DISTRIBUCIONES DE FRECUENCIAS

Podemos distinguir dos tipos fundamentales de distribuciones:

- *Distribuciones de frecuencias con datos no agrupados*, en las que simplemente cada valor de la variable x_i lleva asociado una frecuencia n_i. La utilización de este tipo de distribuciones es adecuada en aquellos casos en los que la variable X toma pocos valores, pero se repiten un gran número de veces.

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
...	...	...	...	...
x_k	n_k	N_k	f_k	$F_k = 1$

Un caso particular de estas distribuciones son las distribuciones unitarias en las que todas las frecuencias absolutas son *unitarias* tal que $n_i = 1$ para todo $i = 1, 2, \ldots, k$:

x_i	n_i	N_i	f_i	F_i
x_1	1	1	$1/N$	$1/N$
x_2	1	2	$1/N$	$2/N$
...	...	...	...	...
x_k	1	N	$1/N$	$F_k = 1$

- *Distribuciones de frecuencias con datos agrupados en intervalos*, en las que los valores de la variable quedan agrupados en intervalos, lo cual puede resultar de utilidad cuando el número de valores de la variable es muy elevado, si bien a cambio deberemos estar dispuestos a perder parte de la información. Evidentemente sólo podremos agrupar en intervalos a las variables de tipo cuantitativo.

  La notación utilizada para designar al intervalo i-ésimo en este tipo de distribuciones es $[L_{i-1}, L_i)$, donde L_{i-1} es el límite inferior (valor de la variable más pequeño en él), y L_i es el límite superior (valor mayor de la variable en él). Por convenio, los intervalos han de ser solapados y semiabiertos por la derecha (cerrados por la izquierda y abiertos por la derecha).

  Por su parte, la amplitud de los intervalos se define como $a_i = L_i - L_{i-1}$, es decir, la diferencia entre el límite superior e inferior. Dicha amplitud puede ser constante, siendo todos los intervalos de la misma amplitud, o variable, variando la amplitud entre intervalos.

  Dado que no es posible operar con los valores de un intervalo, en la práctica recurriremos a la *marca de clase*, denotada por x_i, que se define como el punto medio de un intervalo, tal que $x_i = \dfrac{L_{i-1} + L_i}{2}$.

Las tablas de frecuencias para las distribuciones con datos agrupados por intervalos presentan un formato similar al siguiente:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
$[L_0, L_1)$	x_1	n_1	N_1	f_1	F_1
$[L_1, L_2)$	x_2	n_2	N_2	f_2	F_2
	...	...	...	...	...
$[L_{k-1}, L_k)$	x_k	n_k	N	f_k	$F_k = 1$

Ejemplo 3.2. Supongamos que disponemos de los siguientes datos de gasto en viajes durante las vacaciones navideñas de cinco familias:

2.000, 1.500, 3.000, 2.500, 1.750

Ordenando los valores de menor a mayor podemos construir fácilmente la tabla de esta distribución de tipo unitario:

x_i	n_i	N_i	f_i	F_i
1.500	1	1	0,2	0,2
1.750	1	2	0,2	0,4
2.000	1	3	0,2	0,6
2.500	1	4	0,2	0,8
3.000	1	5	0,2	1

Ejemplo 3.3. Se dispone de las siguientes puntuaciones otorgadas a la calidad del servicio en un hotel por 20 familias, en las que 0 representa que se considera que la calidad del servicio es baja y 4 representa que es alta:

1 0 2 4 1
3 2 0 1 1
1 2 1 1 0
0 1 1 1 2

Ordenando los valores de menor a mayor y contando el número de veces que aparece cada valor de la variable construimos la tabla para la distribución:

x_i	n_i	N_i	f_i	F_i
0	4	4	0,20	0,20
1	10	14	0,50	0,70
2	4	18	0,20	0,90
3	1	19	0,05	0,95
4	1	20	0,05	1

Ejemplo 3.4. Un restaurante ha abierto sus puertas al público durante 25 días del mes ingresando las siguientes cantidades:

1.650	1.005	1.232	1.000	2.254
732	1.380	1.830	1.460	2.500
1.708	1.900	1.190	1.376	1.507
2.021	728	2.120	2.309	2.450
1.580	500	1.305	2.160	1.770

Dado que se dispone de un gran número de valores para la variable y todos tienen frecuencia unitaria, podemos agrupar los valores en intervalos a fin de simplificar la información. A la hora de elegir la amplitud para los intervalos de la distribución se suele utilizar el siguiente procedimiento:

1. Obtenemos el *rango* o *recorrido* de la distribución, es decir, la diferencia entre el mayor y el menor valor que toma la variable:

$$R = 2.500 - 500 = 2.000$$

2. Seguidamente dividimos el rango obtenido entre el número de intervalos que deseamos definir, siempre que el resultado de dicha división sea entero. Por ejemplo, podemos establecer 5 intervalos tal que la amplitud para cada uno de ellos será:

$$a_i = \frac{R}{k} = \frac{2.000}{5} = 400$$

3. Finalmente definimos los extremos de los intervalos:

$$L_0 = 500$$
$$L_1 = 500 + 400 = 900$$
$$L_2 = 900 + 400 = 1.300$$
$$L_3 = 1.300 + 400 = 1.700$$
$$L_4 = 1.700 + 400 = 2.100$$
$$L_5 = 2.100 + 400 = 2.500$$

En base a los resultados anteriores, construimos la tabla para la distribución de frecuencias:

$[L_{i-1}, L_i)$	x_i	n_i	N_i	f_i	F_i
[500, 900)	700	3	3	0,12	0,12
[900, 1.300)	1.100	4	7	0,16	0,28
[1.300, 1.700)	1.500	7	14	0,28	0,56
[1.700, 2.100)	1.900	5	19	0,20	0,76
[2.100, 2.500)	2.300	6	25	0,24	1

Donde obtener las frecuencias absolutas basta con contar el número de valores que pertenecen a cada intervalo. Por ejemplo entre 500 y 899 hay tres valores en la distribución: 500, 728 y 732.

3.3. REPRESENTACIÓN GRÁFICA DE LAS DISTRIBUCIONES

Las representaciones gráficas constituyen un conjunto de métodos mediante los cuales las observaciones estadísticas se representan mediante magnitudes o figuras geométricas. El objetivo fundamental de la representación gráfica es el de proporcionar de forma instantánea una visión global de los datos observados.

Por supuesto, el gráfico no debe considerarse en ningún caso un sustituto de la tabla estadística, sino un complemento, ya que la lectura de un gráfico, al basarse en impresiones visuales, resulta menos precisa que la de una tabla. En última instancia, siempre será el estudio analítico de los datos el que nos proporcionará las conclusiones definitivas acerca del fenómeno objeto de estudio.

En función de la naturaleza de los datos y de la forma en que éstos se presenten existen diferentes tipos de representaciones. A continuación se muestran los más utilizados para la presentación de datos turísticos.

3.3.1. Diagrama de Barras

Los diagramas de barras se utilizan generalmente para representar distribuciones de frecuencias con datos sin agrupar de variables discretas. Para construirlos, en el eje de abscisas posicionamos los diversos valores x_i que puede tomar la variable, y sobre cada uno de ellos levantamos un segmento o barra de altura igual a la frecuencia absoluta o relativa asociada a cada valor. Todas las barras tienen la misma base y sus áreas son proporcionales a las frecuencias absolutas o relativas.

Ejemplo 3.5. Utilizando los datos del Ejemplo 3.3 vamos a representar el diagrama de barras de la distribución:

x_i	0	1	2	3	4
n_i	4	10	4	1	1

3.3.2. Histograma

Los histogramas son un tipo especial de gráfico de barras que se utiliza normalmente cuando los datos están agrupados en intervalos. Para construir este tipo de gráficos debemos levantar un rectángulo cuya base sea igual a la amplitud de los intervalos, mientras que su altura será igual a las frecuencias asociadas a ellos siempre que los intervalos ***sean todos de la misma amplitud***. En caso contrario, cuando la amplitud de los intervalos es diferente, la altura de los rectángulos h_i debe determinarse aplicando la siguiente fórmula:

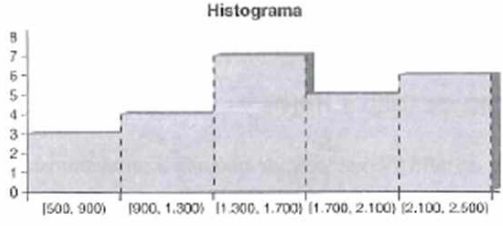
$$h_i = \frac{n_i}{a_i}$$

Donde a_i es la amplitud del intervalo correspondiente. Si no se realizara esta corrección, se produciría una distorsión óptica.

En cualquiera de los dos casos, el área de cada rectángulo es directamente proporcional a su frecuencia absoluta, n_i.

Ejemplo 3.6. Utilizando los datos del Ejemplo 3.4 vamos a elaborar el correspondiente histograma:

$[L_{i-1}, L_i)$	n_i
[500, 900)	3
[900, 1.300)	4
[1.300, 1.700)	7
[1.700, 2.100)	5
[2.100, 2.500)	6



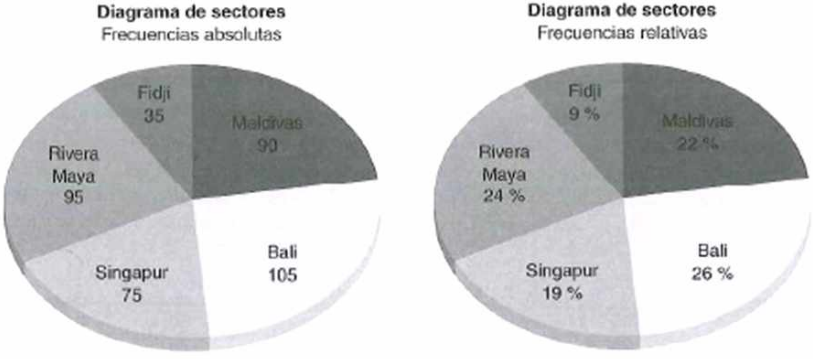
3.3.3. Diagrama de Sectores

En los diagramas de sectores se divide un círculo de tal forma que el área de cada porción es proporcional a la frecuencia absoluta o relativa de cada valor de la variable. Suelen utilizarse habitualmente para representar las distribuciones de atributos.

Ejemplo 3.7. Una agencia de viajes ofrece 5 viajes turísticos a diferentes destinos exóticos. A lo largo del año han vendido 90 viajes a las Maldivas, 105 a Bali, 75 a Singapur, 95 a Riviera Maya y 35 a Fidji. La distribución de frecuencias de los destinos turísticos será por tanto:

x_i	n_i	f_i
Maldivas	90	22%
Bali	105	26%
Singapur	75	19%
Riviera Maya	95	24%
Fidji	35	9%

Utilizando las frecuencias absolutas y relativas de esta distribución procedemos a realizar los correspondientes diagramas de sectores:



3.3.4. Diagrama de Tallo y Hojas

Los diagramas de tallo y hojas permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la hoja) del bloque de cifras restantes (que formará el tallo).

Ejemplo 3.9. Utilizando la siguiente distribución de edades de 20 visitantes de un parque de atracciones, vamos a construir su diagrama de tallo y hojas.

36 25 37 24 39 20 36 45 31 31

39 24 29 23 41 40 33 24 34 40

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

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

Tallos	Hojas
2	5 4 0 4 9 3 4
3	6 7 9 6 1 1 9 3 4
4	5 1 0 0

Finalmente reordenando las hojas, obtenemos el diagrama:

Tallos	Hojas
2	0 3 4 4 4 5 9
3	1 1 3 4 6 6 7 9 9
4	0 0 1 5

3.4. EJERCICIOS DE AUTOEVALUACIÓN

1. Para representar gráficamente una distribución de frecuencias de datos agrupados en intervalos, utilizaremos:

   a) Un diagrama de barras.
   b) Un histograma.
   c) Un diagrama de sectores.
   d) Un diagrama de tallo y hojas.

2. Se dispone de la siguiente distribución de frecuencias relativas acumuladas de la variable X = «Número de viajes vendidos» obtenida a partir de una muestra de 50 empleados de una agencia de viajes.

x_i	F_i
58	0,06
60	0,20
62	0,40
65	0,64
68	0,80
70	0,92
71	1,00

   ¿Cuántos empleados han logrado vender exactamente 62 viajes?:

   a) 20.
   b) 40.
   c) 14.
   d) 10.

3. Utilizando la distribución del ejercicio anterior, ¿qué porcentaje de empleados ha conseguido vender más de 60 viajes?:

   a) 40%.
   b) 94%.
   c) 80%.
   d) 64%.

4. Utilizando la distribución del ejercicio 2, ¿cuál es el número mínimo de viajes que debe haber vendido un empleado para estar entre los diez mejores?:

   a) 65.
   b) 68.
   c) 70.
   d) 71.

5. En un histograma, el área del rectángulo correspondiente al intervalo $[L_{i-1}, L_i)$ es proporcional a:

a) La marca de clase del intervalo.
b) La frecuencia absoluta acumulada del intervalo.
c) La frecuencia absoluta del intervalo.
d) El rango de la distribución.

6. La utilización de distribuciones de frecuencias con datos no agrupados es aconsejable cuando:

a) La variable toma pocos valores, pero se repiten un gran número de veces.
b) La variable toma muchos valores, pero se repiten pocas veces.
c) La variable toma pocos valores y además se repiten pocas veces.
d) La variable toma muchos valores y repiten muchas veces.

7. En una distribución de frecuencias, la proporción de observaciones con valores comprendidos entre los límites de un intervalo concreto se conoce con el nombre de:

a) Frecuencia absoluta.
b) Frecuencia relativa.
c) Frecuencia acumulada.
d) Frecuencia total.

8. Una variable X toma únicamente 4 valores distintos: x_1, x_2, x_3, x_4 (en orden creciente). En una muestra de tamaño 200 se observa que: el 45% de las observaciones toma el valor x_2; la proporción de observaciones en las que el valor de X es menor o como máximo igual a x_3 es 0,90 y 70 observaciones toman el valor x_1. ¿Cuál de las siguientes afirmaciones es verdadera?:

a) La frecuencia relativa de x_1 es 0,70.
b) La frecuencia absoluta de x_2 es 90.
c) La frecuencia absoluta de x_4 es 10.
d) La frecuencia relativa de x_3 es 0,20.

9. El número de observaciones correspondiente a un intervalo concreto se conoce con el nombre de:

a) Frecuencia total.
b) Frecuencia acumulada.
c) Frecuencia relativa.
d) Frecuencia absoluta.

10. Los diagramas de sectores se utilizan generalmente para representar distribuciones de frecuencias de:

a) Variables.
b) Atributos.
c) Variantes.
d) Datos agrupados en intervalos.

LECTURAS RECOMENDADAS

ALEGRE, J. et *al.* (2003). *Análisis Cuantitativo de la Actividad Turística*. Ed. Pirámide.

FERNÁNDEZ, C. (1993). *Manual de Estadística Descriptiva Aplicada al Sector Turístico*. Ed. Síntesis.

RAYA, J. M. (2004). *Estadística Aplicada al Turismo*. Ed. Pearson Prentice Hall.

RONQUILLO, A. (1997). *Estadística Aplicada al Sector Turístico. Técnicas cuantitativas y Cualitativas de Análisis Turístico*. Ed. CEURA.

SANTOS, J. et *al.* (2007). *Estadística para Estudios de Turismo*. Ediciones Académicas.

URIEL, E. y MUÑIZ, M. (1988). *Estadística Económica y Empresarial*. Editorial AC.

PALABRAS CLAVE

- Distribución de frecuencias
- Frecuencia absoluta
- Frecuencia relativa
- Datos no agrupados
- Datos agrupados en intervalos
- Diagrama de barras
- Histograma
- Diagrama de sectores
- Diagrama de tallo y hojas

CAPÍTULO 4

Medidas de posición, dispersión, forma y concentración

ESQUEMA

OBJETIVOS

Al finalizar el estudio de este capítulo, el alumno deberá ser capaz de:

1. Describir e interpretar los datos estadísticos desde una óptica descriptiva.
2. Calcular las diferentes medidas de posición, dispersión, forma y concentración presentadas a lo largo del capítulo.
3. Utilizar de manera adecuada las medidas estudiadas en la resolución de diversos problemas.

4.1. INTRODUCCIÓN

Como acabamos de ver en el capítulo anterior, todo análisis estadístico se inicia con una primera fase descriptiva de los datos, mediante la que se trata de organizar la información mediante la elaboración de tablas de frecuencias y representaciones gráficas. Una vez organizados los datos, el siguiente paso en el análisis es tratar de resumir toda la información contenida en las tablas de frecuencias a través de una serie de medidas que resuman toda esa información y que, de alguna forma, caracterizan a la distribución.

Dichas medidas, denominadas también ***estadísticos***, pueden ser de cuatro tipos:

1. Medidas de posición, las cuales sintetizan toda la información obtenida reduciéndola a un solo valor. Dentro de ellas podemos distinguir:
   - Medidas de posición central, en las que se hace referencia a un número «central» que se considera representativo de toda la muestra o población. Tal es el caso de la media aritmética, la media geométrica, la media armónica, la mediana y la moda.
   - Medidas de posición no central, que permiten conocer otros aspectos característicos de la distribución que no están relacionados con los valores centrales. Entre las medidas de posición no central más importantes están los cuantiles.
2. Medidas de dispersión, también llamadas de variabilidad, muestran la variabilidad de una distribución, indicando numéricamente si los diferentes valores de una variable están muy alejados con respecto a una medida de posición central, como puede ser la media aritmética de la distribución. Dentro de este grupo están el rango, el recorrido intercuartílico, la desviación media, la varianza, la desviación típica y el coeficiente de variación de Pearson.
3. Medidas de forma, las cuales permiten establecer una tipología de distribuciones comparando su representación gráfica con la de la distribución normal. Dentro de las medidas de forma, podemos distinguir a su vez medidas de asimetría y medidas de apuntamiento o curtosis.
4. Medidas de concentración, cuyo objetivo es cuantificar el grado de desigualdad en el reparto o distribución de una variable (generalmente de tipo económico como por ejemplo renta, beneficios, etc.), entre un número determinado de unidades (individuos, familias, empresas, etc.). Dentro de este grupo estudiaremos el índice de Gini y la curva de Lorenz.

Pasamos a continuación a ver en detalle el cálculo y las propiedades de cada una de estas medidas.

4.2. MEDIDAS DE POSICIÓN

4.2.1. Media Aritmética

La media aritmética de un conjunto finito de números se obtiene sumando de todas las observaciones y dividiendo el resultado por el tamaño de la muestra. La denotaremos por $\bar{x}$ y la fórmula para obtenerla es la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N}$$

En el caso de distribuciones de frecuencias con datos agrupados por intervalos, los valores x_i utilizados para el cálculo de la media aritmética serán las marcas de clase.

Es recomendable utilizar la media aritmética cuando los datos sean de naturaleza aditiva (rentas, salarios, beneficios, pesos, estaturas, puntos, etc.), de tal forma que su suma representa el total de los recursos repartidos entre todos los elementos de la distribución.

Ejemplo 4.1. Disponemos de los datos del número de días de estancia de 20 turistas:

Número de días (x_i)	Número de turistas (n_i)	$x_i n_i$
5	6	30
8	8	64
12	4	48
16	2	32
	20	174

La media aritmética será:

$$\bar{x} = \frac{5 \cdot 6 + 8 \cdot 8 + 12 \cdot 4 + 16 \cdot 2}{20} = 8{,}7 \text{ días}$$

Esto implica que por término medio los turistas incluidos en esta muestra han permanecido entre 8 y 9 días en el lugar de destino de sus vacaciones.

Ejemplo 4.2. Deseamos saber el peso medio embarcado por familia en los vuelos realizados en Semana Santa. Para ello, contamos con la siguiente distribución del peso en kilogramos del equipaje:

Peso (x_i)	Número de familias (n_i)	$x_i n_i$
54	2	108
59	3	177
63	4	252
64	1	64
	10	601

$$\bar{x} = \frac{54 \cdot 2 + 59 \cdot 3 + 63 \cdot 4 + 64 \cdot 1}{10} = 60{,}1 \text{ kg}$$

Sin embargo, en ocasiones puede que no todos los datos tengan la misma importancia para la investigación que estemos realizando por lo que puede resultar útil otorgar pesos o valores a los datos. En esos casos podemos utilizar la ***media aritmética ponderada***, $\bar{x}_w$, en la que cada valor de la variable x_i recibe una ponderación o peso independientemente de su frecuencia. Su cálculo se realiza mediante la siguiente fórmula:

$$\bar{x}_w = \frac{\sum_{i=1}^{k} w_i x_i n_i}{\sum_{i=1}^{k} w_i n_i} = \frac{w_1 x_1 n_1 + w_2 x_2 n_2 + \ldots + w_k x_k n_k}{w_1 n_1 + w_2 n_2 + \ldots + w_k n_k}$$

siendo w_i la ponderación de la variable x_i, y $\sum_{i=1}^{k} w_i$ la suma de todas las ponderaciones.

Ejemplo 4.3. Para superar la asignatura de estadística, un alumno debe ser evaluado en distintas pruebas referentes a la misma: test, problemas y práctica, cada una de ellas ponderada según su importancia o contribución en la nota final. Así, los pesos de cada prueba serán del 30%, 50% y 20% respectivamente. Sabiendo que las notas obtenidas por el alumno en cada prueba son 7, 3 y 5 respectivamente, ¿cuál es la nota global en la asignatura?

El primer paso que debemos realizar es construir la correspondiente tabla de frecuencias para agrupar toda la información:

Nota (x_i)	Número de exámenes (n_i)	Ponderación (w_i)	$x_i n_i$
7	1	0,3	2,1
3	1	0,5	1,5
5	1	0,2	1,0
	3	1	4,6

Si calculamos la media aritmética simple, tenemos que:

$$\bar{x} = \frac{7 + 3 + 5}{3} = 5 \text{ puntos}$$

En cambio, si calculamos la media ponderada, obtenemos:

$$\bar{x}_w = \frac{2{,}1 + 1{,}5 + 1{,}0}{0{,}3 + 0{,}5 + 0{,}2} = 4{,}6 \text{ puntos}$$

Como podemos observar el resultado es notablemente distinto, ya que el alumno pasa de aprobado a suspenso si se utiliza la media ponderada para calcular su nota media final, debido al mal resultado obtenido en la parte de problemas, que es la que más pondera.

Ejemplo 4.4. Para ocupar un puesto de trabajo vacante en una agencia de viajes se realizan diferentes pruebas a los aspirantes, cada una de ellas con una importancia determinada. El resultado de las pruebas por parte de dos aspirantes es la siguiente:

Prueba	Nota Aspirante 1 (x_i)	Nota Aspirante 2 (x_i)	Importancia (w_i)
Geografía	9	7	1
Contabilidad	6	5	3
Idiomas	7	10	6
Ofimática	10	4	2
			12

Para saber qué aspirante obtendrá el puesto de trabajo, calculamos las medias ponderadas para cada aspirante y, después, comparamos los resultados:

Aspirante 1			Aspirante 2		
Nota (x_i)	Importancia (w_i)	$w_i x_i n_i$	Nota (x_i)	Importancia (w_i)	$w_i x_i n_i$
9	1	9	7	1	7
6	3	18	5	3	15
7	6	42	10	6	60
10	2	20	4	2	8
	12	89		12	90

$$\bar{x}_w(\text{Aspirante 1}) = \frac{1 \cdot 9 \cdot 1 + 3 \cdot 6 \cdot 1 + 6 \cdot 7 \cdot 1 + 2 \cdot 10 \cdot 1}{1 \cdot 1 + 3 \cdot 1 + 6 \cdot 1 + 2 \cdot 1} = \frac{89}{12} = 7{,}4 \text{ puntos}$$

$$\bar{x}_w(\text{Aspirante 2}) = \frac{1 \cdot 7 \cdot 1 + 3 \cdot 5 \cdot 1 + 6 \cdot 10 \cdot 1 + 2 \cdot 4 \cdot 1}{1 \cdot 1 + 3 \cdot 1 + 6 \cdot 1 + 2 \cdot 1} = \frac{90}{12} = 7{,}5 \text{ puntos}$$

El Aspirante 2 logra hacerse con el puesto de trabajo si bien por muy poco. Sin duda, ha contribuido a lograrlo el hecho de que en la prueba que más pondera, la de idiomas, haya sacado un 10.

Propiedades de la media aritmética

La media aritmética presenta las siguientes propiedades:

- La suma de las desviaciones de los valores de la variable con respecto a la media aritmética es igual a cero.
- Si a todos los valores de la variable se les suma o resta una misma cantidad, la media aritmética queda aumentada o disminuida también en dicha cantidad (es lo que se conoce como *cambio de origen*).

- Si todos los valores de la variable se multiplican o dividen por una misma constante, la media aritmética queda multiplicada o dividida también por dicha constante (es lo que se conoce como *cambio de escala*).
- Si una variable Y es transformación lineal de otra variable X, tal que $Y = a + bX$, entonces la media aritmética de la variable Y sigue la misma transformación lineal con respecto a la media aritmética de la variable X, tal que:

$$\bar{y} = \frac{\sum y_i n_i}{N} = \frac{\sum (a + bx_i) n_i}{N} = \frac{\sum (an_i + bx_i n_i)}{N} = \frac{a \sum n_i}{N} + \frac{b \sum x_i n_i}{N} = a + b\bar{x}$$

Ventajas e inconvenientes de la media aritmética

Las ventajas que podemos destacar de esta medida de posición son:

- Su cálculo es muy sencillo.
- Se puede calcular en las variables de naturaleza cuantitativa.
- Para su cálculo se utilizan todos los valores de la distribución.
- Está perfectamente definida de forma objetiva y es única para cada distribución de frecuencias.
- Tiene un claro significado, ya que al ser el centro de gravedad de toda la distribución nos representa a todo el conjunto de valores observados.

Sin embargo, la media aritmética también presenta algunos inconvenientes:

- Es una medida de posición muy sensible a los valores extremos de la distribución, con lo que puede llegar a ser poco representativa del conjunto si la dispersión de los datos es muy elevada. Por ejemplo, si el nivel de gasto al día de cinco turistas es 28 €, 20 €, 18 €, 29 € y 108 €, el gasto medio es de 40,6 €. Observamos claramente que el valor 108 condiciona la obtención de una media poco representativa.
- No puede ser calculada cuando la variable es de tipo cualitativo.
- Podemos tener dificultades para su cálculo en distribuciones con intervalos abiertos.

4.2.2. Media Geométrica

La media geométrica, denotada por G, es una medida de posición central utilizada generalmente cuando los valores de la variable no son de naturaleza aditiva, sino acumulativa o con efectos multiplicativos. Tal es el caso de tipos de interés, porcentajes, tasas, números índices, etc.

En muchas ocasiones los valores de la distribución no son de naturaleza propiamente aditiva; tal es el caso de los números índice o de los porcentajes, los cuales representan la evolución de una característica con respecto al valor que tiene en un período o situación que llamamos base. Cuando se desea obtener promedios de magnitudes tales como tipos de interés, tasas, porcentajes, números índice, etc., la media aritmética pierde la propiedad de tener un claro significado, ya que la suma de dichas magnitudes no representa un total de recursos como en las magnitudes de naturaleza aditiva. En estos casos es aconsejable utilizar la media geométrica, pues se trata de la medida de posición central más representativa cuando la variable presenta variaciones acumulativas.

Su valor se obtiene como la raíz enésima del productorio[1] de los valores de la variable elevados a sus frecuencias respectivas. Matemáticamente:

$$G = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}} = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}} = (x_1^{n_1} \cdot x_2^{n_2} \cdot x_k^{n_k})^{\frac{1}{N}}$$

En ocasiones la complejidad de cálculo de la expresión anterior obliga a tomar logaritmos tal que:

$$\log G = \frac{1}{N} \log (x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}) = \frac{1}{N} (n_1 \log x_1 + n_2 \log x_2 + \ldots + n_k \log x_k) =$$

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

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

Aplicando antilogaritmos en la expresión anterior obtenemos el valor de la media geométrica:

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

[1] El símbolo productorio, denotado por una pi mayúscula (Π), es un operador matemático que representa la multiplicación finita o infinita de factores. Su funcionamiento es análogo al caso del sumatorio para la suma.

Ejemplo 4.5. Empleamos los datos del Ejemplo 4.1, relativos al número de días de estancia de 20 turistas, para obtener la media geométrica:

Número de días (x_i)	Número de turistas (n_i)
5	6
8	8
12	4
16	2
	20

Aplicando la fórmula anterior tenemos que:

$$G = \sqrt[20]{5^6 \cdot 8^8 \cdot 12^4 \cdot 16^2} = \sqrt[20]{1{,}3916 \cdot 10^{18}} = 8{,}0756$$

Valor ligeramente inferior al obtenido en el ejemplo anterior y que verifica la propiedad de que, para datos no negativos, siempre se cumple que $G \leq \overline{X}$.

Alternativamente podemos calcular la media geométrica aplicando logaritmos. Para ello calculamos dos columnas adicionales, $\log x_i$ y $n_i \log x_i$:

Número de días (x_i)	Número de turistas (n_i)	$\log x_i$	$n_i \log x_i$
5	6	0,6990	4,1938
8	8	0,9031	7,2247
12	4	1,0792	4,3167
16	2	1,2041	2,4082
	20		18,1435

El logaritmo de la media geométrica sería:

$$\log G = \frac{18{,}1435}{20} = 0{,}907175$$

Por tanto, tomando antilogaritmos tenemos que la media geométrica es:

$$G = \textit{anti}\log 0{,}907175 = 8{,}0756 \text{ días}$$

Resultado que, tal y como cabía esperar, coincide con el resultado obtenido aplicando la otra fórmula.

Ventajas e inconvenientes de la media geométrica

Las principales ventajas de la media geométrica son:

- Si su cálculo es posible, está definida de forma objetiva y es única.
- Tiene en cuenta en su cálculo a todos los valores de la distribución.
- Los valores extremos tienen menor influencia que en la media aritmética por estar definida a través de productos en vez de sumas.
- Es más representativa que la media aritmética cuando la variable evoluciona de forma acumulativa con efectos multiplicativos.

Sin embargo, la metodología de cálculo de esta media hace que presente ciertas desventajas:

- Su cálculo es más complicado que el de la media aritmética.
- Si algún valor de la variable es igual a cero, el resultado obtenido no es representativo al obtenerse una media geométrica nula.
- Asimismo, si la variable presentara valores negativos podría darse el caso de que no fuera posible calcularla, ya que se obtendrían soluciones imaginarias.

4.2.3. Media Armónica

Existen situaciones en las que no es adecuado utilizar la media aritmética ni la geométrica, ya que los datos observados no son de naturaleza aditiva ni multiplicativa. Esto ocurre en los casos en los que se desea promediar velocidades, rendimientos, productividades, etc., es decir, aquellos casos en los que la variable está medida en unidades relativas. Para este tipo de variables resulta más apropiado el uso de la media armónica.

La media armónica H es la inversa de la media aritmética de los inversos de los valores de la variable tal que:

$$H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}}$$

Ejemplo 4.6. Tomando los datos del Ejemplo 4.1 calculamos la media armónica:

Número de días (x_i)	Número de turistas (n_i)
5	6
8	8
12	4
16	2
	20

$$H = \frac{20}{\frac{6}{5} + \frac{8}{8} + \frac{4}{12} + \frac{2}{16}} = 7{,}52$$

El resultado obtenido es sensiblemente inferior al obtenido en el caso de la media aritmética y geométrica. En general, para una misma distribución de frecuencias con todos sus datos positivos y no nulos, se verifica siempre que $H \leq G \leq \overline{X}$.

Ejemplo 4.7. Un banco ha realizado para sus clientes a lo largo de la semana los siguientes cambios de euros a dólares estadounidenses:

Tipo de cambio (x_i)	Volumen (n_i)
1,3590	2.000
1,3570	1.500
1,3563	3.000
1,3577	2.100
1,3610	2.500
1,3630	2.300
1,3585	1.000
	14.400

La media armónica sería:

$$H = \frac{14.400}{\frac{2.000}{1,3590} + \frac{1.500}{1,3570} + \frac{3.000}{1,3563} + \frac{2.100}{1.3577} + \frac{2.500}{1,3610} + \frac{2.300}{1,3630} + \frac{1.000}{1,3585}} = 1,3590$$

El cambio medio aplicado por el banco es de 1,3590 dólares por euro.

Ventajas e inconvenientes de la media armónica

Las principales ventajas de la media armónica son:

- Está definida de forma objetiva y es única.
- Intervienen todos los valores de la distribución.
- Es más representativa que las otras medias en los casos de obtener promedios en velocidades, rendimientos y productividades.

Por su parte los inconvenientes de este tipo de media son:

- Si algún valor de la variable es nulo, no es posible calcular la media armónica.
- La presencia valores de la variable muy pequeños pueden provocar que sus inversos aumenten muchísimo haciendo despreciable frente a ellos la información de otros valores mayores de x_i.

4.2.4. Mediana

Las medidas de posición que hemos visto hasta el momento permiten representar al conjunto de valores observados de la distribución equilibrando los más elevados, los intermedios y los pequeños, ya que en su cómputo intervienen todos ellos. El problema que tienen estas medidas es que son sensibles a los valores extremos muy altos o muy bajos y cuando existe mucha dispersión son poco representativas del conjunto de observaciones.

Con objeto de superar estas dificultades vamos a definir otra medida de posición central denominada *mediana* en cuyo cálculo no intervienen todos los valores de la variable y en la que, en vez de equilibrar valores de la variable para determinar el centro de gravedad de la distribución, se equilibra las frecuencias observadas a ambos lados de su valor.

La mediana (Me) es aquel valor tal que, tras ordenar los valores de la variable en orden creciente, deja a su izquierda y a su derecha el mismo número de frecuencias.

Para el cálculo de la mediana debemos diferenciar si los valores se encuentran o no agrupados en intervalos:

A) Cálculo de la mediana en distribuciones con valores no agrupados en intervalos

En el caso de que los valores de la variable no estén agrupados en intervalos, el cálculo de la mediana se realiza de la siguiente manera:

- Si la distribución de frecuencias es unitaria:
  - En el caso de que el número de datos N sea impar, la mediana será el valor central de la distribución.
  - Por el contrario, si el número de datos N es par, existirán dos valores centrales por lo que la mediana será la media aritmética entre ellos.

Ejemplo 4.8. Las edades de menor a mayor de dos grupos de jóvenes que han realizado actividades de senderismo durante el fin de semana son las siguientes:

Grupo 1: 22, 23, 24, 25, 28
Grupo 2: 20, 21, 22, 24, 27, 29

En el Grupo 1 hay 5 personas ($N = 5$) por lo que la mediana es simplemente en este caso el valor central de la distribución, es decir, $Me_1 = 24$.

Sin embargo, en el caso del Grupo 2 tenemos 6 individuos por lo que N es par de tal forma que para calcular la mediana debemos tomar los dos valores centrales de la distribución y hallar su media: $Me_2 = \dfrac{22 + 24}{2} = 23$.

- Si la distribución de frecuencias no es unitaria, deberemos calcular el valor $N/2$ y compararlo con la columna de frecuencias absolutas acumuladas. Se observa

cuál es la primera frecuencia acumulada que supera o iguala $N/2$, distinguiéndose dos casos:

- Si $N/2$ coincide con algún valor de la columna de frecuencias absolutas acumuladas, entonces la mediana será la media aritmética entre el valor de la variable cuya frecuencia absoluta acumulada es $N/2$ y el siguiente valor de la variable.
- Si $N/2$ no coincide con ningún valor de la columna de frecuencias absolutas acumuladas, entonces la mediana será el primer valor de la variable cuya frecuencia absoluta acumulada sea superior a $N/2$.

Ejemplo 4.9. A continuación se muestran la distribución de los días de pernoctación (x_i) en dos hoteles de la Costa del Sol, así como sus respectivas frecuencias absolutas acumuladas (N_i).

Hotel 1		
x_i	n_i	N_i
1	3	3
2	4	7
5	9	16
7	10	26
10	6	32
	$N = 32$	

Hotel 2		
x_i	n_i	N_i
1	3	3
2	4	7
5	9	16
7	10	26
10	7	33
13	2	35
	$N = 35$	

En el caso del Hotel 1, tenemos que $N/2 = 16$, valor que podemos encontrar en la columna de frecuencias absolutas acumuladas. En este caso la mediana es la media aritmética entre el valor de la variable cuya frecuencia absoluta acumulada es $N/2 = 16$ y el siguiente valor de la variable, es decir, $Me_1 = \dfrac{5+7}{2} = 6$.

En el caso del Hotel 2, el valor $N/2$ es igual a 17,5, valor que no se encuentra en la columna de frecuencias absolutas acumuladas. En este caso la mediana coincide con el valor inmediato posterior al valor de la variable $N/2 = 17{,}5$, es decir, $Me_2 = 7$.

B) Cálculo de la mediana en distribuciones con valores agrupados en intervalos

En el caso de que los valores de la variable estén agrupados en intervalos, actuaremos de forma similar al caso anterior. Una vez determinado el intervalo cuya frecuencia absoluta acumulada es igual o mayor a $N/2$, el procedimiento para obtener el valor de la mediana es el siguiente:

- Si se cumple que $N_i = N/2$ coincide con algún valor de la columna de frecuencias absolutas acumuladas, entonces por convención el valor de la mediana será el extremo superior del intervalo que verifica dicha condición.
- Si no hubiera ningún valor N_i igual a $N/2$, entonces el intervalo que contiene a la mediana será el primer valor de la variable cuya frecuencia absoluta acumulada Ni sea mayor que $N/2$.

Si bien en el caso de que $N_i = N/2$ es posible establecer directamente un valor para la mediana, en el segundo tan sólo podremos determinar cuál es el intervalo que contiene a la mediana, por lo que debemos establecer un criterio para determinar su valor numérico. Si bien existen varios como, por ejemplo, utilizar la marca de clase del intervalo, el más extendido es el que resulta de aplicar la siguiente fórmula:

$$Me = L_i + \frac{\frac{N}{2} - N_{i-1}}{n_i} \cdot a_i$$

Donde L_i el extremo inferior del intervalo cuya frecuencia absoluta acumulada es superior a $N/2$, N_{i-1} es la frecuencia absoluta acumulada en el intervalo inmediatamente anterior y a_i es la amplitud del intervalo con frecuencia absoluta acumulada superior a $N/2$.

Ejemplo 4.10. La distribución por edades de una muestra de 900 turistas que han visitado el templo de Debod es la siguiente:

Intervalos de edades $[L_i, L_{i+1})$	**Número de turistas** (n_i)	**Número de Turistas acumulado** (N_i)
[20, 25)	100	100
[25, 30)	150	250
[30, 35)	200	450
[35, 40)	180	630
[40, 45)	270	900
	$N = 900$	

Para obtener la mediana de esta distribución calculamos $N/2 = 900/2 = 450$. En la columna de frecuencias absolutas acumuladas encontramos que en el intervalo [30, 35) se verifica que $N/2 = N_i = 450$, por lo que la mediana será el extremo superior del intervalo tal que $Me = 35$.

Ejemplo 4.11. Disponemos de la siguiente distribución de edades de una muestra de 797 turistas que han visitado la Sagrada Familia:

Intervalos de edades $[L_i, L_{i+1})$	**Número de turistas** (n_i)	**Número de Turistas acumulado** (N_i)
[0, 26)	211	211
[26, 35)	185	396
[35, 47)	201	597
[47, 75)	200	797
	$N = 797$	

Para calcular la mediana primero determinamos $N/2 = 797/2 = 398{,}5$. Ahora simplemente debemos buscar un valor igual o mayor que 398,5 en la columna de frecuencias absolutas acumuladas, lo que se verifica para el intervalo [35, 47).

Aplicando la expresión para obtener el valor concreto que toma la mediana dentro de este intervalo tenemos que:

$$Me = 35 + \frac{398{,}5 - 396}{201} \cdot 12 = 35{,}15$$

Por supuesto, en lugar de este criterio podríamos haber utilizado la marca de clase del intervalo en cuyo caso la mediana sería:

$$Me^* = \frac{35 + 47}{2} = 41$$

Valor sensiblemente distinto del obtenido aplicando la fórmula. En todo caso, cabe señalar que ambos valores pueden ser considerados válidos, debiendo ser el investigador el que determine cuál es el más adecuado en funciones de los objetivos que pretenda alcanzar.

Ventajas e inconvenientes de la mediana

Como ventajas de la mediana cabe señalar las siguientes:

- Se trata de la medida más representativa en el caso de variables que solo admiten la escala ordinal.
- Su interpretación y cálculo son sencillos.
- No es sensible a valores extremos de la variable ya que en su cálculo solo intervienen los valores centrales de la distribución.

El principal inconveniente que presenta la mediana es que en su determinación no intervienen todos los valores de la variable, por lo que no se utiliza toda la información presente en la distribución.

4.2.5. Moda

La moda (Mo) se define como el valor de la variable que más veces se repite, es decir, se trata de aquel valor de la variable que presenta la mayor frecuencia absoluta. Dado que la moda está definida en relación a valores de la variable asociados a sus distintas frecuencias con lo que no tiene sentido hablar de moda en las distribuciones de frecuencias de tipo unitario.

El cálculo de la moda depende de si los valores de la variable se encuentran o no agrupados en intervalos:

A) CÁLCULO DE LA MODA EN DISTRIBUCIONES CON VALORES NO AGRUPADOS EN INTERVALOS

En el caso de que los valores de la variable no estén agrupados en intervalos, la moda será simplemente aquel valor de la variable que presente la máxima frecuencia.

Ejemplo 4.12. Las siguientes tablas nos muestran el precio de diferentes viajes a Tánger y el número de viajes vendidos por dos agencias de viajes distintas, durante los meses de verano del último año:

Agencia de Viajes 1		Agencia de Viajes 2	
Precio (x_i)	**N.º de Viajes Vendidos (n_i)**	**Precio (x_i)**	**N.º de Viajes Vendidos (n_i)**
290	32	290	20
370	25	370	20
480	12	480	8
550	10	550	5
720	5	720	1

En el caso de la Agencia n.º 1, la moda es $Mo = 290$, al ser este valor el de mayor frecuencia ($n_1 = 32$). Sin embargo, en el caso de la Agencia n.º 2 nos encontramos con la circunstancia de que tenemos dos valores modales, $Mo_1 = 290$ y $Mo_2 = 370$, ya que ambos valores poseen la mayor frecuencia ($n_1 = n_2 = 20$). En este caso diremos que la distribución de frecuencias es bimodal. Del mismo modo, cuando haya tres modas en la distribución diremos que la distribución es trimodal y, en caso de que hubiera más de tres, multimodal.

En ocasiones también hablaremos de la ***moda relativa***, que se define como aquel valor (o valores) de la variable cuya frecuencia absoluta no es superada por la de sus valores contiguos.

Ejemplo 4.13. A continuación se recogen las calificaciones de 120 alumnos del grado de Turismo:

Nota (x_i)	N.º de alumnos (n_i)
1	20
3	30
4	20
5	40
7	7
9	3

Es evidente que en esta distribución la moda es $Mo = 5$ al ser este valor el que mayor frecuencia absoluta presenta ($n_4 = 40$). Sin embargo también existe una moda relativa para $x_2 = 3$ ya que las frecuencias absolutas de los valores anterior y posterior (20 en ambos casos) no superan a la frecuencia de x_2, que es 30.

B) Cálculo de la moda en distribuciones con valores agrupados en intervalos

En este caso no hablaremos de un valor para la moda sino de un intervalo modal, que será aquel que presente la mayor frecuencia, siempre que la amplitud de todos los intervalos sea la misma. En caso contrario, la moda será aquel intervalo que presente la mayor densidad de frecuencia d_i en relación a la amplitud del intervalo a_i, es decir, aquel que presente el mayor valor $d_i = \frac{n_i}{a_i}$.

No obstante, es posible asignar un valor puntual a la moda dentro del intervalo. Si bien algunos autores prefieren utilizar la marca de clase del intervalo para obtener un valor puntual, el convenio más extendido entre los estadísticos es realizar un prorrateo partiendo de la hipótesis de que existe cierta uniformidad en la distribución de valores dentro de cada intervalo. Para ello, suponiendo que todos los intervalos de la distribución tienen la misma amplitud aplicaremos la siguiente fórmula:

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

Donde L_i es el extremo inferior del intervalo con frecuencia absoluta, n_{i-1} y n_{i+1} son respectivamente las frecuencias absolutas del intervalo anterior y posterior al de mayor frecuencia absoluta y a_i es la amplitud del intervalo de mayor frecuencia.

En el caso de que la amplitud de cada intervalo fuera diferente, la fórmula anterior pasaría a ser:

$$Mo = L_i + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a_i$$

Donde L_i es el extremo inferior del intervalo con mayor densidad de frecuencia, a_i es la amplitud de dicho intervalo y d_{i-1} y d_{i+1} son respectivamente las densidades de frecuencia del intervalo anterior y posterior al de mayor densidad de frecuencia.

Ejemplo 4.14. Vamos a calcular la moda de la siguiente distribución de edades de una muestra de 119 visitantes del museo Guggenheim de Bilbao:

Intervalo $[L_i, L_{i+1})$	Frecuencia (n_i)
[10, 20)	11
[20, 30)	14
[30, 40)	21
[40, 50)	30
[50, 60)	18
[60, 70)	15
[70, 80)	7
[80, 90)	3
	119

La amplitud de todos los intervalos es la misma (10) por lo que el cálculo de la moda es tan sencillo como determinar cuál es el intervalo con mayor frecuencia absoluta, que es [40, 50) y utilizar la fórmula correspondiente:

$$Mo = 40 + \frac{9}{9+12} \cdot 10 = 44{,}29$$

La moda de esta distribución es 44,29, lo que indica que la mayor parte de las personas que visitan el Guggenheim tiene en torno a 44 años.

Ejemplo 4.15. Vamos a calcular la moda de la siguiente distribución de edades de una muestra de 791 turistas que han visitado Atenas:

Intervalo $[L_i, L_{i+1})$	N.º turistas (n_i)	Amplitud (a_i)	Densidad de frecuencia (d_i)
[0, 26)	180	26	6,92
[26, 35)	237	9	26,33
[35, 47)	180	12	15,00
[47, 75)	194	28	6,93
	791		

Dado que los intervalos presentan diferentes amplitudes, es necesario calcular una columna adicional con las densidades de frecuencias para cada intervalo. Analizando los valores obtenidos en dicha columna resulta evidente que el intervalo modal es el [26, 35) al ser este el que presenta la mayor densidad de frecuencia (26,33). Una vez determinado el intervalo debemos obtener un valor puntual para la moda; para ello aplicamos la fórmula anterior:

$$Mo = 26 + \frac{15{,}00}{6{,}92 + 15{,}00} \cdot 9 = 32{,}16$$

La moda de esta distribución es 32,16, es decir, la gente que visita Atenas generalmente tiene 32 años.

Ventajas e inconvenientes de la moda

La principal ventaja de la moda, aparte de su sencillez de cálculo y su fácil interpretación estadística, es que es posible calcularla tanto para variables cuantitativas como para variables cualitativas. Como principal inconveniente cabe señalar que en su determinación no intervienen todos los valores de la distribución, centrándonos sólo en la mayor frecuencia absoluta de un determinado valor de la variable o de la modalidad de los atributos.

Ejemplo 4.16. A continuación se muestra la distribución de una muestra de 481 viajeros que han utilizado los servicios de diferentes compañías aéreas en el aeropuerto de Barajas:

Compañía (x_i)	N.º de viajeros (n_i)
Air Europa	50
Air France	75
British Airways	35
EasyJet	62
Iberia	150
Ryanair	54
Spanair	55

La compañía aérea que presenta la mayor frecuencia es Iberia, por lo que dicha compañía es la moda de la distribución.

4.2.6. Cuantiles

Los cuantiles Q_i son aquellos valores que dividen la distribución en un cierto número de partes iguales, de manera que en cada una de ellas hay el mismo porcentaje de valores de la variable. Los cuantiles más utilizados son los cuartiles, los deciles y los percentiles:

- Los cuartiles C_i son tres valores que dividen la distribución en cuatro partes iguales (C_1, C_2 y C_3), por lo que dentro de cada uno está incluido el 25% de los valores. Los cuartiles corresponden pues al 25%, 50% y 75%.
- Los deciles D_i son nueve valores que dividen la distribución en diez partes iguales (D_1, D_2, ..., D_9), por lo que dentro de cada uno está incluido el 10% de los valores. Los deciles corresponden pues al 10%, 20%, ..., 90%.
- Los percentiles P_i son noventa y nueve valores que dividen la distribución en cien partes iguales (P_1, P_2, ..., P_{99}), por lo que dentro de cada uno está incluido el 1% de los valores. Los percentiles corresponden al 1%, 2%, 3%, ..., 99%.

Todos los cuantiles presentan una característica común y es que su valor intermedio coincide siempre con el de la mediana, de tal forma que para una misma distribución, el valor de la variable correspondiente al segundo cuartil, al quinto decil y al quincuagésimo percentil son el mismo y coinciden con el de la mediana.

El cálculo de los cuantiles será diferente dependiendo de si los datos de la distribución están agrupados en intervalos o no.

A) Cálculo de cuantiles en distribuciones con valores no agrupados en intervalos

El primer paso para calcular cuantiles en distribuciones con valores no agrupados es obtener los diferentes valores teóricos de las frecuencias acumuladas de cuantil.

Para ello utilizaremos la expresión $\frac{rN}{q}$, en la que r es el cuantil correspondiente, q es el número de intervalos con iguales frecuencias en que se divide la distribución utilizando dicho cuantil y N es el total de datos.

Seguidamente comparamos el resultado obtenido con la columna de frecuencias absolutas acumuladas de la distribución, procediendo de manera similar a como lo hacíamos en el caso de la mediana:

- Si el valor $\frac{rN}{q}$ coincide con algún valor de la columna de frecuencias acumuladas, entonces el valor del cuantil será igual a la media aritmética del valor de la variable cuya frecuencia absoluta acumulada es $\frac{rN}{q}$ y el siguiente valor de la variable.
- Por el contrario, si no existe ninguna frecuencia acumulada igual a $\frac{rN}{q}$, el valor del cuantil será el primer valor de la variable cuya frecuencia absoluta acumulada sea superior a $\frac{rN}{q}$.

Ejemplo 4.17. A continuación se presenta la distribución del número de pernoctaciones en una casa rural obtenida a través de una encuesta a 20 turistas. Utilizando estos datos vamos a calcular el primer y tercer cuartil, C_1 y C_3, el cuarto decil, D_4, y el nonagésimo percentil, P_{90}.

x_i	n_i	N_i
5	3	3
10	7	10
15	5	15
20	3	18
25	2	20
	$N = 20$	

- Primer cuartil, C_1

  Calculamos la frecuencia acumulada teórica correspondiente a este cuartil tal que $\frac{rN}{q} = \frac{1 \cdot 20}{4} = 5$.

  Seguidamente comparamos el valor obtenido con la columna de frecuencias absolutas de la distribución. En este caso, ningún valor de dicha columna es igual a la frecuencia teórica por lo que el valor del primer cuartil es $C_1 = 10$, es decir, el valor de la variable cuya frecuencia acumulada es mayor que $\frac{rN}{q}$.

- Tercer cuartil, C_3

  La frecuencia acumulada teórica correspondiente a este cuartil es $\frac{rN}{q} = \frac{3 \cdot 20}{4} = 15$. Dado que este valor coincide con la frecuencia absoluta acumulada para $x_3 = 15$, tenemos que el valor del tercer cuartil será la media aritmética de este valor y del siguiente valor de la variable, $x_4 = 20$, tal que $C_3 = \frac{15 + 20}{2} = 17{,}5$.

- Cuarto decil, D_4

  El valor teórico para el cuarto decil es $\frac{rN}{q} = \frac{4 \cdot 20}{10} = 8$.

  Dado que ningún valor de la columna de frecuencias absolutas es igual a la frecuencia teórica, el valor del cuarto decil es $D_4 = 10$, es decir, el valor de la variable cuya frecuencia acumulada es mayor que $\frac{rN}{q}$.

- Nonagésimo percentil, P_{90}

  En este caso tenemos que $\frac{rN}{q} = \frac{90 \cdot 20}{100} = 18$, valor que coincide con la frecuencia absoluta acumulada asociada a $x_4 = 20$, por lo que el valor del nonagésimo percentil es $P_{90} = \frac{20 + 25}{2} = 22{,}5$.

B) Cálculo de cuantiles en distribuciones con valores agrupados en intervalos

En el caso de distribuciones con valores agrupados en intervalos, el procedimiento es similar al utilizado para obtener el valor de la mediana. En primer lugar determinaremos el intervalo donde estará el cuantil aplicando el procedimiento que acabamos de ver en el apartado anterior. Una vez localizado el intervalo, utilizaremos la siguiente expresión para determinar el valor del cuantil:

$$Q = L_i + \frac{\frac{rN}{q} - N_{i-1}}{n_i} \cdot a_i$$

Donde L_i el extremo inferior del intervalo cuya frecuencia absoluta acumulada es superior a $\frac{rN}{q}$, N_{i-1} es la frecuencia absoluta acumulada en el intervalo inmediatamente anterior y a_i es la amplitud del intervalo con frecuencia absoluta acumulada superior a $\frac{rN}{q}$.

Ejemplo 4.18. La siguiente distribución muestra datos del número de habitaciones obtenidos a partir de 25 cuestionarios enviados a hoteles de cuatro estrellas en la provincia de Almería. Con estos datos vamos a calcular el segundo cuartil, el séptimo decil y el octogésimo tercer percentil:

$[L_i, L_{i+1})$	n_i	N_i
[0, 60)	5	5
[60, 200)	5	10
[200, 400)	9	19
[400, 600)	6	25
	$N = 25$	

- Segundo cuartil, C_2

  La frecuencia acumulada teórica para el segundo cuartil es $\frac{rN}{q} = \frac{2 \cdot 25}{4} = 12{,}5$.

  Ningún valor de la columna de frecuencias absolutas es igual a la frecuencia teórica el intervalo en el que estará situado el segundo cuartil será [200, 400), es decir, el primer intervalo cuya frecuencia acumulada es mayor que $\frac{rN}{q}$.

  Finalmente calculamos el valor numérico de C_2:

$$C_2 = 200 + \frac{12{,}5 - 10}{9} \cdot 200 = 255{,}56$$

- Séptimo decil, D_7

  En este caso tenemos que $\frac{rN}{q} = \frac{7 \cdot 25}{100} = 17{,}5$. Al no existir ningún valor de la columna de frecuencias absolutas igual a este valor, el séptimo decil estará situado en el intervalo [200, 400), siendo su valor:

$$D_7 = 200 + \frac{17{,}5 - 10}{9} \cdot 200 = 366{,}67$$

- Octogésimo tercer percentil, P_{83}

  Para el sexagésimo tercer percentil tenemos que $\frac{rN}{q} = \frac{83 \cdot 25}{100} = 20{,}75$. El intervalo cuya frecuencia absoluta acumulada supera este valor es [400, 600), por lo que el octogésimo tercer percentil estará contenido en este intervalo, siendo su valor:

$$P_{83} = 400 + \frac{20{,}75 - 19}{6} \cdot 200 = 458{,}33$$

4.3. MEDIDAS DE DISPERSIÓN

Mientras que las medidas de posición tienen como objetivo sintetizar los datos de una distribución en un único valor representativo, el objetivo de las medidas de dispersión es decirnos hasta qué punto las medidas de posición son realmente representativas de los datos. Para ello, mediante las medidas de dispersión podemos cuantificar la separación o la variabilidad de los valores de la distribución con respecto a un valor central. Cuanto mayor sea la dispersión, menos representativa de la distribución será la medida de posición.

Podemos clasificar las medidas de dispersión en:

- Absolutas, que son aquellas cuyo valor está expresado en las unidades de medida de la variable y que, por lo tanto, no son comparables entre diferentes distribuciones. Tal es el caso del rango, el recorrido intercuartílico, la desviación media, la varianza y la desviación típica.
- Relativas, cuyo resultado está expresado sin unidades de medida por lo que sirven para comparar la dispersión de distribuciones de frecuencias distintas. Un ejemplo de medida de dispersión relativa es el coeficiente de variación de Pearson.

4.3.1. Medidas de Dispersión Absoluta

4.3.1.1. Rango

El rango R, también denominado recorrido, se define como la diferencia entre el mayor y el menor valor de la variable, de tal forma que si ordenamos los valores de la variable de manera creciente tenemos que el rango se calcula como:

$$R = x_k - x_1$$

Sin duda, la ventaja fundamental de esta medida de dispersión es su sencillez de cálculo, pero presenta un claro inconveniente: se trata de una medida imprecisa, puesto que sólo tiene en cuenta el máximo y el mínimo de la distribución, sin tener en cuenta la frecuencia de cada valor.

Ejemplo 4.19. Calculamos el rango de la distribución de frecuencias de la variable «gasto por persona y día» de una muestra de 797 turistas entrevistados:

Máximo: 601,01 €

Mínimo: 0,32 €

$$R = x_k - x_1 = 601{,}01 - 0{,}32 = 600{,}69 \text{ €}$$

4.3.1.2. Recorrido Intercuartílico

El recorrido intercuartílico *RI* se define como la diferencia entre el tercer y el primer cuartil de la distribución tal que:

$$RI = C_3 - C_1$$

Esta medida evita la influencia que tienen en la dispersión los valores más extremos, ya que recoge el 50% central de las observaciones.

4.3.1.3. Desviación Absoluta

Para medir la representatividad de una determinada medida de posición P parece razonable utilizar como medida las distancias de todas las observaciones con respecto a ella, ya que cuanto más agrupados estén los valores en torno a una medida de posición, más representativa será ésta. Sin embargo, no podemos utilizar las distancias con su signo ya que al agregar todas se compensarían entre sí y no sería posible obtener una medida de la dispersión.

Para evitar este problema, una posible solución consiste en definir las distancias a la medida de posición en valor absoluto[2] y obtener un promedio de ellas, obteniendo así el estadístico denominado *desviación absoluta*, el cual denotaremos por D. En términos matemáticos podemos expresar D como:

$$D = \frac{\sum_{i=1}^{k} |x_i - P|\, n_i}{N}$$

Donde las barras verticales en el numerador (|...|) indican que las diferencias entre los diferentes valores de la variable X y la medida de posición P son consideradas en valor absoluto.

Si sustituimos P por medidas de posición concretas, obtenemos diferentes medidas de dispersión:

- Desviación absoluta respecto a la media ($D_{\bar{x}}$), que se define como la media de los valores absolutos de las desviaciones de los valores de la variable con respecto a la media aritmética de la distribución. Matemáticamente se expresa como:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}|\, n_i}{N}$$

Ejemplo 4.20. Durante una semana se hace valorar a los clientes de un hotel los servicios ofrecidos. Las puntuaciones varían de 1 a 5 (siendo 1 muy deficiente y 5 excelente). Los resultados obtenidos son los siguientes:

Puntuaciones	n_i	$x_i n_i$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert\, n_i$
1	20	20	2,6	52
2	30	60	1,6	48
3	100	300	0,6	60
4	120	480	0,4	48
5	80	400	1,4	112
	$N = 350$	1.260		320

[2] En matemáticas, el valor absoluto de un número real se define como su valor numérico sin tener en cuenta su signo. Así, por ejemplo, 3 es el valor absoluto tanto de 3 como de –3.

La media aritmética es:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{1.260}{350} = 3{,}6$$

Y la desviación absoluta con respecto a esta media:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{5} |x_i - \bar{x}|\, n_i}{N} = \frac{320}{350} = 0{,}91$$

- Desviación absoluta respecto a la mediana (D_{Me}), la cual viene dada por la siguiente expresión:

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

Ejemplo 4.21. Utilizando los datos del ejemplo anterior:

Puntuaciones	n_i	$x_i n_i$	$\lvert x_i - Me \rvert$	$\lvert x_i - Me \rvert\, n_i$
1	20	20	3	60
2	30	60	2	60
3	100	300	1	100
4	120	480	0	0
5	80	400	1	80
	$N = 350$	1.260		300

Para obtener la mediana, calculamos $N/2 = 350/2 = 175$. Al ser N par, $Me = 4$. La desviación media con respecto a la mediana es:

$$D_{Me} = \frac{\sum_{i=1}^{5} |x_i - Me|\, n_i}{N} = \frac{300}{350} = 0{,}86$$

4.3.1.4. Varianza

Otra alternativa para resolver el problema de compensación entre desviaciones de diferente signo que mencionábamos en el epígrafe anterior consiste en elevarlas al cuadrado, obteniendo así el estadístico denominado *desviación cuadrática*, D^2:

$$D^2 = \frac{\sum_{i=1}^{k} (x_i - P)^2\, n_i}{N}$$

A la desviación cuadrática respecto a la media se la denomina varianza. Se trata de una medida de dispersión muy utilizada en Estadística y cuyo resultado está expresado en las mismas unidades que los valores de la variable, pero elevadas al cuadrado. Matemáticamente la expresión de la varianza, que denotaremos por S^2 es:

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

Desarrollando el sumatorio del numerador, podemos obtener una expresión alternativa para la varianza que resultará más manejable a efectos prácticos:

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

$$= \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} + \bar{x}^2 \cdot \frac{N}{N} - 2\bar{x}^2 \cdot \bar{x} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} + \bar{x}^2 - 2\bar{x}^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2$$

Al igual que sucede con el resto de medidas de dispersión, cuanto más elevado sea su valor, más dispersión existirá, por lo que la media será menos representativa. Tiene la ventaja de que es una buena medida de dispersión cuando se ha utilizado la media como medida de posición. Sin embargo, presenta el inconveniente de que viene expresada en distinta unidad que la variable (la unidad de la variable al cuadrado).

La varianza está acotada inferiormente de tal forma que nunca puede ser negativa, al ser una suma de cuadrados, verificándose que $0 \leq S^2 \leq \infty$. En el caso particular $S^2 = 0$ estaremos ante una situación de nula dispersión lo que indicará que todos los valores de la variable son iguales a una constante.

4.3.1.5. Desviación Típica

La desviación típica se define como la raíz cuadrada de la varianza, tomando el resultado con signo positivo. Presenta la ventaja frente a la varianza de que su valor viene expresado en la misma unidad que la variable. Matemáticamente la desviación típica, denotada por S, se expresa como:

$$S = +\sqrt{S^2} = +\sqrt{\frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \, n_i}{N}}$$

Ejemplo 4.22. Utilizamos los datos del Ejemplo 4.20 para calcular la varianza y la desviación típica:

Puntuaciones	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
1	20	20	1	20
2	30	60	4	120
3	100	300	9	900
4	120	480	16	1.920
5	80	400	25	2.000
	$N = 350$	1.260		4.960

Realizamos el cálculo de la varianza con la fórmula transformada, por ser más sencilla:

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i n_i}{N} = \frac{1.260}{350} = 3{,}6$$

$$S^2 = \frac{\sum_{i=1}^{5} x_i^2 n_i}{N} - \bar{x}^2 = \frac{4.960}{350} - (3{,}6)^2 = 1{,}21$$

Calculando la raíz cuadrada de este resultado obtenemos la desviación típica tal que:

$$S = +\sqrt{1{,}21} = 1{,}1$$

4.3.2. Medidas de Dispersión Relativas

Como acabamos de ver, los valores de las medidas de dispersión absolutas dependen directamente de la unidad en la que esté medida la variable por lo que necesitamos recurrir a otro tipo de medidas de dispersión para poder establecer comparaciones entre diferentes niveles de dispersión en torno a un determinado parámetro de dos distribuciones distintas. Para ello utilizaremos medidas de dispersión relativas, cuyos valores se expresan adimensionalmente lo que posibilita la comparación de distribuciones medidas en distintas unidades. Dentro de estas medidas, la más utilizada en Estadística es el Coeficiente de Variación de Pearson, que pasamos a ver a continuación.

4.3.2.1. Coeficiente de Variación de Pearson

El coeficiente de variación de Pearson, representado por γ, se define como el cociente entre la desviación típica y el valor absoluto de la media tal que:

$$\gamma = \frac{S}{|\bar{X}|}$$

Podemos expresar el coeficiente en términos porcentuales simplemente multiplicando por 100 la expresión anterior tal que:

$$\gamma = \frac{S}{|\bar{X}|} \cdot 100$$

Esta medida nos da la dispersión en porcentaje, por lo que permite su interpretación de forma fácil y, además, facilita poder comparar la dispersión de varias distribuciones aunque las variables estén en diferentes unidades de medida.

El valor mínimo del coeficiente de variación de Pearson es cero, lo cual quiere decir que $S = 0$, es decir, todos los valores coinciden con la media y, por tanto, no hay dispersión. Por convención, se considera que la dispersión es óptima si γ es igual o menor al 30%. Por el contrario, si el valor de γ es superior al 50%, podemos considerar que la media es muy poco representativa.

El único problema que podemos encontrarnos a la hora de calcular este coeficiente es que la media de la variable analizada sea igual a cero, lo que impediría obtener un valor numérico para el coeficiente. En este caso, sería adecuado realizar un cambio de origen de la variable para solventar el problema.

Ejemplo 4.23. Siguiendo con los datos del Ejemplo 4.20, tenemos que:

$$\gamma = \frac{S}{|\bar{X}|} \cdot 100 = \frac{1{,}1}{3{,}6} \cdot 100 = 30{,}56\%$$

4.3.2.2. Tipificación de Variables

Otra posibilidad para facilitar la comparación de valores que pertenecen a diferentes distribuciones es la *tipificación de variables*, procedimiento mediante el cual podemos transformar cualquier variable en una nueva que denominaremos Z con media igual a cero y varianza igual a uno.

Para obtener los valores tipificados, simplemente basta con restar a cada valor de la variable la media de la distribución, y dividir el resultado por la desviación típica tal que:

$$z_i = \frac{x_i - \bar{x}}{S}$$

Ejemplo 4.22. Disponemos de la distribución del número de billetes de avión vendidos por cuatro agencias de viajes para la temporada de Semana Santa.

x_i	x_i^2
200	40.000
225	50.625
240	57.600
250	62.500
915	210.725

Calculamos la media y la desviación de la distribución:

$$\overline{x} = \frac{\sum_{i=1}^{4} x_i n_i}{N} = \frac{915}{4} = 228{,}75$$

$$S_X^2 = \frac{\sum_{i=1}^{4} x_i^2 n_i}{N} - \overline{x}^2 = \frac{210.725}{4} - (228{,}8)^2 = 354{,}69 \Rightarrow S_X = +\sqrt{354{,}69} = 18{,}83$$

Para hallar la distribución de la variable tipificada Z, restamos de cada valor x_i la media y después dividimos el resultado por la desviación tal que:

$$Z_1 = \frac{200 - 228{,}75}{18{,}83} = -1{,}53$$

$$Z_2 = \frac{225 - 228{,}75}{18{,}83} = -0{,}20$$

$$Z_3 = \frac{240 - 228{,}75}{18{,}83} = 0{,}60$$

$$Z_4 = \frac{250 - 28{,}75}{18{,}83} = 1{,}13$$

Por tanto la distribución de la nueva variable tipificada Z es:

z_i	z_i^2
−1,53	2,33
−0,20	0,04
0,60	0,36
1,13	1,27
0	4

Realizando los cálculos pertinentes podemos verificar que la media de esta distribución es igual a cero y la desviación igual a uno tal que:

$$\overline{Z} = \frac{0}{4} = 0$$

$$S_Z^2 = \frac{4}{4} - (0)^2 = 1 \Rightarrow S_Z = +\sqrt{1} = 1$$

Ejemplo 4.25. El número de viajes a Kenia vendidos por una agencia de viajes de Madrid la semana pasada es de 20. La media de viajes de este tipo vendidos por todas las agencias de Madrid durante el mismo periodo es de 25, con una desviación típica igual a 3.

Por otro lado, una agencia de viajes de Valencia ha vendido 14 viajes a Kenia, siendo la media de ventas de viajes a Kenia de las agencias de Valencia de 15 con una desviación igual a 1. ¿Qué agencia ocupa una mejor posición relativa?

Para resolver este problema, basta con tipificar las ventas de cada agencia a fin de poder compararlas tal que:

$$z_{Madrid} = \frac{20 - 25}{3} = -1{,}67$$

$$z_{Valencia} = \frac{14 - 15}{1} = -1$$

Los dos valores obtenidos son negativos, al ser los valores a comparar eran inferiores a la media en ambos casos. Dado que se trata de valores negativos, –1 es mayor que –1,67, por lo que la agencia de Valencia ha obtenido un mejor resultado en términos relativos.

4.4. MEDIDAS DE FORMA

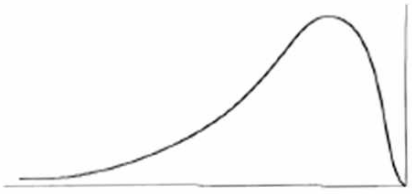
Cuando dos distribuciones coinciden en sus medidas de posición y dispersión, no tenemos datos analíticos para ver si son distintas, por lo que la única opción que nos queda es compararlas mediante su forma. Para ello podemos recurrir a diversas medidas estadísticas que nos van a permitir, sin necesidad de representar gráficamente la distribución, establecer su grado relativo de simetría así como de apuntamiento o curtosis.

4.4.1. Medidas de Asimetría

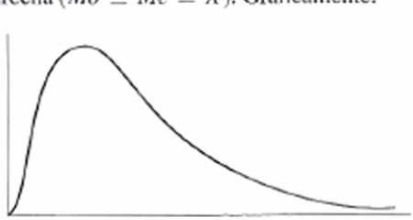
El objetivo de la medida de la asimetría es estudiar la deformación horizontal de los valores de la variable respecto al valor central de la media. En términos generales diremos que una distribución es simétrica cuando a la derecha y a la izquierda de su media existe el mismo número de valores, equidistantes dos a dos de la media, y además con la misma frecuencia. Matemáticamente, una distribución es simétrica siempre verifica que $\overline{X} = Me = Mo$.

En caso de que no satisfaga esta condición diremos que la distribución es asimétrica, pudiendo ser de dos tipos:

- Asimétrica a la izquierda ($Mo \geq Me \geq \overline{X}$). Gráficamente:



- Asimétrica a la derecha ($Mo \leq Me \leq \overline{X}$). Gráficamente:



Para medir el grado de asimetría de una distribución o compararlo con el de otra, podemos utilizar los dos coeficientes siguientes: el Coeficiente de Asimetría de Pearson y el Coeficiente de Asimetría de Fisher.

4.4.1.1. Coeficiente de Asimetría de Pearson

Este coeficiente, denotado por A_P se calcula como el cociente entre la diferencia de la media aritmética y la moda, y la desviación típica. Matemáticamente:

$$A_P = \frac{\overline{x} - Mo}{S}$$

El signo del resultado dependerá del numerador, ya que la desviación típica siempre es positiva. Su interpretación es la siguiente:

- Si $A_P > 0$, la distribución es asimétrica a la derecha.
- Si $A_P = 0$, la distribución es simétrica.
- Si $A_P < 0$, la distribución es asimétrica a la izquierda.

Si bien se trata de una medida muy sencilla de calcular, sólo puede utilizarse en el caso de que la distribución sea unimodal y campaniforme. Asimismo, al basarse sólo en la distancia entre media y moda, no es adecuado para medir asimetrías leves.

Ejemplo 4.26. La distribución de pernoctaciones de una muestra de 70 turistas con una edad igual o superior a 35 años es la siguiente:

Pernoctaciones (x_i)	Frecuencia (n_i)	$x_i n_i$	$x_i^2 n_i$
5	2	10	50
6	2	12	72
7	19	133	931
8	1	8	64
10	1	10	100
12	1	12	144
13	3	39	507
14	32	448	6.272
15	1	15	225
19	1	19	361
21	3	63	1.323
30	2	60	1.800
70	2	140	9.800
	$N = 70$	969	21.649

Calculamos la media aritmética, la moda y la desviación típica de esta distribución:

$$\bar{x} = \frac{\sum_{i=1}^{13} x_i n_i}{N} = \frac{969}{70} = 13{,}84$$

$$Mo = 14$$

$$S^2 = \frac{\sum_{i=1}^{13} x_i^2 n_i}{N} - \bar{x}^2 = \frac{21.649}{70} - (13{,}84)^2 = 117{,}73 \Rightarrow S_X = +\sqrt{117{,}73} = 10{,}85$$

Al ser la media aritmética 13,84, la moda 14 y la desviación típica 10,85, el coeficiente de asimetría de Pearson es:

$$A_P = \frac{\bar{x} - Mo}{S} = \frac{13{,}84 - 14}{10{,}85} = -0{,}015$$

El resultado indica que esta distribución es prácticamente simétrica, aunque presenta una ligera asimetría a la izquierda.

4.4.1.2. Coeficiente de Asimetría de Fisher

Si la distribución no fuera unimodal y campaniforme, debemos recurrir al coeficiente de asimetría de Fisher, cuya expresión matemática es:

$$g_1 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^3 \, n_i}{\frac{N}{S^3}}$$

Como la desviación típica es siempre positiva, el signo de g_1 dependerá del signo del numerador; al tener éste exponente impar, cada sumando del numerador será positivo o negativo según x_i esté por encima o por debajo de $\overline{x}$ pudiendo darse los siguientes resultados:

- Si las diferencias positivas superan a las negativas, el numerador será positivo por lo que $g_1 > 0$ y la distribución será asimétrica a la derecha.
- Por el contrario, si las diferencias negativas superan a las positivas el numerador será negativo, por lo que $g_1 < 0$, y la distribución será asimétrica a la izquierda.
- Finalmente si las diferencias positivas se compensaran con las negativas anulando el numerador tal que $g_1 = 0$, la distribución será simétrica.

Generalmente el coeficiente de Fisher es más preciso que el de Pearson, aunque su cálculo no resulte tan inmediato como el de éste.

Ejemplo 4.27. Con los datos del Ejemplo 4.24 planteamos el cálculo del coeficiente de Fisher. Para ello calculamos previamente las diferencias con respecto a la media elevadas al cubo multiplicadas por su respectiva frecuencia:

Pernoctaciones (x_i)	Frecuencia (n_i)	$(x_i - \overline{x})$	$(x_i - \overline{x})^3 \, n_i$
5	2	–8,84	–1.382,95
6	2	–7,84	–964,83
7	19	–6,84	–6.087,88
8	1	–5,84	–199,47
10	1	–3,84	–56,75
12	1	–1,84	–6,26
13	3	–0,84	–1,80
14	32	0,16	0,12
15	1	1,16	1,55
19	1	5,16	137,16
21	3	7,16	1.099,87
30	2	16,16	8.435,75
70	2	56,16	354.197,10
	$N = 70$		355.171,61

Sustituyendo los valores obtenidos, calculamos el coeficiente de asimetría de Fisher:

$$g_1 = \frac{\frac{355.171,61}{70}}{(10,85)^3} = 3,97$$

Como podemos ver, el resultado obtenido con el coeficiente de asimetría de Fisher es el opuesto al obtenido en el ejemplo anterior, presentando ahora la distribución una ligera asimetría a la derecha. Dado que la forma de la distribución no es ni mucho menos campaniforme (algo que se aprecia observando la tabla a simple vista, ya que la mayor parte de las frecuencias no se agrupan en el centro de la misma), daríamos por válido este último resultado.

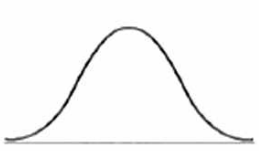
4.4.2. Medidas de Apuntamiento o Curtosis

Las medidas de apuntamiento o curtosis tratan de estudiar la distribución de frecuencias en la zona central de la distribución. Una mayor o menor concentración de frecuencias en torno a la media de la distribución dará lugar a una distribución más o menos apuntada.

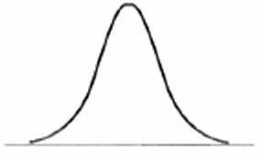
Para estudiar el apuntamiento hay que definir una distribución tipo que nos sirva de referencia. Esta distribución es conocida como la distribución Normal y se corresponde con numerosos fenómenos de la naturaleza. Su forma es la de una campana en donde la gran mayoría de los valores se encuentran concentrados alrededor de la media, siendo escasos los valores que están, en ambos extremos, muy distanciados de ésta.

Utilizando como referencia esta distribución normal es posible establecer una tipología de distribuciones en base a su apuntamiento:

- Una distribución es Mesocúrtica si la distribución de sus datos es la misma que la de la variable Normal.
- La distribución es Leptocúrtica si está más apuntada que la Normal.
- Si la distribución está menos apuntada que la Normal, entonces es Platicúrtica.



Distribución Mesocúrtica



Distribución Leptocúrtica

Distribución Platicúrtica

4.4.2.1. Coeficiente de Apuntamiento de Fisher

El coeficiente más utilizado para medir la curtosis es el coeficiente de apuntamiento de Fisher, el cual se calcula de acuerdo con la siguiente expresión:

$$g_2 = \frac{\dfrac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N}}{S^4} - 3$$

En función de los resultados obtenidos con este coeficiente, podemos encontrarnos en alguna de las siguientes situaciones:

- Si $g_2 > 0$, la distribución es leptocúrtica.
- Si $g_2 = 0$, la distribución es mesocúrtica.
- Si $g_2 < 0$, la distribución es platicúrtica.

Ejemplo 4.28. Continuando con los datos del Ejemplo 4.24, procedemos a calcular el coeficiente de curtosis de Fisher:

Pernoctaciones (x_i)	Frecuencia (n_i)	$(x_i - \bar{x})$	$(x_i - \bar{x})^4 n_i$
5	2	–8,84	12.229,27
6	2	–7,84	7.567,06
7	19	–6,84	41.658,49
8	1	–5,84	1.165,47
10	1	–3,84	218,08
12	1	–1,84	11,53
13	3	–0,84	1,51
14	32	0,16	0,01
15	1	1,16	1,79
19	1	5,16	707,35
21	3	7,16	7.871,91
30	2	16,16	136.297,61
70	2	56,16	19.890.697,42
	$N = 70$		20.098.427,52

$$g_2 = \frac{\dfrac{20.098.427,52}{70}}{(10,85)^4} = 20,72$$

Al ser el coeficiente de apuntamiento mayor de 3, la distribución es leptocúrtica.

4.5. MEDIDAS DE CONCENTRACIÓN

Las medidas de concentración tratan de poner de manifiesto el mayor o menor grado de igualdad en el reparto total de los valores de la variable[3]. Son por tanto, indicadores del grado de equidistribución de la variable; es por ello que este tipo de medidas proceden del campo de la Economía, aplicándose a distribuciones de rentas, salarios, etc.

La idea intuitiva sobre la que se basan la mayor parte de las medidas de concentración es la siguiente: sean k individuos cuyos valores de la variable (rentas, salarios, etc.) son $x_1, x_2, \ldots, x_k$, siendo $P = \sum_{i=1}^{k} x_i$ el dinero total repartido entre los individuos. En términos de concentración podemos encontrarnos con las siguientes situaciones extremas:

- Concentración máxima, o menor equidad en el reparto. En este caso, un solo individuo percibe el total y los demás nada. En este caso nos encontramos ante un reparto no equitativo tal que: $x_1 = x_2 = x_{k-1} = 0$ y $x_k = P$. El hecho de ser el individuo k el que más cobra (el único) se debe a que los x_i se expresan siempre de manera creciente.
- Concentración mínima o mayor equidad en el reparto. Ahora el conjunto total de valores de la variable está repartido por igual, es decir, se trata de una situación de reparto equitativo verificándose que $x_1 = x_2 = \ldots = x_k = \frac{P}{k}$.

Para medir la concentración se utilizan fundamentalmente dos medidas: la curva de Lorenz y el índice de Gini, que pasamos a ver a continuación.

4.5.1. Curva de Lorenz

Una forma de estudiar gráficamente la concentración es a través de la curva de Lorenz, la cual se construye representado en el eje de abscisas el porcentaje de frecuencias acumuladas y en el eje de ordenadas los porcentajes acumulados del total de la variable. Al unir los puntos resultantes obtenemos la curva, cuya forma nos permitirá determinar el nivel de concentración.

Para obtener la curva, en primer lugar debemos crear una tabla con las siguientes columnas:

- Una primera columna con los valores de la variable, x_i siendo $i = 1, 2, \ldots, k$.
- Una segunda columna con las frecuencias n_i de cada valor de la variable.
- Los productos de los valores de cada variable por su frecuencia, $x_i n_i$.
- Las frecuencias absolutas acumuladas N_i.

[3] Conviene tener en cuenta que si bien tienden a confundirse erróneamente, desde el punto de vista estadístico los términos dispersión y concentración no son opuestos por lo que es erróneo decir que una gran concentración equivale a una pequeña dispersión o al contrario.

- Los totales acumulados u_i, que se definen como la suma acumulativa de los productos de los valores de cada variable por su frecuencia tal que:

$$u_1 = x_1 n_1$$
$$u_2 = x_1 n_1 + x_2 n_2$$
$$\vdots$$
$$u_k = x_1 n_1 + x_2 n_2 + \ldots + x_k n_k = \sum_{i=1}^{k} x_i n_i$$

- La columna total de frecuencias acumuladas relativas p_i, que vendrá expresado en tanto por ciento por:

$$p_i = \frac{N_i}{N} \cdot 100$$

Por tanto, p_i es el porcentaje que representan los N_i primeros individuos sobre el total de individuos N.

- La columna de valores q_i que se define como:

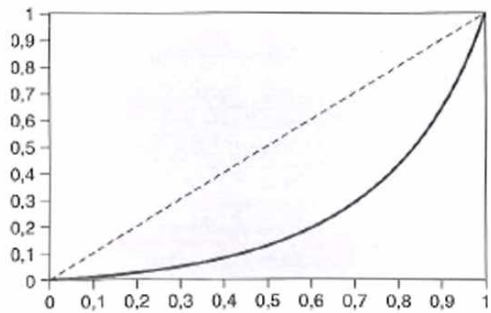
$$q_i = \frac{u_i}{u_k} \cdot 100$$

- Una última columna con las diferencias entre los valores p_i y q_i, es decir, $p_i - q_i$.

La tabla construida tendrá una forma como la siguiente:

x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{u_i}{u_k} \cdot 100$	$p_i - q_i$
x_1	n_1	$x_1 n_1$	N_1	u_1	p_1	q_1	$p_1 - q_1$
x_2	n_2	$x_2 n_2$	N_2	u_2	p_2	q_2	$p_2 - q_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$x_k n_k$	N_k	u_k	$p_k = 100$	$q_k = 100$	$p_k - q_k = 0$
	N	u_k					

Para representar la curva de Lorenz, basta con dibujar un cuadrado cuyos lados están divididos en una escala de 0% a 100%. Representando en el gráfico los pares de valores (p_i, q_i) y uniendo los puntos resultantes obtenemos la curva de Lorenz. Gráficamente su aspecto es el siguiente:



Generalmente la curva se representa junto con la diagonal del cuadrado, denominada *línea de equidad*. La curva de Lorenz siempre se sitúa por debajo de ella por verificarse que $p_i = q_i$ y estar los valores de la variable considerada ordenados de menor a mayor. Asimismo, la curva siempre es creciente (al ser p_i y q_i valores acumulados) y convexa.

La diagonal nos resultará útil para determinar el nivel de concentración de la distribución, pudiendo darse dos casos extremos:

- Concentración mínima: la curva coincide con la diagonal, verificándose por tanto que $p_i = q_i$ para todo i. Se trata de una situación de máxima equidad.
- Concentración máxima: la curva coincide con los lados del cuadrado, verificándose que $q_i = 0$ para $i = 1, 2, \ldots, k-1$ y $q_k = 100$. En este caso, no existe equidad alguna en el reparto.

4.5.2. Índice de Gini

El índice de Gini cuantifica el grado de aproximación existente entre la curva de Lorenz y la línea de equidad. Matemáticamente se expresa como:

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

Es importante resaltar que el sumatorio termina en $k-1$ porque en el numerador, aunque el sumatorio abarcase hasta k, sólo tendríamos $(k-1)$ sumandos, al ser $p_k = q_k = 100$, con lo que $p_k - q_k$ es siempre igual a cero.

La expresión anterior puede interpretarse en términos geométricos como el cociente entre el área comprendida entre la curva de Lorenz y la línea de equidad, y el área comprendida entre la línea de máxima concentración y la de equidad.

En función de los resultados que podemos obtener al calcular este índice, podemos distinguir dos casos extremos:

- Concentración mínima ($I_G = 0$): al verificarse que $p = q_i$, tenemos que

$$I_G = \frac{0}{\sum_{i=1}^{k-1} p_i} = 0$$

- Concentración máxima ($I_G = 1$): al verificarse que $q_i = 0$ para $i = 1, 2, \ldots, k-1$ y $q_k = 100$, obtenemos que

$$I_G = \frac{\sum_{i=1}^{k-1} p_i}{\sum_{i=1}^{k-1} p_i} = 1$$

Por tanto, el índice de Gini oscila entre 0 y 1. Cuanto más próximo esté su valor a cero, menor será la concentración, es decir, mayor equidad habrá en el reparto de la variable entre los individuos; por el contrario, cuanto más próximo esté a la unidad, mayor será la concentración.

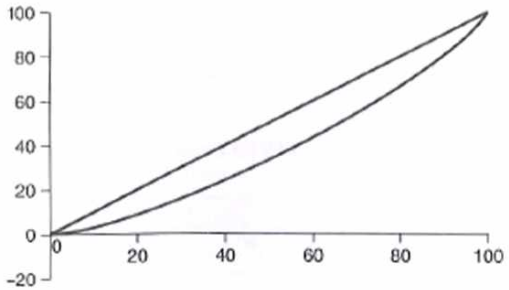
Ejemplo 4.27. Dada la siguiente distribución de ingresos medios diarios obtenidos en verano por 260 puestos de helados de la Comunidad Valenciana, calcular el índice de Gini y la curva de Lorenz.

$L_i - L_{i+1}$	x_i	n_i	$x_i n_i$	N_i	u_i	$p_i = \frac{N_i}{N} \cdot 100$	$q_i = \frac{u_i}{u_k} \cdot 100$	$p_i - q_i$
0-50	25	23	575	23	575	8,85	1,48	7,37
50-100	75	72	5.400	95	5.975	36,54	15,38	21,16
100-150	125	62	7.750	157	13.725	60,38	35,33	25,06
150-200	175	48	8.400	205	22.125	78,85	56,95	21,90
200-250	225	19	4.275	224	26.400	86,15	67,95	18,20
250-300	275	8	2.200	232	28.600	89,23	73,62	15,61
300-350	325	14	4.550	246	33.150	94,62	85,33	9,29
350-400	375	7	2.625	253	35.775	97,31	92,08	5,22
400-450	425	5	2.125	258	37.900	99,23	97,55	1,68
450-500	475	2	950	260	38.850	100,00	100,00	0,00
		$N = 260$	38.850			651,15		125,48

Sustituyendo los valores, calculamos el índice de Gini:

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{125,48}{651,15} = 0,193$$

El valor obtenido, cercano a cero, indica que existe poca concentración en los valores de la distribución, hecho que confirma la curva de Lorenz al estar cerca de la línea de equidad como puede apreciarse en el siguiente gráfico. El resultado obtenido indica que los ingresos obtenidos por los puestos de helados están distribuidos de manera equitativa en la Comunidad Valenciana.



4.6. EJERCICIOS DE AUTOEVALUACIÓN

1. Indique cuál de las siguientes afirmaciones es falsa:

   a) La media aritmética es una medida de posición central y en su cálculo interviene toda la información muestral.
   b) La mediana y la media aritmética nunca pueden coincidir, pero aportan información complementaria de la distribución.
   c) El valor de la media aritmética siempre está comprendido entre el menor y el mayor valor observado.
   d) La media aritmética es más sensible a los valores extremos o anómalos que la mediana.

2. La distribución de frecuencias de las líneas de autobús que utilizan un grupo de turistas en su desplazamientos visitando la ciudad se presenta a continuación:

N.º de días (x_i)	N.º de turistas (n_i)
7	12
33	8
54	5
60	16
67	4
68	3
74	25
75	2

   La medida de posición adecuada para resumir esta distribución es:

   a) La mediana, es decir, la línea número 60.
   b) La media aritmética, es decir, la línea número 54.
   c) La moda o la línea número 74.
   d) Ninguna de las anteriores.

3. Se han realizado las siguientes operaciones de cambio de dólares a euros:

Importe de la operación en dólares	Tipo de cambio dólares/euros
12.000	1,2000
23.000	1,1500
35.000	1,1700
9.760	1,2200

   El tipo medio de cambio en Euros/Dólar ha sido, aproximadamente:

   a) 0,9980.
   b) 0,8443.
   c) 1,1805.
   d) 0,8512.

4. Se sabe que el 40% de los viajeros que transporta una compañía aérea *low cost* realiza viajes dentro de la Península y que el beneficio medio por viajero es 30 euros por viaje. El resto de viajeros realiza viajes a Europa un beneficio medio de 25 euros por viaje. El beneficio medio de la compañía por viajero es:

   a) 27 euros.
   b) 28 euros.
   c) 26 euros.
   d) 29 euros.

5. De la encuesta realizada a los estudiantes de primer curso del grado de Turismo se han obtenido las siguientes medidas estadísticas de la variable «Gasto en ocio» diferenciando entre los que no trabajan y los que sí lo hacen.

Estadístico	No trabajan	Sí trabajan
Media	20,37	36,15
Mediana	20,00	35,00
Varianza	66,58	425,38
Desviación típica	8,16	21,27
Rango	25,00	80,00

   En base a estos resultados podemos afirmar que:

   a) La media de esta variable es más representativa en el colectivo que trabaja.
   b) La dispersión relativa de esta variable en el colectivo que trabaja es del 32,5%.
   c) La dispersión relativa de esta variable en el colectivo que no trabaja es del 40%.
   d) La relación entre los rangos nos permite concluir que la dispersión es más del doble en el colectivo que trabaja que en el que no trabaja.

6. La media aritmética y la varianza del ingreso medio por habitación en hoteles de cuatro estrellas en España son, respectivamente, 100 y 1.250. Si en la escala estandarizada la puntuación de la Comunidad Valenciana es –0,66, el ingreso medio por habitación en los hoteles de cuatro estrellas de esta comunidad es:

   a) 125,16.
   b) 123,33.
   c) 153,75.
   d) No se puede saber con esta información.

7. De 170 empresas del sector turístico se dispone de información sobre la distribución de la variable X = «Gasto realizado en publicidad (en miles de euros)» para el último ejercicio. A partir de la distribución de frecuencias de X se ha obtenido que su segundo decil es igual a 514, su 82º centil es 587 y su mediana es 549. ¿Cuál de las siguientes afirmaciones es falsa?:

a) El gasto máximo del 18% de las empresas que menos gastan en publicidad es 587 miles de euros.
b) El gasto máximo del 50% de las empresas que menos gastan en publicidad es 549 miles de euros.
c) El gasto máximo del 82% de las empresas que menos gastan en publicidad es 587 miles de euros.
d) El gasto máximo del 20% de las empresas que menos gastan en publicidad es 514 miles de euros.

8. Tenemos la información de los precios (agrupados en intervalos) y del número de vehículos alquilados en la isla de Tenerife en este último año. Si con respecto al año anterior los precios de los automóviles se han mantenido y los deciles 7, 8 y 9 se han desplazado considerablemente a la derecha, entonces podemos afirmar que en el último año:

a) Se ha alquilado un mayor porcentaje de coches de gama baja.
b) Se ha alquilado un mayor porcentaje de coches de gama media.
c) Se ha alquilado un mayor porcentaje de coches de gama alta.
d) Como los precios no han variado los deciles no pueden haber aumentado.

9. Un artículo de prensa afirmaba que «en las autopistas del Levante el 40% de los conductores supera la velocidad permitida, es decir, 120 km/h». El valor 120 en la distribución de frecuencias de la variable X = «Velocidad» es:

a) La media aritmética.
b) La mediana.
c) El cuarto decil.
d) El sexto decil.

10. En una muestra de 100 estudiantes se ha observado la variable X «Gasto anual en campings».

Gasto en campings	N.º de estudiantes
0-50	45
50-100	25
100-130	20
130-150	10

La media, la varianza y la desviación estándar son, respectivamente:

a) 70; 2162,5 y 46,5.
b) 67; 5572,5 y 74,65.
c) 70; 7062,5 y 84,04.
d) 67; 1803,5 y 42,47.

11. En tres secciones (A, B y C) se ha medido el tiempo en minutos que se tarda en empaquetar los productos, obteniendo las siguientes medias aritmética y varianzas.

Sección	Media aritmética	Varianza
A	12,3	25,45
B	13,4	28,85
C	10,3	24,45

Podemos decir que:

a) El tiempo que se tarda en empaquetar los productos presenta mayor dispersión relativa en la sección B.
b) La media de esta variable es más representativa en la sección C.
c) El tiempo que se tarda en empaquetar los productos presenta mayor dispersión relativa en la sección C.
d) El tiempo que se tarda en empaquetar los productos presenta mayor dispersión relativa en la sección A.

12. En un establecimiento de comida rápida situado en la playa se ha observado que sus 3 trabajadores preparan los bocadillos a una velocidad de: 80, 60 y 40 bocadillos por hora. Se contratan 2 empleados más con velocidad de 60 bocadillos por hora. Entonces podemos afirmar que:

a) La media de bocadillos por hora disminuye y la varianza sigue siendo la misma.
b) La media de bocadillos por hora y la varianza disminuyen.
c) La media de bocadillos por hora y la varianza aumentan.
d) La media de bocadillos por hora no varía y la varianza disminuye.

13. Indique cuál de las siguientes afirmaciones es falsa:

a) Al tipificar una distribución modificamos el origen y la escala de la variable.
b) Los cambios de origen no afectan al coeficiente de variación.
c) La media y la varianza de una distribución quedan afectadas por los cambios de unidades.
d) Si estandarizamos una distribución el número de orden que le corresponde a un elemento determinado no queda modificado.

14. Para fijar los precios de los tratamientos de belleza, un spa incrementa el coste de su personal en un 20% y le suma una cantidad fija de 100 euros. Si el precio medio de los tratamientos ha resultado ser de 280 euros, ¿cuáles será el coste medio de los tratamientos para el spa?

a) 150.
b) 436.
c) 156.
d) 250.

15. Se ha observado los niveles de ventas de tres empleados de una agencia de viajes a lo largo de 10 días, disponiéndose de los siguientes datos:

Empleado 1: $\sum_{i=1}^{10} x_i = 330$; $\sum_{i=1}^{10} (x_i - \bar{x})^2 = 2.262{,}00$

Empleado 2: $\sum_{i=1}^{10} x_i = 361$; $\sum_{i=1}^{10} x_i^2 = 14.853{,}26$

Empleado 3: $\sum_{i=1}^{10} x_i = 390$; $S_X^2 = 244{,}20$

¿Cuál de las siguientes afirmaciones es verdadera?

a) Las ventas medias del empleado 1 son mayores que las del 2 y son más homogéneas.
b) Las ventas medias del empleado 2 son mayores que las del 3 y son más homogéneas.
c) Las ventas medias del empleado 3 son mayores que las del 1, pero son menos homogéneas.
d) Las ventas medias del empleado 3 son mayores que las del 2, pero son menos homogéneas.

16. La distribución del número de horas que dedica un alumno a estudiar Estadística en el grado de Turismo es la siguiente:

Horas	N.º de estudiantes
1	10
2	12
3	15
4	8
5	6
6	9

Los valores para la media aritmética, la media geométrica, la media armónica, la mediana y la moda de esta distribución son respectivamente:

a) 3,95; 3,79; 3,23; 3,00; 3,00.
b) 3,25; 2,79; 2,33; 3,00; 3,00.
c) 3,00; 2,90; 2,33; 2,00; 3,00.
d) 2,25; 2,79; 2,00; 3,00; 6,00.

17. Dados los resultados estadísticos del siguiente cuadro:

Estadístico	Valor
Tamaño Muestral	81
Media	20,13
Mediana	25,65
Varianza	79,78
Asimetría	–3,93
Curtosis	–2,05
Rango	35,50

Podemos asegurar que la distribución de frecuencias es:

a) Asimétrica hacia la izquierda y platicúrtica.
b) Asimétrica hacia la derecha y platicúrtica.
c) Asimétrica hacia la izquierda y leptocúrtica.
d) Asimétrica hacia la derecha y mesocúrtica.

18. Se dispone de los siguientes datos, relativos a la distribución de los salarios mensuales de cuatro categorías profesionales:

Salarios (euros)	N.º de empleados
1.000	25
2.000	10
3.000	4
4.000	1

En relación a estos datos, ¿cuál de las siguientes afirmaciones es falsa?:

a) La curva de Lorenz está muy separada de la línea de equidad.
b) La curva de Lorenz está muy cerca de la línea de equidad.
c) El valor del índice de Gini está cerca de cero.
d) La distribución de los salarios es muy equitativa.

19. A continuación se presentan datos referidos a la variable X = «N.º de días de vacaciones en Agosto» para tres grupos de individuos, junto con sus frecuencias absolutas acumuladas:

Grupo A		Grupo B		Grupo C	
SMS	N_i	SMS	N_i	SMS	N_i
5	10	3	5	5	15
7	25	6	35	8	27
8	45	7	47	9	37
11	54	9	72	10	45
13	60	15	80	14	50

A la vista de los datos podemos concluir que:

a) La distribución del Grupo A es más homogénea que la del grupo B y presenta menor media aritmética.
b) La distribución del Grupo A es menos homogénea que la del grupo C y presenta mayor media aritmética.
c) La media aritmética de la distribución del Grupo A es inferior a la del grupo B y es más representativa.
d) La media aritmética de la distribución del Grupo B es inferior a la del grupo C y la distribución es más heterogénea en B que en C.

20. La línea aérea Iberspan tiene una media de retrasos de 30 minutos, con una desviación típica de 3, mientras que su competidora, Aerolia, presenta una media de 25 minutos de retraso en las salidas de sus vuelos, con una desviación típica de 5 minutos. ¿Cuál de las dos compañías presenta una mayor variabilidad en sus retrasos?

a) Iberspan.
b) Aerolia.
c) Las dos tienen la misma variabilidad.
d) No se puede saber.

LECTURAS RECOMENDADAS

ALEGRE, J. *et al.* (2003). *Análisis Cuantitativo de la Actividad Turística*. Ed. Pirámide.
FERNÁNDEZ, C. (1993). *Manual de Estadística Descriptiva Aplicada al Sector Turístico*. Ed. Síntesis.
RAYA, J. M. (2004). *Estadística Aplicada al Turismo*. Ed. Pearson Prentice Hall.
RONQUILLO, A. (1997). *Estadística Aplicada al Sector Turístico. Técnicas cuantitativas y Cualitativas de Análisis Turístico*. Ed. CEURA.
SANTOS, J. *et al.* (2007). *Estadística para Estudios de Turismo*. Ediciones Académicas.
URIEL, E. y MUÑIZ, M. (1988). *Estadística Económica y Empresarial*. Editorial AC.

PALABRAS CLAVE

- Media aritmética
- Media geométrica
- Media armónica
- Mediana
- Moda
- Cuantiles
- Dispersión
- Varianza
- Desviación típica
- Coeficiente de variación de Pearson
- Tipificación
- Asimetría
- Curtosis
- Concentración
- Curva de Lorenz
- Índice de Gini

CAPÍTULO 5

Distribuciones de frecuencias bidimensionales. Regresión y correlación

ESQUEMA

OBJETIVOS

Al finalizar el estudio de este capítulo, el alumno deberá ser capaz de:

1. Analizar el comportamiento simultáneo de dos variables.
2. Establecer relaciones de dependencia y causalidad entre variables.
3. Estimar relaciones lineales entre dos variables e interpretar el resultado obtenido.
4. Utilizar las estimaciones realizadas para realizar predicciones.

5.1. INTRODUCCIÓN

En capítulos anteriores hemos estudiado distribuciones unidimensionales en las que se analizaba una única característica para cada individuo. Sin embargo, ello no excluye la posibilidad de analizar simultáneamente varias características de los individuos. Concretamente en este capítulo nos dedicaremos al estudio simultáneo de dos características diferentes de los individuos que componen una población o muestra. Por ejemplo, supongamos que deseamos analizar qué factores influyen en el gasto en vacaciones que realizan los individuos varones mayores de 18 años de la Comunidad de Madrid. Además de la mencionada variable (gasto en vacaciones), seguramente también nos interese medir otras características que quizás podrían estar relacionadas con ella, tales como ingresos del individuo, edad, estado civil, medio de locomoción utilizado en las vacaciones, etc.

Todas estas características influirán en distinto grado en los niveles de gasto que realice cada individuo, permitiéndonos explicar en parte su comportamiento. A priori, cabe esperar que, por ejemplo, cuanto mayor sea su nivel de ingresos mayor gasto realizará también en sus vacaciones; o que cuanta menor sea su edad, menos gasto realice. Por supuesto, es posible realizar un estudio por separado de cada variable como hemos venido haciendo en los dos capítulos anteriores, pero el análisis conjunto de las variables nos va a permitir estudiar sus relaciones y dar respuesta a cuestiones tales como: ¿en qué medida el nivel de ingresos determina el gasto en vacaciones? ¿Gastan más los solteros o los casados? A lo largo del capítulo veremos las herramientas necesarias para dar respuesta a tales preguntas.

5.2. TABULACIÓN DE DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES

De la misma forma que hablamos de una distribución de frecuencias en el caso unidimensional, podemos introducir de manera similar el concepto análogo para el caso de una distribución de frecuencias bidimensional, si bien ahora las frecuencias absolutas hacen referencia al número de veces que un determinado par de datos aparece.

Consideremos el caso de una población compuesta por N individuos, donde cada uno de ellos presenta dos características, representadas mediante las variables X e Y. La variable X presenta r modalidades, $i = 1, 2, \ldots, r \Rightarrow x_i = x_1, x_2, \ldots, x_r$ mientras que la variable Y cuenta con s modalidades, $j = 1, 2, \ldots, s \Rightarrow y_j = y_1, y_2, \ldots, y_s$. Podemos disponer las observaciones en una tabla de doble entrada, denominada *tabla de correlación* cuando las variables son de tipo cuantitativo y *tabla de contingencia* cuando las variables son de tipo cualitativo, como la siguiente:

X/Y	y_1	y_2	...	y_s	$n_{i\cdot}$
x_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	$N = n_{\cdot\cdot}$

A partir de esta tabla podemos definir los siguientes conceptos:

- Frecuencia absoluta conjunta (n_{ij}): es el número de veces que se presenta conjuntamente el par (x_i, y_j). Por ejemplo, n_{12} indica el número de veces que se ha presentado x_1 conjuntamente con y_2. La suma de las frecuencias absolutas conjuntas es igual al número total de observaciones tal que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = N$$

- Frecuencia relativa conjunta (f_{ij}): Es el cociente entre la frecuencia absoluta conjunta, n_{ij} y la total, N, es decir

$$f_{ij} = \frac{n_{ij}}{N}$$

Lógicamente se verifica que

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}}{N} = \frac{N}{N} = 1$$

- Frecuencias absolutas marginales ($n_{i\cdot}$, $n_{\cdot j}$): en la última fila y en la última columna de la tabla se totalizan las frecuencias correspondientes a cada uno de los valores de las variables. En particular:

  - La frecuencia absoluta marginal del valor x_i, $n_{i\cdot}$, es el número de veces que se presenta el valor x_i con independencia de los valores de la variable Y, es decir,

$$n_{i\cdot} = n_{i1} + n_{i2} + \ldots + n_{is} = \sum_{j=1}^{s} n_{ij}$$

  - La frecuencia absoluta marginal del valor y_j, $n_{\cdot j}$, representa el número de veces que se presenta el valor y_j con independencia de los valores de la variable X, es decir,

$$n_{\cdot j} = n_{1j} + n_{2j} + \ldots + n_{rj} = \sum_{i=1}^{r} n_{ij}$$

En ambos casos se verifica que:

$$\sum_{i=1}^{r} n_{i\cdot} = n_{1\cdot} + n_{2\cdot} + \ldots + n_{r\cdot} = N$$

$$\sum_{j=1}^{s} n_{\cdot j} = n_{\cdot 1} + n_{\cdot 2} + \ldots + n_{\cdot s} = N$$

- Frecuencias relativas marginales ($f_{i\cdot}$, $f_{\cdot j}$): es el cociente entre la frecuencia absoluta marginal y la frecuencia total, es decir:

$$f_{i\cdot} = \frac{n_{i\cdot}}{N} \qquad f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

Cabe señalar que cuando la distribución tenga pocas observaciones, aunque la tabla de correlación sigue siendo válida, resulta más cómodo tabular directamente los datos en columnas de la siguiente forma:

x_i	x_j	n_i
x_1	y_1	n_1
x_2	y_2	n_2
...	...	...
x_r	y_s	n_r
		N

En el caso de que las frecuencias fueran unitarias, incluso podríamos prescindir de la última columna:

x_i	x_j
x_1	y_1
x_2	y_2
...	...
x_r	y_s

Ejemplo 5.1. Se dispone de los siguientes datos de 100 turistas relativos al gasto realizado durante su estancia en Granada (X) y su edad (Y):

X/Y	**25**	**30**	**35**	**40**	$n_{i\cdot}$
500	15	11	18	0	44
600	12	14	0	12	38
700	0	3	7	8	18
$n_{\cdot j}$	27	28	25	20	$N = 100$

Procedemos a interpretar la tabla:

- Podemos observar por ejemplo que hay 14 individuos que gastan 600 euros en su estancia y cuya edad es de 30 años; este número es, pues, la frecuencia del par de características (600, 30). También puede observarse que no existe ningún individuo que gaste 500 euros y que tenga 40 años. El resto de datos contenidos en la tabla se interpretaría de manera análoga.
- La suma de todas las frecuencias absolutas conjuntas es igual al tamaño muestral, es decir, 100.

- La frecuencia conjunta relativa del par (600, 30) sería

$$f_{22} = \frac{n_{22}}{N} = \frac{14}{100} = 0{,}14$$

Si sumáramos todas las frecuencias conjuntas relativas se verificaría que

$$\sum_{i=1}^{3}\sum_{j=1}^{4}\frac{n_{ij}}{N} = \frac{100}{100} = 1$$

- Si analizamos la tabla por columnas o por filas obtenemos las correspondientes frecuencias absolutas marginales. Por ejemplo, de los 100 turistas entrevistados vemos que 44 de ellos han gastado 500 euros (resultado obtenido sumando las frecuencias de la primera fila); 38 personas se han gastado 600 euros y 18, 700 euros. Del mismo modo, vemos que 27 turistas tienen 25 años (valor que se obtiene sumando las frecuencias de la primera columna); 28, tienen la edad de 30 años; 25 tienen 35 años, y 20 tienen 40 años. Lógicamente la suma de las frecuencias absolutas marginales para cada variable debe coincidir con el tamaño total de la muestra, verificándose en efecto que

$$44 + 38 + 18 = 27 + 28 + 25 + 20 = 100$$

- Por último, si queremos obtener las frecuencias relativas marginales, bastará simplemente con dividir su frecuencia absoluta marginal por el tamaño de la muestra. Así, la frecuencia relativa marginal para $x_2 = 600$ sería

$$f_{\cdot 2} = \frac{38}{100} = 0{,}38$$

mientras cue para $y_2 = 30$, tendríamos que

$$f_{\cdot 2} = \frac{28}{100} = 0{,}28$$

5.3. DISTRIBUCIONES MARGINALES Y CONDICIONADAS

5.3.1. Distribuciones Marginales

A partir de las frecuencias marginales absolutas definidas en el epígrafe anterior podemos obtener lo que se conoce como distribuciones marginales para cada variable que compone la distribución bidimensional, mediante las cuales vamos a poder examinar el comportamiento individual de cada una de ellas. No obstante, debemos tener en cuenta que el proceso inverso (reconstruir una distribución bidimensional a partir de dos distribuciones marginales) en general no es posible, ya que considerando las variables de manera aislada, podrían no mantener ninguna relación de homogeneidad entre sí o, aún manteniéndola, desconocer exactamente cuál es su forma.

Expresadas en columnas, las distribuciones marginales de frecuencias tendría el siguiente formato:

X		Y	
x_i	$n_{i\cdot}$	y_j	$n_{\cdot j}$
x_1	$n_{1\cdot}$	y_1	$n_{\cdot 1}$
x_2	$n_{2\cdot}$	y_2	$n_{\cdot 2}$
...	...	...	...
x_r	$n_{r\cdot}$	y_s	$n_{\cdot s}$
	N		N

Al ser las distribuciones marginales de tipo unidimensional, podemos calcular las medidas de posición, dispersión forma y concentración que vimos en el capítulo anterior. Por ejemplo, podemos definir las medias marginales y varianzas marginales para las variables X e Y de la siguiente forma:

$$\overline{x} = \frac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N}$$

$$\overline{y} = \frac{\sum_{j=1}^{s} y_j n_{\cdot j}}{N}$$

$$S_X^2 = \frac{\sum_{i=1}^{r} (x_i - \overline{x})^2 n_{i\cdot}}{N} = \frac{\sum_{i=1}^{r} x_i^2 n_{i\cdot}}{N} - \overline{x}^2$$

$$S_Y^2 = \frac{\sum_{j=1}^{s} (y_j - \overline{y})^2 n_{\cdot j}}{N} = \frac{\sum_{j=1}^{s} y_j^2 n_{\cdot j}}{N} - \overline{y}^2$$

Ejemplo 5.2. Utilizando los datos del ejemplo anterior, obtenemos la distribución de frecuencias marginales de X e Y:

X		Y	
x_i	$n_{i\cdot}$	y_j	$n_{\cdot j}$
500	44	25	27
600	38	30	28
700	18	35	25
		40	20
	100		100

A partir de esta tabla podemos calcular las medias marginales:

$$\bar{x} = \frac{\sum_{i=1}^{3} x_i n_{i\cdot}}{N} = \frac{500 \cdot 44 + 600 \cdot 38 + 700 \cdot 18}{100} = 574$$

$$\bar{y} = \frac{\sum_{j=1}^{4} y_i n_{\cdot j}}{N} = \frac{25 \cdot 27 + 30 \cdot 28 + 35 \cdot 25 + 40 \cdot 20}{100} = 31{,}90$$

Por su parte, las varianzas marginales son:

$$S_X^2 = \frac{\sum_{i=1}^{3} x_i^2 n_{i\cdot}}{N} - \bar{x}^2 = \frac{500^2 \cdot 44 + 600^2 \cdot 38 + 700^2 \cdot 18}{100} - (574)^2 = 5.524$$

$$S_Y^2 = \frac{\sum_{j=1}^{4} y_j^2 n_{\cdot j}}{N} - \bar{y}^2 = \frac{25^2 \cdot 27 + 30^2 \cdot 28 + 35^2 \cdot 25 + 40^2 \cdot 20}{100} - (31{,}90)^2 = 29{,}39$$

La obtención de las desviaciones típicas marginales resulta inmediata a partir de las varianzas:

$$S_X = +\sqrt{5.524} = 74{,}32$$
$$S_Y = +\sqrt{29{,}39} = 5{,}42$$

5.3.2. Distribuciones Condicionadas

En ocasiones el investigador puede estar interesado en el comportamiento de una de las variables sujeto a un determinado valor de la otra. Para ello, podemos recurrir a las distribuciones condicionadas, en las que el conjunto de valores que toma una de las variables está delimitado por el valor que toma la otra. Por ejemplo, supongamos que

X está condicionada a que Y tome el valor y_3; en ese caso, la tabla correspondiente de la distribución condicionada sería:

$x_i/Y = y_3$	$n_i/Y = y_3$
x_1	n_{13}
x_2	n_{23}
...	...
x_r	n_{r3}
	$n_{\cdot 3}$

Obsérvese que ahora la frecuencia total de esta distribución no es N sino $n_{\cdot 3}$, pues partimos de la condición de que Y toma el valor y_3.

Del mismo modo, podríamos condicionar Y a que X tome el valor x_2, de tal forma que ahora la tabla tendría la siguiente estructura:

$y_i/X = x_2$	$n_i/X = x_2$
y_1	n_{21}
y_2	n_{22}
...	...
y_s	n_{2s}
	$n_{2\cdot}$

De manera análoga al ejemplo anterior, dado que partimos de la condición de que X toma el valor x_2, la frecuencia total de esta distribución es $n_{2\cdot}$.

En general, la forma de la distribución de X condicionada a $Y = y_j$ y de la distribución de Y condicionada a $X = x_i$ será la siguiente:

X		Y	
$x_i/Y = y_j$	n_{ij}	$y_i/X = x_i$	n_{ji}
x_1	n_{1j}	y_1	n_{j1}
x_2	n_{2j}	y_2	n_{j2}
...	...	...	...
x_r	n_{rj}	y_s	n_{js}
	$n_{\cdot j}$		$n_{j\cdot}$

Por su parte, las frecuencias relativas condicionadas se definen como:

$$f_{ij} = \frac{n_{ij}}{n_{\cdot j}} = \frac{\frac{n_{ij}}{N}}{\frac{n_{\cdot j}}{N}} = \frac{f_{ij}}{f_{\cdot j}} \qquad f_{ji} = \frac{n_{ji}}{n_{j\cdot}} = \frac{\frac{n_{ji}}{N}}{\frac{n_{j\cdot}}{N}} = \frac{f_{ji}}{f_{j\cdot}}$$

Por supuesto, dado el carácter univariante de las distribuciones condicionadas también es posible calcular las diferentes medidas de posición, dispersión, forma y concentración que ya vimos en el capítulo anterior.

Ejemplo 5.3. Utilizando los datos del ejemplo 5.1, vamos a obtener algunas distribuciones condicionadas. Por ejemplo, la distribución de las edades de los turistas condicionada a que el gasto sea igual a 600 será:

$y_i/X = 600$	$n_i/X = 600$
25	12
30	14
35	0
40	12
	38

Mientras que la distribución del gasto condicionada a que los turistas tengan 35 años es:

$x_i/Y = 35$	$n_i/Y = 35$
500	18
600	0
700	7
	25

5.4. DEPENDENCIA ESTADÍSTICA

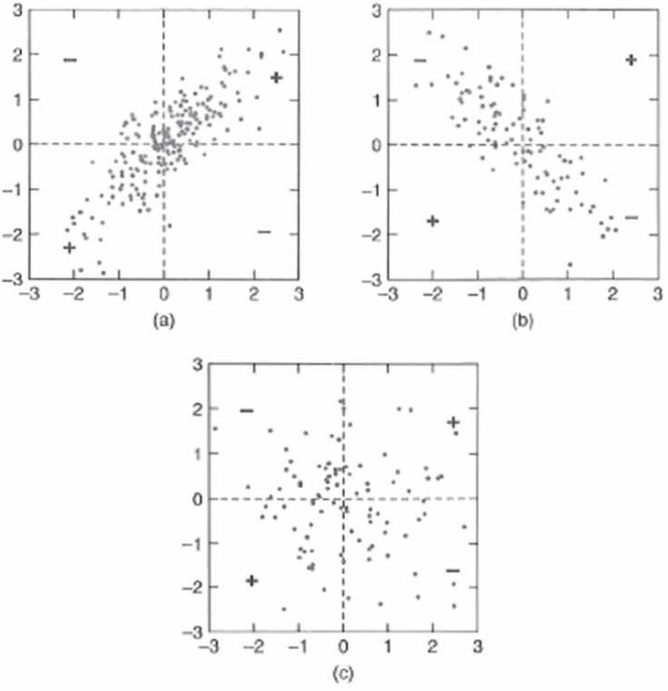
Una de las cuestiones que mayor interés ofrece para el investigador dentro de los fenómenos representados mediante distribuciones de carácter bidimensional es la de conocer el grado de relación existente entre las variables que la componen. El estudio de dicha relación se puede plantear desde dos puntos de vista diferentes:

- Por un lado, podemos estudiar la intensidad y el signo de la relación existente entre dos variables, lo que se conoce con el nombre de *correlación*.
- Por otro lado, podemos tratar de explicar el comportamiento de una variable (denominada dependiente, endógena o explicada) a partir del comportamiento de otra variable (denominada independiente, exógena o explicativa) a través de un modelo matemático que describe la relación entre las variables; es lo que se denomina *regresión*, sobre la que hablaremos en el siguiente epígrafe.

En numerosas ocasiones resulta sencillo establecer la presencia de una relación de dependencia entre dos variables simplemente estudiando su diagrama de dispersión de dos variables. Recordemos que un diagrama de dispersión es un tipo de diagrama que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.

Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal y el valor de la otra variable determinado por la posición en el eje vertical.

Podemos ver algunos ejemplos de diagramas de dispersión en la siguiente figura:



A la vista de los gráficos anteriores, podemos establecer el sentido de la relación que existe entre las dos variables:

- En el gráfico (a) la relación entre las variables es *directa*, ya que a medida que aumentan los valores de X lo hacen los de Y y viceversa.
- En el gráfico (b), por el contrario, se dice que la relación es *inversa* ya un aumento en el valor de la variable X implica una reducción en el valor de la variable Y y viceversa.
- Finalmente en el gráfico (c) no parece existir *a priori* una relación muy evidente entre las variables.

5.4.1. Covarianza y Correlación

Una medida que nos va a permitir conocer el signo de la relación existente entre dos variables es la ***covarianza***, denotada por S_{XY} y cuya expresión es:

$$S_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \overline{x})(y_j - \overline{y})n_{ij}}{N}$$

Alternativamente, utilizando un procedimiento similar a la que vimos en el caso de la fórmula de la varianza, podemos simplificar la fórmula anterior tal que:

$$S_{XY} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{s}x_i y_j n_{ij}}{N} - \overline{x}\,\overline{y}$$

En función del signo de la covarianza, la interpretación de la relación entre las variables será la siguiente:

- Si $S_{XY} > 0$, diremos que hay dependencia directa o positiva, es decir, las variables varían en el mismo sentido. Éste sería el caso del gráfico (a) que vimos anteriormente.
- Si $S_{XY} < 0$, entonces la dependencia será inversa o negativa, de tal forma que las variables varían en sentido opuesto. Un ejemplo de distribución con covarianza negativa lo tenemos en el gráfico (b) de la figura anterior.

a) Propiedades de la covarianza

La covarianza posee las siguientes propiedades:

1. Se verifica siempre que $S_{XY} = S_{YX}$.
2. Si a todos los valores de la variable X les sumamos una constante a y a todos los valores de la variable Y les sumamos una constante b, el valor de la covarianza no varía.
3. Si multiplicamos todos los valores de la variable X por una constante a y a todos los valores de la variable Y los multiplicamos por una constante b, la covarianza queda multiplicada por el producto de las constantes a y b.
4. Si aplicamos transformaciones lineales a las variables X e Y tal que $Z = a + bX$ y $T = c + dY$, la covarianza entre las variables transformadas T y Z se relaciona con la anterior de la forma: $S_{ZT} = bdS_{XY}$.

Como consecuencia de estas propiedades, podemos ver que el principal inconveniente de la covarianza es su dependencia de las unidades, al ser su valor sensible a los cambios de escala; asimismo, su valor no está acotado, lo que impide la comparación

entre las covarianzas de diferentes distribuciones. Por todo ello, en la práctica la utilidad de la covarianza se limita a establecer el sentido de la relación entre las variables, si bien en el caso de que ambas variables estén medidas en las mismas unidades, podemos afirmar que cuanto mayor sea la covarianza, mayor será su relación de dependencia.

Ejemplo 5.4. A partir de los datos del ejemplo 5.1 calculamos la covarianza. Para ello transformamos la tabla de doble entrada en otra más sencilla que nos permitirá realizar los cálculos con mayor agilidad, combinando cada uno de los valores de la variable X con todos los de la variable Y:

x_i	y_j	n_{ij}	$x_i y_j n_{ij}$
500	25	15	187.500
500	30	11	165.000
500	35	18	315.000
500	40	0	0
600	25	12	180.000
600	30	14	252.000
600	35	0	0
600	40	12	288.000
700	25	0	0
700	30	3	63.000
700	35	7	171.500
700	40	8	224.000
		100	1.846.000

Dado que en el ejemplo 5.2 ya habíamos calculado la media de ambas variables:

$$\bar{x} = 574$$

$$\bar{y} = 31{,}90$$

Disponemos de toda la información necesaria para calcular la covarianza entre ambas variables:

$$S_{XY} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{4} x_i y_j n_{ij}}{N} - \bar{x}\,\bar{y} = \frac{1.846.000}{100} - (574 \cdot 31{,}90) = 149{,}40$$

El resultado presenta signo negativo lo que nos indica que cuanto mayor sea el gasto turístico en Granada, mayor edad tendrá probablemente el turista y vice-versa.

Debido a los inconvenientes que presenta la covarianza, resulta necesario definir una nueva medida que no se vea afectada por cambios en las unidades de medida y que nos permita determinar de manera objetiva la intensidad de la relación; ello se logra dividiendo el valor de la covarianza por el producto de las desviaciones típicas de las

variables X e Y obteniendo así el ***coeficiente de correlación lineal de Pearson***, el cual se denota por r_{XY} y cuya expresión es:

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

El valor de este coeficiente es adimensional y siempre está comprendido entre –1 y +1. Su interpretación es la siguiente:

- Si el coeficiente de correlación lineal es positivo ($r_{XY} > 0$), la relación entre las variables será directa. La intensidad de dicha relación será tanto más fuerte cuanto más se aproxime a +1, siendo aceptables niveles a partir de +0,75.
- Por el contrario, si el coeficiente de correlación lineal es negativo ($r_{XY} < 0$), la relación entre las variables es inversa. La intensidad de dicha relación será tanto más fuerte cuanto más se aproxime a –1, considerándose aceptables valores menores de –0,75.
- Finalmente, si $r_{XY} = 0$ diremos que no existe correlación lineal entre las variables, si bien ello no impide que pueda existir otras formas de correlación (parabólica, exponencial, etc.)

En todo caso, debe tenerse en cuenta que, si bien viene expresado en términos numéricos, este coeficiente posee carácter cualitativo, lo que significa que si, por ejemplo, obtenemos un coeficiente $r_{XY} = 0{,}2$, y con otras variables diferentes obtenemos $r_{X'Y'} = 0{,}6$, podemos afirmar que en el segundo caso la intensidad de la relación es mayor que en el primero, pero no que sea tres veces superior.

b) Propiedades del coeficiente de correlación

Podemos destacar las siguientes propiedades del coeficiente de correlación de Pearson:

1. Si multiplicamos todos los valores de una de las variables por una constante a, el valor del coeficiente de correlación no varía, alterando únicamente su signo si $a < 0$.
2. Si existe una relación exacta entre ambas variables tal que $Y = a + bX$ entonces se cumple que $r_{XY} = 1$ si $b > 0$ y $r_{XY} = -1$ si $b < 0$.
3. Si dos variables son independientes, se verifica que $r_{XY} = 0$; sin embargo, su recíproco no siempre es cierto.

Ejemplo 5.5. Continuando con los datos del ejemplo anterior tenemos que las desviaciones típicas de las variables son:

$$S_X = 74{,}32$$

$$S_Y = 5{,}42$$

Por lo que el valor del coeficiente de correlación lineal es:

$$r_{XY} = \frac{149{,}40}{74{,}32 \cdot 5{,}42} = 0{,}3709$$

Lo que nos indica que, si bien existe una relación directa entre el gasto de los turistas y su edad, su intensidad no es elevada.

5.5. REGRESIÓN LINEAL

Tal y como acabamos de ver, utilizando el coeficiente de correlación lineal podemos determinar el signo y la intensidad de la relación entre dos variables. Sin embargo, dicho coeficiente no nos permite decir nada acerca del sentido de la relación entre ellas, es decir, no es posible determinar una relación de ***causalidad*** de una variable respecto a otra.

Para comprender mejor el concepto de causalidad, veamos un ejemplo ilustrativo: supongamos que disponemos de los datos anuales de temperatura y el número de matrimonios que se han celebrado en una localidad. Si calculamos el coeficiente de correlación entre ambas variables, es muy probable que obtengamos un valor muy elevado. Sin embargo, resulta evidente que las altas temperaturas no causan los matrimonios; mucho menos lógico parece pensar que un aumento en el número de matrimonios implique unas temperaturas más cálidas, por lo que es muy probable que la elevada correlación obtenida se deba a que los matrimonios tienden a producirse en verano debido a otros motivos, tales como una mayor disponibilidad de tiempo libre, celebración de banquetes al aire libre, etc. Por tanto, parece evidente que no existe una relación causal directa entre ambas variables. En general, a este tipo de correlaciones se las denomina ***espurias*** y su origen suele hallarse en otra variable que presenta una relación de dependencia con las variables observadas. En otros casos en los que no sea posible establecer una relación indirecta con una tercera variable, simplemente diremos que la correlación observada es fruto de la ***casualidad***.

Por el contrario, si consideramos el caso de una agencia de viajes que reduce sus precios y que, posteriormente, aumenta su nivel de ventas es evidente que existe una alta probabilidad de que dicho aumento se deba a la rebaja de precios como consecuencia de la ley de oferta y demanda. En este caso, sí estaríamos ante una relación de tipo causal.

En conclusión, el hecho de que dos variables estén estadísticamente relacionadas no implica necesariamente que una sea causa de la otra. Para poder concluir que X causa a Y se deben cumplir, al menos, tres condiciones:

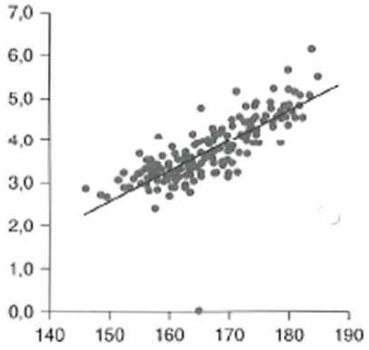
1. X debe preceder a Y.
2. Y no debe ocurrir cuando X no ocurre.
3. Y debe ocurrir cada vez que X ocurra.

Esta reflexión previa acerca de la causalidad resulta necesaria cuando se aborda el tema de la regresión, ya que a la hora de establecer la formulación del modelo que

relaciona a ambas variables deberemos haber especificado previamente cuál es la variable que actúa como causa y cuál es la que representa el efecto en base a algún modelo teórico.

Una vez establecida la relación de causalidad, debemos seleccionar la forma del modelo matemático que relaciona a las variables. Si bien ésta puede de ser de muy diversos tipos (exponencial, logarítmica, parabólica, polinómica, etc.), en el presente capítulo nos vamos a centrar en el estudio del caso en el que una línea recta es la función que mejor describe la dependencia entre las variables.

No obstante, antes de pasar a describir el procedimiento para obtener la recta que mejor se ajusta a nuestros datos, conviene explicar de manera intuitiva cuál es el significado de la recta de regresión y el objetivo que se persigue con su obtención. Supongamos que a la vista del diagrama de dispersión consideramos que la función que mejor se ajusta a la nube de puntos es la de una recta, tal y como se muestra en la siguiente figura. Una vez obtenida la ecuación de dicha recta, el objetivo que perseguimos al disponer de una recta que se ajusta bien a nuestros datos, es el de poder realizar predicciones de la variable dependiente a partir de valores predeterminados de la variable independiente.



Pasamos a continuación a examinar en detalle el proceso mediante el que se determina la ecuación de la recta que mejor se ajusta a la nube de puntos.

5.5.1. El Método de Mínimos Cuadrados

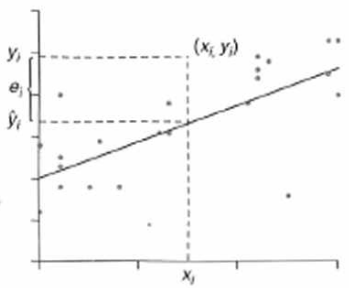
La forma más sencilla de escribir la ecuación de una recta es la siguiente:

$$Y = a + bX$$

En consecuencia, nuestro objetivo será encontrar los valores óptimos para los parámetros de la recta, a y b, que reciben el nombre de ordenada en el origen y pendiente respectivamente.

Si bien existen diferentes criterios para determinar su valor (por ejemplo, podríamos elegir la recta que pase por los dos puntos más alejados entre sí o utilizar aquella recta que pasa por el centro de gravedad de la distribución), el criterio de Mínimos Cuadrados es el utilizado habitualmente debido a que produce una recta con buenas propiedades estadísticas y permite obtener el valor de los parámetros mediante expresiones sencillas. Aplicando dicho criterio, se minimiza la suma de los cuadrados de las diferencias entre el valor observado para la variable dependiente y el obtenido al sustituir en la ecuación de la recta el valor de la variable independiente.

Veamos gráficamente en qué consiste el método de ajuste mediante Mínimos Cuadrados. En la siguiente figura se muestra un gráfico de dispersión junto con su correspondiente recta de regresión. En ella podemos observar que a la pareja de datos (x_i, y_i) podemos hacerle corresponder sobre la recta otro punto cuyas coordenadas son $(x_i, \widehat{y_i})$, siendo $\widehat{y_i}$ la predicción que la recta nos da para la variable Y sustituyendo el valor x_i en la recta.



Entre esta predicción y el valor observado realmente para Y existe una diferencia que denominamos *residuo* o *error*, el cual denotamos por e_i tal que:

$$e_i = y_i - \widehat{y_i}$$

El objetivo del método de Mínimos Cuadrados consiste en encontrar valores para los parámetros de la recta, a y b, tales que la suma de residuos al cuadrado SC_e sea mínima. Es decir, se trata de minimizar la siguiente expresión:

$$SC_e = \sum_{i=1}^{k} e_i^2 = \sum_{i=1}^{k} (y_i, \widehat{y_i})^2 = \sum_{i=1}^{k} (y_i - (a + bx_i))^2$$

Utilizando métodos de optimización se obtiene que, de todas las rectas de la forma $Y = a + bX$, la que minimiza SC_e es aquella que cumple que:

$$b = \frac{S_{XY}}{S_X^2}$$

$$a = \overline{y} - b\overline{x} = \overline{y} - \frac{S_{XY}}{S_X^2} \cdot \overline{x}$$

La solución anterior es válida para el caso en el que estemos calculando la *recta de regresión de Y sobre X*, es decir, cuando consideramos a Y como la variable dependiente y a X como la independiente.

Si en lugar de ello tomáramos a Y como variable independiente y a X como variable dependiente, estaríamos calculando la recta de regresión que minimiza errores con respecto a X, es decir, *la recta de regresión de X sobre Y*. En este caso la recta sería ahora:

$$X = a + bY$$

Y la solución óptima para los parámetros de la recta es ahora:

$$b = \frac{S_{XY}}{S_Y^2}$$

$$a = \overline{x} - b\overline{y} = \overline{x} - \frac{S_{XY}}{S_Y^2} \cdot \overline{y}$$

En ambos casos, dado que las varianzas son positivas por definición, el signo de las pendientes de cada recta, $\frac{S_{XY}}{S_X^2}$ y $\frac{S_{XY}}{S_Y^2}$, será el mismo que el de la covarianza, por lo lo que las rectas serán crecientes o decrecientes, dependiendo de si la covarianza es positiva o negativa, respectivamente.

Ejemplo 5.6. Se dispone de la siguiente distribución bidimensional de frecuencia unitaria, en la que la variable X representa la renta anual de 5 individuos e Y representa el gasto que realiza cada uno de ellos al año en vacaciones; ambas variables están expresadas en miles de euros.

x_i	y_i
28,00	0,56
32,00	0,81
36,70	1,28
39,00	1,88
40,00	2,88

Utilizando estos datos vamos a obtener la recta de regresión de Y sobre X, es decir, supondremos que la relación causal más lógica es pensar que cuanta más renta tiene un individuo, mayor gasto realizará en vacaciones. Por tanto la forma de la recta será $Y = a + bX$.

En base a la tabla anterior realizamos algunos cálculos que nos resultarán de utilidad para obtener la varianza de X y la covarianza entre X e Y:

x_i	y_j	x_i^2	$x_i y_i$
28,00	0,56	784,00	15,68
32,00	0,81	1.024,00	25,92
36,70	1,28	1.346,89	46,98
39,00	1,88	1.521,00	73,32
40,00	2,88	1.600,00	115,2
		6.275,89	277,10

Por otro lado, las medias marginales de X e Y son:

$$\overline{x} = \frac{\sum_{i=1}^{r} x_i n_{i\cdot}}{N} = \frac{175,7}{5} = 35,14$$

$$\overline{y} = \frac{\sum_{j=1}^{s} y_j n_{\cdot j}}{N} = \frac{7,41}{5} = 1,48$$

Con todos estos datos podemos calcular la varianza marginal de X:

$$S_X^2 = \frac{\sum_{i=1}^{r} x_i^2 n_{i\cdot}}{N} - \overline{x}^2 = \frac{6.275,89}{5} - (35,14)^2 = 20,36$$

Por su parte la covarianza entre ambas variables es:

$$S_{XY} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{5} x_i y_j n_{ij}}{N} - \overline{x}\,\overline{y} = \frac{277,10}{5} - (35,14 \cdot 1,48) = 3,34$$

Con todos estos valores podemos obtener los parámetros de la recta de regresión:

$$b = \frac{S_{XY}}{S_X^2} = \frac{3,34}{20,36} = 0,1641$$

$$a = \overline{y} - \frac{S_{XY}}{S_X^2} \cdot \overline{y} = 1,48 - 0,1641 \cdot 35,14 = -4,2860$$

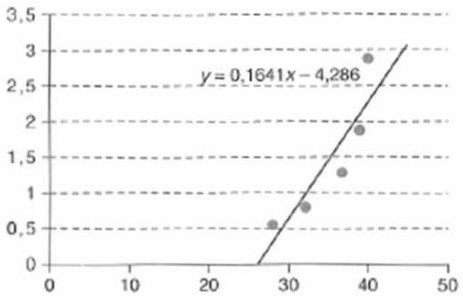
Por tanto, la recta obtenida es:

$$Y = -4,2860 + 0,1641X$$

El valor obtenido para la pendiente de la recta, $b = 0,1641$, indica que cuando la renta de los individuos aumenta, su gasto en vacaciones también lo hace si bien en menor cuantía. Dicho de otro modo, por cada 1.000 euros de renta adicional que obtiene el individuo, su gasto aumenta en 164,1 euros. Por su parte el término indepen-

diente, $a = -4{,}286$, puede interpretarse como que, cuando el individuo no tiene rentas, éste «ahorra» 4.286 euros en vacaciones.

Gráficamente, se presenta en la siguiente figura la recta ajustada a la nube de puntos:



Asimismo, utilizando la ecuación obtenida podemos elaborar una tabla con los errores cometidos:

x_i	Valor Real (y_i)	Valor previsto ($\widehat{y}_i$)	$e_i = y_i - \widehat{y}_i$
28,00	0,56	0,31	0,25
32,00	0,81	0,97	−0,16
36,70	1,28	1,74	−0,46
39,00	1,88	2,12	−0,24
40,00	2,88	2,28	0,60

5.5.2. Bondad del Ajuste

Una vez realizado un ajuste de la recta de regresión, pasamos a analizar en qué medida queda explicada la variable endógena por la variable exógena en base al ajuste realizado.

Para ello, generalmente se utiliza el ***coeficiente de determinación*** R^2 con el que se mide la proporción de variabilidad de la variable dependiente respecto a su media que es explicada por el modelo de regresión. Dicho coeficiente se obtiene mediante la siguiente expresión:

$$R^2 = r_{XY}^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}$$

En la que podemos observar que para obtener el coeficiente de determinación basta con elevar al cuadrado el coeficiente de correlación lineal de Pearson. Por ello, al igual que éste, su valor varía entre 0 y 1; cuando su valor es cero, la representatividad de la ecuación de regresión será nula, mientras que cuando su valor es igual a la unidad, el ajuste de la recta obtenido a la nube de puntos será perfecto. Como criterio general, el valor del coeficiente de determinación a partir del cual se considera que la

recta es aceptablemente representativa de la relación entre las variables debe ser superior a 0,75.

Ejemplo 5.7. Continuando con el ejemplo 5.6, vamos a calcular el coeficiente de determinación de la regresión. Para ello previamente debemos obtener el valor de la varianza marginal de Y:

x_i	y_i	x_j^2
28,00	0,56	0,3136
32,00	0,81	0,6561
36,70	1,28	1,6384
39,00	1,88	3,5344
40,00	2,88	8,2944
		14,4369

Por tanto:

$$S_Y^2 = \frac{\sum_{j=1}^{c} x_j^2 n_{*j}}{N} - \bar{y}^2 = \frac{14,4369}{5} - (1,48)^2 = 0,6911$$

Dado que $S_X^2 = 20,36$ y $S_{XY} = 3,34$, podemos calcular el valor de R^2 tal que:

$$R^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2} = \frac{(3,34)^2}{20,36 \cdot 0,69} = 0,7937$$

El valor obtenido indica que fiabilidad o confianza del ajuste realizado para representar la relación entre las variables es aceptable al ser superior a 0,75.

5.5.3. Interpolación y Extrapolación

Una vez hemos determinado una relación causal entre dos variables, hemos ajustado una recta utilizando el método de Mínimos Cuadrados y hemos comprobado que el ajuste es aceptable utilizando el coeficiente de determinación, ha llegado el momento de proceder a la realización de previsiones con el modelo obtenido. Para ello, obtendremos valores previstos para la variable dependiente utilizando para ello valores dados de la variable independiente así como los coeficiente a y b estimados. En particular, utilizando la ecuación estimada podemos:

- Pronosticar los valores de la variable dependiente a partir de valores de la variable independiente que pertenecen al intervalo de variación de los datos observados; es lo que se conoce como ***interpolación***.

- Predecir valores de la variable dependiente a partir de valores de la variable independiente que estén situados fuera de dicho intervalo, operación que recibe el nombre de ***extrapolación***.

A la hora de realizar predicciones, debemos tener en cuenta que su fiabilidad dependerá de dos factores:

- La calidad del ajuste: nuestras predicciones serán más fiables cuanto mejor sea el ajuste, es decir, cuanto mayor sea el valor del coeficiente R^2.
- Los valores de la variable independiente: la fiabilidad de la predicción disminuirá a medida que nos alejemos del rango que comprende a los datos de partida, ya que desconocemos cómo es la relación entre las variables a partir de determinados valores.

Ejemplo 5.8. Utilizando la ecuación obtenida en el ejemplo 5.6, vamos a realizar algunas predicciones:

- Previsión para $X = 30{,}00$ (interpolación)
  $-4{,}2860 + 0{,}1641 \cdot 30{,}00 = 0{,}6383$.
- Previsión para $X = 35{,}00$ (interpolación)
  $-4{,}2860 + 0{,}1641 \cdot 35{,}00 = 1{,}4590$.
- Previsión para $X = 45{,}00$ (extrapolación)
  $-4{,}2860 + 0{,}1641 \cdot 45{,}00 = 3{,}1005$.
- Previsión para $X = 55{,}00$ (extrapolación)
  $-4{,}2860 + 0{,}1641 \cdot 55{,}00 = 4{,}7419$.

5.6. EJERCICIOS DE AUTOEVALUACIÓN

1. Dos variables X e Y presentan una fuerte dependencia lineal. Sabiendo que $\overline{x} = 5$, $\overline{y} = 4\overline{x}$ y que $S_Y^2 = 8S_{XX}$, ¿qué valor cabe esperar para Y si $X = 1$?

   a) –8.
   b) 19,5.
   c) –12.
   d) 25.

2. Sea $Y = 4{,}736 + 1{,}088X$ la recta de regresión de Y sobre X. La varianza de la variable dependiente es 36, mientras que la media y la varianza de la variable independiente son, respectivamente, 3 y 25. En ese caso, la recta de regresión de X sobre Y es:

   a) $X = 0{,}755 + 3{,}04Y$.
   b) $X = -3{,}04 + 0{,}755Y$.
   c) $X = 0{,}755 - 3{,}04Y$.
   d) $X = 0{,}755 + 0{,}755Y$.

3. Se han obtenido las siguientes rectas de regresión de Y sobre X y X sobre Y:

$$Y = 5 - 0{,}4Y$$

$$X = 1 - 0{,}9Y$$

   En ese caso los valores para el coeficiente de determinación R^2 y el coeficiente de correlación de Pearson de estas regresiones es:

   a) $R^2 = 0{,}36$ y $r_{XY} = 0{,}6$.
   b) $R^2 = -0{,}6$ y $r_{XY} = -0{,}36$.
   c) $R^2 = 0{,}36$ y $r_{XY} = -0{,}6$.
   d) $R^2 = 0{,}6$ y $r_{XY} = 0{,}36$.

4. Señale la afirmación verdadera:

   a) El coeficiente de correlación lineal está comprendido entre 0 y 1.
   b) Si las variables son independientes, el coeficiente de correlación lineal es igual a cero.
   c) Si el coeficiente de correlación lineal es igual a cero, las variables son independientes.
   d) Ninguna de las anteriores.

5. El coeficiente de correlación lineal de una variable consigo misma tiene un valor de:

   a) +1.
   b) –1.
   c) 0.
   d) No puede calcularse.

6. En una distribución bidimensional se sabe que $S_{XY} = 4{,}5$, $S_Y^2 = 9$ y que el término independiente de regresión de Y sobre X es igual a 1. En ese caso:

   a) Si $\overline{x} = 2$ e $\overline{y} = 5$, la recta de regresión de X sobre Y es $X = 3 - Y$.
   b) El coeficiente de correlación lineal es igual a +0,5.
   c) Si $\overline{x} = 2$ e $\overline{y} = 5$, la recta de regresión de Y sobre X es $X = -0{,}5 + 0{,}5Y$.
   d) El coeficiente de correlación lineal es igual a –0,5.

7. Señale cuál de las siguientes relaciones nunca podría darse en una regresión lineal:

   a) $R^2 > 0$ y $r_{XY} = -0{,}9$.
   b) $r_{XY} = 0$ y $S_{XY} = -15$.
   c) $R^2 > 0$ y $S_{XY} = -15$.
   d) $R^2 > 0$ y $r_{XY} = 0$.

8. Sabiendo que $S_X^2 = 12{,}5$, $S_Y^2 = 20$, que la relación entre X e Y es inversa y que la proporción de variabilidad explicada por regresión lineal de Y sobre X es del 90%, el valor de la covarianza entre las dos variables será:

   a) –15.
   b) +15.
   c) –225.
   d) +225.

9. Señale cuál de las siguientes afirmaciones es correcta:

   a) La sensibilidad de la variable dependiente ante cambios unitarios en la variable independiente está medida por la pendiente de la recta de regresión incrementado en el valor del término constante.
   b) Si la pendiente de la recta de regresión es negativo, la relación entre las variables no es significativa.
   c) La sensibilidad de la variable dependiente ante cambios unitarios en la variable independiente está medida por el término constante de la recta de regresión.
   d) La sensibilidad de la variable dependiente ante cambios unitarios en la variable independiente está medida por la pendiente de la recta de regresión.

10. ¿Cuál de las siguientes situaciones referidas a la relación entre dos variables X e Y no podría darse nunca?:

    a) $Y = 1{,}8 - 0{,}4X$ y $R^2 = 0{,}8$.
    b) $Y = 1{,}8 - 0{,}4X$ y $X = 5 - 2{,}6Y$.
    c) $R^2 = 0{,}8$ y $S_{XY} = -3{,}45$.
    d) Ninguna de las anteriores.

11. Dada la siguiente distribución bidimensional de frecuencias:

X/Y	3	4	8	$n_{i\cdot}$
5	4	2	2	8
6	2	1	2	5
7	1	2	4	7
$n_{\cdot j}$	7	5	8	$N = 20$

¿Cuáles serían los valores de $\overline{x}$, $\overline{y}$, S_X^2 y S_{XY}?

a) $\overline{x} = 6$, $\overline{y} = 5{,}25$, $S_X^2 = 36{,}15$, $S_{XY} = 35{,}80$.
b) $\overline{x} = 5{,}95$, $\overline{y} = 7{,}01$, $S_X^2 = 36{,}15$, $S_{XY} = 25{,}31$.
c) $\overline{x} = 6$, $\overline{y} = 5{,}25$, $S_X^2 = 32{,}30$, $S_{XY} = 31{,}85$.
d) $\overline{x} = 5{,}95$, $\overline{y} = 5{,}25$, $S_X^2 = 36{,}15$, $S_{XY} = 31{,}85$.

12. Utilizando los datos del ejercicio anterior, ¿cuál sería la recta de regresión de Y sobre X?

a) $Y = 0{,}15 + 1{,}14X$.
b) $Y = -0{,}15 + 1{,}14X$.
c) $Y = 0{,}01 + 0{,}88X$.
d) $Y = -0{,}01 + 0{,}88X$.

LECTURAS RECOMENDADAS

ALEGRE, J. *et al.* (2003). *Análisis Cuantitativo de la Actividad Turística*. Ed. Pirámide.
FERNÁNDEZ, C. (1993). *Manual de Estadística Descriptiva Aplicada al Sector Turístico*. Ed. Síntesis.
RAYA, J.M. (2004). *Estadística Aplicada al Turismo*. Ed. Pearson Prentice Hall.
RONQUILLO, A. (1997). *Estadística Aplicada al Sector Turístico. Técnicas cuantitativas y Cualitativas de Análisis Turístico*. Ed. CEURA.
SANTOS, J. *et al.* (2007). *Estadística para Estudios de Turismo*. Ediciones Académicas.
URIEL, E. y Muñiz, M. (1988). *Estadística Económica y Empresarial*. Editorial AC.

PALABRAS CLAVE

- Distribución bidimensional
- Frecuencia absoluta conjunta
- Frecuencia absoluta relativa
- Frecuencia marginal conjunta
- Frecuencia marginal relativa
- Distribución marginal
- Distribución condicionada
- Dependencia estadística
- Covarianza
- Correlación
- Causalidad
- Espurio
- Regresión
- Coeficiente de determinación
- Interpolación
- Extrapolación

Soluciones a los ejercicios propuestos

CAPÍTULO 1. INTRODUCCIÓN. CONCEPTOS BÁSICOS

1. b y d
2. a, e y f son discretas, b, c y d son continuas.
3. c, d, g, i y j son variables, mientras que a, b, e, f, h y k son atributos.
4. Los atributos a, c y e so n ordinales; por su parte, los atributos b, d y f son ordinales.
5. c

CAPÍTULO 2. FUENTES DE INFORMACIÓN ESTADÍSTICA DE INTERÉS PARA EL SECTOR TURÍSTICO

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

CAPÍTULO 3. DISTRIBUCIONES DE FRECUENCIAS UNIDIMENSIONALES

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

CAPÍTULO 4. MEDIDAS DE POSICIÓN, DISPERSIÓN, FORMA Y CONCENTRACIÓN

1. b	5. c	9. d	13. b	17. a
2. c	6. b	10. d	14. a	18. a
3. d	7. a	11. c	15. d	19. d
4. a	8. c	12. d	16. b	20. b

CAPÍTULO 5. DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES. REGRESIÓN Y CORRELACIÓN

1. b	4. b	7. b	10. b
2. c	5. a	8. a	11. d
3. c	6. c	9. d	12. c

Bibliografía

ALEGRE, J. *et al.* (2003). *Análisis Cuantitativo de la Actividad Turística*. Ed. Pirámide.

CASAS, J.M. y SANTOS, J. (2002). *Introducción a la Estadística para Economía y Administración de Empresas*. Ed. CEURA.

FERNÁNDEZ, C. (1993). *Manual de Estadística Descriptiva Aplicada al Sector Turístico*. Ed. Síntesis.

FERNÁNDEZ, A. y LACOMBA, B. (2003). *Técnicas Estadísticas para el Turismo*. Ed. Ágora

GONICK, L. y SMITH, W. (2002). *La Estadística en Cómic*. Ed. Zendrera.

GONZÁLEZ, J.M. (1994). *El Azar y la Historia*. Ed. Planeta.

Instituto de Estudios Turísticos (2008). *Balance de Resultados de Demanda Turística Internacional 2004-2007 desde la Óptica de los Mercados Emisores*.

Instituto de Estudios Turísticos (2008). *El Turismo Español en Cifras* 2007.

Instituto de Estudios Turísticos (2009). *España en Europa. El Comportamiento Turístico de los Residentes en la Unión Europea*.

LÓPEZ, M. (1996). *Fundamentos y Métodos de Estadística*. Ed. Pirámide.

MARTÍN-GUZMÁN, M.P. y MARTÍN, F.J. (1993). *Curso Básico de Estadística Económica*. Ed. AC.

RAO, C.R. (1994). *Estadística y Verdad (Aprovechando el Azar)*. Ed. PPU.

RAYA, J.M. (2004). *Estadística Aplicada al Turismo*. Ed. Pearson Prentice Hall.

RONQUILLO, A. (1997). *Estadística Aplicada al Sector Turístico. Técnicas cuantitativas y Cualitativas de Análisis Turístico*. Ed. CEURA.

SÁNCHEZ, J. (1975). *Historia de la Estadística como Ciencia en España (1500-1900)*. Instituto Nacional de Estadística.

SANTOS, J. *et al.* (2007). *Estadística para Estudios de Turismo*. Ediciones Académicas.

TANUR, J.M. *et al.* (1992). *La Estadística. Una Guía de lo Desconocido*. Alianza Editorial

URIEL, E. y MUNIZ, M. (1988). *Estadística Económica y Empresarial*. Editorial AC.

INTRODUCCIÓN A LA ESTADÍSTICA PARA TURISMO

Alberto Muñoz Cabanes
Alfonso Herrero de Egaña y Espinosa de los Monteros
Azahara Muñoz Martínez

En la actualidad la Estadística constituye una herramienta fundamental a la hora de tomar decisiones en el campo de la empresa turística, pues permite analizar e interpretar datos económicos del sector turístico, y sirve como apoyo para el aprendizaje de otras materias que se encuentran en el programa del grado de Turismo tales como Marketing o Economía.

Utilizando un lenguaje claro y sencillo en la exposición de los conceptos pero sin renunciar al rigor matemático, los autores introducen al estudiante en el razonamiento estadístico, haciendo especial énfasis en la resolución de problemas de índole económica y empresarial, preferentemente relacionados con la actividad en el sector turístico, presentando numerosos ejemplos que permiten una rápida comprensión de las ideas presentadas.

En particular, el estudio del presente manual permitirá al alumno:

- Comprender la importancia que representa el uso de métodos estadísticos para el sector turístico.
- Conocer y analizar las principales fuentes de información estadística para describir el sector turístico español.
- Resumir la información mediante su representación a través de tablas y gráficos.
- Calcular e interpretar el valor de diferentes medidas de posición, dispersión, forma y concentración que permiten sintetizar la información relevante de la actividad económica y turística.
- Medir y modelizar relaciones de dependencia entre variables estadísticas desde una óptica descriptiva.

Al término de cada capítulo se presenta una serie de ejercicios de autoevaluación de tipo test, con el propósito de que el alumno pueda afianzar el aprendizaje de los conocimientos adquiridos y cuyas soluciones se pueden encontrar al final del libro. También se ha incluido una selección de lecturas recomendadas mediante las cuales es posible ampliar información sobre los contenidos tratados en cada capítulo, así como una lista de palabras clave, cuyo conocimiento resulta fundamental para una adecuada comprensión del tema.

EDICIONES ACADÉMICAS

UNED

ISBN: 978-84-9
9 788492