Introducción

La función principal de la Estadística es la recopilación de datos relevantes para construir, a partir de estos, informes que nos permitan analizar un tema concreto de estudio. La Estadística se convierte, entonces, en una disciplina matemática que nos habla de cantidades, donde la información cuantitativa que nos brinda nos permite conocer mucho mejor a una sociedad, como por ejemplo, cuántas personas viven en un país, cuál es la tasa de desempleo, cuál es la tasa de indigencia o pobreza, cuál es el nivel promedio de educación de esa sociedad, etc. Todos estos datos numéricos son utilizados por los responsables del Estado para realizar proyectos de diferente tipo que estudien cómo mejorar esa situación o mantenerla en el caso de que sea favorable. Así, la Estadística permite la toma de decisiones dentro del ámbito gubernamental, pero también en el mundo de los negocios y el comercio. La Estadística es, por tanto, la ciencia que estudia métodos y procedimientos para recoger, organizar, resumir y analizar un conjunto de datos recopilados, así como para obtener conclusiones válidas y tomar decisiones razonables basadas en tal análisis. Podemos, por tanto, clasificar la Estadística en «Descriptiva» e «Inferencial».

La *Estadística Descriptiva* describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen la información. Así, hablamos de Estadística Descriptiva cuando los resultados del análisis no pretenden ir más allá del conjunto de datos.

La *Inferencia Estadística* estudia los métodos para establecer conclusiones y tomar decisiones sobre una población a partir de una muestra de la misma. Así, hablamos de Estadística Inferencial cuando el objetivo del estudio es derivar las conclusiones obtenidas en los datos recopilados a un conjunto de datos más amplio. Esta toma de decisiones va acompañada de un margen de error, cuya probabilidad está determinada. De ahí que el paso del Análisis de Datos a la Inferencia Estadística requiera el manejo de conceptos y resultados relacionados con la Probabilidad.

En consecuencia, los pasos a la hora de realizar un estudio estadístico son los siguientes:

1. Definir el problema en términos precisos, indicando la población que se quiere investigar y las características que queremos analizar (variables).

2. Decidir qué datos recoger (muestra). Aunque la muestra puede incluir toda la población, generalmente será un subconjunto de esta.

3. Recoger los datos.

4. Describir (resumir) los datos obtenidos: Estadística Descriptiva.

5. Extrapolar o inferir las conclusiones obtenidas en los datos recogidos a un conjunto más amplio de la población, cuantificando la confianza que tienen dichas conclusiones: Inferencia Estadística.

Este manual pretende ser un punto de partida en Estadística para el estudiante de las carreras de empresariales, mostrándole cómo llevar a cabo un análisis descriptivo de datos, tanto unidimensional como bidimensional, a lo que están orientados los capítulos 1 y 2, respectivamente, para pasar al estudio de la Probabilidad en los capítulos 4, 5, 6 y 7, estando el tercero dedicado al estudio de números índices, algo esencial desde el punto de vista de la Economía, las Finanzas, el Marketing, la Contabilidad y la Administración de Empresas.

Almería, 23 de noviembre de 2024 Los autores

Capítulo 1

Análisis de datos unidimensionales

Siguiendo el orden lógico secuencial de un estudio estadístico, comenzaremos abordando la materia correspondiente a la parte de «Estadística Descriptiva», que de forma resumida se puede definir como un un conjunto de técnicas numéricas y gráficas orientadas a describir de manera sintética un conjunto de datos recopilados de una serie de individuos o elementos, con el fin de hacer más comprensibles esos datos. A ese conjunto de individuos o elementos (con ciertas características comunes) de los que se quiere estudiar cierta información se le llama **población**, mientras que la propiedad o cualidad respecto a la cual queremos hacer el estudio se conoce como **carácter**.

En relación a la población de estudio, el problema más habitual al que nos enfrentamos en la práctica es que dicha población puede tener un tamaño demasiado grande, lo que hace que no siempre sea factible observar de manera exhaustiva todos y cada uno de sus elementos (por razones de coste, rapidez en la obtención de la información,...) En estos casos, el procedimiento más usual consiste en realizar una observación parcial a partir de un grupo representativo de individuos o elementos de la población al que llamaremos **muestra**. Así, en el caso de que nuestro estudio se haga sobre la totalidad de la población se dice que hemos realizado un **censo**, mientras que si el estudio se limita a la observación de una muestra de la población, se dice que hemos realizado una **encuesta** o **estudio muestral**.

Por otro lado, asociado a la característica que queremos estudiar de los individuos o elementos de la población surge el siguiente concepto:

Definición 1.1. Una **variable** es la característica de la muestra o población que se está observando y que varía entre los diferentes individuos o elementos del estudio. Se suele denotar por letras mayúsculas (X, Y, Z,...) y en ella se recogen todos los posibles valores que toma la característica de interés. Dependiendo de sus valores, estas se clasifican en:

- Cualitativas: son cualidades o atributos de los individuos. No son medibles numéricamente; no podemos operar con sus valores. Estas, a su vez, se subdividen en:

    1. **Nominales**: cuando no se puede establecer ningún tipo de ordenación (en magnitud) en los valores que toma la variable, por ejemplo el sexo (hombre, mujer), o la nacionalidad del individuo (española, italiana, francesa,...).
    2. **Ordinales**: cuando sí podemos establecer algún tipo de ordenación (en magnitud) en los valores que toma la variable, como por ejemplo la satisfacción del individuo (insatisfecho, satisfecho), la motivación de los empleados (muy desmotivado, desmotivado, motivado, muy motivado), la calidad del servicio prestado (muy deficiente, malo, regular, bueno, excelente), etc.

    Una práctica muy común a la hora de introducir las observaciones correspondientes de una variable cualitativa de estudio en una base de datos para su posterior tratamiento informático, es la de codificar sus distintos valores. La codificación consiste en asignar un número o una letra a cada una de las distintas modalidades de la variable, para agilizar el proceso manual de introducción de datos. Por ejemplo, si la variable cualitativa es el sexo, podemos asignar a la modalidad «hombre» el número 0 o la letra H y a la modalidad «mujer» el número 1 o la letra M. En el caso de optar por números para la codificación, es importante no olvidar que la naturaleza de la variable es de tipo cualitativo, por lo que no podemos realizar operaciones aritméticas con esos números.

- Cuantitativas: miden algo cuantificable en cada individuo. Toman valores numéricos con los que se puede operar. Según las propiedades del conjunto de valores que toman pueden ser:

    1. **Discretas**: si la variable cuantitativa solo puede tomar una cantidad finita o numerable de valores. En la práctica, se corresponde con aquellas variables que solo admiten valores enteros que se repiten un gran número de veces entre los distintos individuos de la población. Ejemplos: número de hijos de una familia, número de días de la semana que el individuo práctica deporte, o el número de goles de un equipo en cada partido.
    2. **Continuas**: si la variable puede tomar una cantidad infinita no numerable de valores. En general, cualquier variable que puede tomar cualquier valor dentro de un intervalo se considera de tipo continuo, por ejemplo, la estatura o peso de la persona. En la práctica, se consideran también variables continuas aquellas que pueden tomar un gran número de valores diferentes, aún cuando

sean valores enteros, como por ejemplo, el salario de los trabajadores de una empresa.

La distinción entre los distintos tipos de variables es importante porque las técnicas a aplicar a cada una pueden ser muy diferentes, y muchos parámetros y cálculos tienen sentido para las variables de un tipo y no para las de otro. Hay que tener en cuenta también que una misma variable de la realidad puede venir expresada de diversas maneras, incluso como cualitativa o como cuantitativa, dependiendo de que usemos valores numéricos o solo modalidades. Piénsese, por ejemplo, en que la variable «calificación obtenida en una asignatura» puede expresarse numéricamente (variable cuantitativa continua) o bien expresada bajo las categorías «suspenso», «aprobado», «notable», «sobresaliente» o «matrícula de honor» (variable cualitativa ordinal). En estos casos, debe quedar claro que la variable es, en esencia, cuantitativa y que su tratamiento como cualitativa supone una pérdida de calidad en la información, solo admisible si no podemos disponer de los datos numéricos.

Ejemplo 1.1. *Clasificar las siguientes variables, relativas a la información solicitada mediante encuesta a los estudiantes de primer curso universitario en la Universidad de Almería:*

a) *Nota de acceso a la universidad.*

b) *Número de asignaturas en las que se ha matriculado el estudiante en el primer cuatrimestre.*

c) *Nivel de inglés acreditado del estudiante (ninguna acreditación, A1, A2, B1, B2, C1, C2).*

d) *Número de idiomas que habla el estudiante.*

e) *Asistencia a clases particulares para la preparación de alguna asignatura (SI, NO).*

f) *Tiempo, en horas, dedicado semanalmente al estudio.*

g) *Grado en el que está matriculado el estudiante.*

h) *Nivel de satisfacción global del estudiante con el Grado en el que está matriculado (Muy insatisfacho, Insatisfecho, Satisfecho, Muy satisfecho).*

Solución.

a) Variable cuantitativa continua.

b) Variable cuantitativa discreta.

c) Variable cualitativa ordinal.

d) Variable cuantitativa discreta.

e) Variable cualitativa nominal.

f) Variable cuantitativa continua.

g) Variable cualitativa nominal.

h) Variable cualitativa ordinal.

■

Por otro lado, si tenemos en cuenta el número de características que estudiamos en los individuos o elementos de la población, las variables pueden ser:

- **Unidimensionales**: si solo analizamos una característica en los individuos o elementos de la población. Por ejemplo, la edad (X) de los estudiantes de cierta clase universitaria, $X = \{18, 19, 20, \ldots, 25, \ldots\}$.
- **Bidimensionales**: si analizamos dos características en los individuos o elementos de la población de forma conjunta. Por ejemplo, la edad (X) y la altura (Y), en cm, de los estudiantes de una clase, $(X, Y) = \{(18, 165), \ldots, (19, 170), \ldots, (20, 174), \ldots, (25, 168), \ldots\}$.
- **Pluridimensionales**: si analizamos más de dos características en los individuos o elementos de la población de forma conjunta. Por ejemplo, la edad (X), altura (Y), en cm, y peso (W), en kg, de los estudiantes de una clase, $(X, Y, W) = \{(18, 165, 52), \ldots, (19, 170, 62), \ldots, (20, 174, 65), \ldots, (25, 168, 58), \ldots\}$.

En este manual, en concordancia con el contenido de la asignatura al que está enfocado, se abordará únicamente el estudio de variables unidimensionales y bidimensionales.

Cualquier estudio estadístico requiere un proceso de recolección de datos, pero dicha recopilación de datos revela muy poca información por sí sola. Para determinar su significancia, los datos deben organizarse de manera que, con un simple vistazo, se pueda tener una idea de lo que pueden decirnos. Así, la descripción de los valores observados es la primera etapa de cualquier análisis estadístico. En este capítulo veremos los métodos que utiliza la Estadística Descriptiva para describir grandes conjuntos de datos cuando solo

estudiamos una característica de los individuos o elementos de la población (variable unidimensional). La selección de la herramienta más adecuada depende del tipo de variable que estemos analizando. De aquí en adelante, seguiremos las siguientes notaciones:

- N será el tamaño total de la población objeto de estudio.
- X será la variable unidimensional objeto de estudio, y $x_1, x_2, \ldots, x_n$ los distintos valores que toma dicha variable, ordenados de menor a mayor.

1.1. Distribuciones de frecuencias

Para resumir y organizar mejor la información obtenida tras el estudio de una variable, se utilizan las llamadas **tablas de frecuencias**, que se basan en una técnica de conteo. Dependiendo de la tipología de la variable de estudio, esta adquiere un formato diferente.

1.1.1. Tabla de frecuencias con datos no agrupados

En el caso de que la variable de estudio sea cualitativa (toma como valores distintas modalidades) o cuantitativa discreta (toma valores enteros que se repiten un gran número de veces), la tabla de frecuencias está compuesta por las columnas siguientes:

- **Valores de la variables** (x_i): diferentes valores o modalidades que toma la variable.
- **Frecuencia absoluta** (n_i): es el número de veces que se repite cada valor de la variable x_i. La suma de todas las frecuencias absolutas n_i coincide siempre con el número total de datos N.
- **Frecuencia relativa** (f_i): es la proporción de individuos de la población que presentan el valor x_i, es decir, $f_i = \dfrac{n_i}{N}$. Suelen expresarse en % (porcentaje de individuos de la población que toma el valor x_i), multiplicándolas por 100. La suma de todas las frecuencias relativas, f_i, es siempre 1 (o 100 si se expresan en %).

Además, para variables de tipo cuantitativas o cualitativas ordinales, cuyos valores se ordenan siempre de menor a mayor, se pueden definir las frecuencias acumuladas:

- **Frecuencia absoluta acumulada** (N_i): es el número de individuos que presentan valores menores o iguales a x_i, es decir, $N_i = n_1 + n_2 + \ldots + n_i$. La última frecuencia absoluta acumulada coincide siempre con el total de datos N.

- **Frecuencia relativa acumulada** (F_i): es la proporción de individuos de la población que presentan un valor menor o igual a x_i, es decir:

$$F_i = \frac{N_i}{N} = f_1 + f_2 + \ldots + f_i.$$

También suelen expresarse en %, multiplicándolas por 100, en cuyo caso, indican el % de individuos de la población que presentan un valor inferior o igual a x_i. La última frecuencia relativa acumulada siempre vale 1 (o 100 en el caso de expresarlas en %).

Observación 1.1. *Las frecuencias acumuladas carecen de sentido en el caso de variables cualitativas nominales, puesto que sus valores no se pueden ordenar de menor a mayor.*

De este modo, la tabla de frecuencias asociada a una **variable cualitativa nominal** adquiere el siguiente formato:

Valores	Frecuencia absoluta	Frecuencia relativa
x_i	n_i	f_i
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
x_n	n_n	f_n
	N	1

mientras que para el caso de una **variable cualitativa ordinal** o **cuantitativa discreta** su configuración es la siguiente:

Valores	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada
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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	N_i	f_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	n_n	N	f_n	1
	N		1	

Ejemplo 1.2 (Distribución de frecuencias de una variable cualitativa nominal)**.** *La siguiente tabla de frecuencias recoge el Grado en el que están matriculados los 40 estudiantes de un determinado grupo de la asignatura de Estadística:*

Grado	n_i	$f_i(\%)$
Economía		*22,5*
FYCO	*10*	
ADE	*15*	
Marketing		

Completar la tabla de frecuencias anterior y añadir las columnas que proceda.

Solución. Estamos ante una variable cualitativa nominal, pues las categorías en las que queda agrupada no admiten una ordenación natural. Con objeto de completar los datos faltantes en la columna de frecuencias absolutas, n_i, debemos tener en cuenta que

$$f_i = \frac{n_i}{40} \cdot 100 \Rightarrow n_i = \frac{40}{100} \cdot f_i.$$

Por tanto, si el 22,5 % son estudiantes de Economía, estamos hablando de un total de $n_1 = \frac{40}{100} \cdot 22{,}5 = 9$ estudiantes matriculados en dicho grado. En segundo lugar, como hay 10 estudiantes de FYCO y un total de 40 en la población objeto de estudio, el porcentaje correspondiente es $\frac{10}{40} \cdot 100 = 25\,\%$. Razonando análogamente, hay un 37,5 % de estudiantes de ADE. Finalmente, conocidos los porcentajes de todas las categorías a excepción de Marketing, podemos concluir que a este le corresponde un 15 % del total, puesto que la suma de todos los porcentajes debe ser el 100 %. Así, $f_4 = 100 - (22{,}5 + 25 + 37{,}5) = 15\,\%$, que se corresponde con un total de 6 personas como frecuencia absoluta. Además, por la naturaleza de la variable, no tiene sentido añadir las columnas correspondientes a las frecuencias acumuladas.

Grado	n_i	$f_i(\%)$
Economía	9	22,5
FYCO	10	25
ADE	15	37,5
Marketing	6	15
	40	100

■

Ejemplo 1.3 (Distribución de frecuencias de una variable cualitativa ordinal)**.** *La siguiente tabla de frecuencias recoge la calificación obtenida en Estadística por los 90 estudiantes de un determinado grupo:*

Calificación	n_i
No presentado	*20*
Suspenso	*24*
Aprobado	*32*
Notable	*12*
Sobresaliente	*1*
Matrícula de Honor	*1*

Completar la tabla calculando las frecuencias que procedan para este tipo de variable. Con la información obtenida, contesta las siguientes cuestiones:

a) ¿Qué porcentaje de estudiantes no ha superado la asignatura?

b) ¿Cuántos estudiantes han obtenido una calificación superior a «aprobado»?

Solución. Estamos ante una variable cualitativa ordinal, pues las categorías en las que queda agrupada admiten una ordenación natural. En consecuencia, además de calcular la frecuencia relativa f_i, tiene sentido añadir las columnas correspondientes a las frecuencias acumuladas N_i y F_i. Teniendo en cuenta la relación entre las distintas frecuencias, la tabla resultante es la siguiente:

Calificación	n_i	$f_i(\%)$	N_i	$F_i(\%)$
No presentado	20	22,2	20	22,2
Suspenso	24	26,7	44	48,9
Aprobado	32	35,6	76	84,5
Notable	12	13,3	88	97,8
Sobresaliente	1	1,1	89	98,9
Matrícula de Honor	1	1,1	90	100
	90	100		

a) La respuesta a esta pregunta nos la proporciona $F_2 = 48{,}9\,\%$, puesto que indica el porcentaje de estudiantes que han obtenido una calificación inferior al aprobado. Se observa, por tanto, que casi la mitad de los estudiantes de ese grupo no han superado la asignatura.

b) La información proporcionada por $N_3 = 76$, indica que hay 76 estudiantes que ha obtenido una calificación inferior o igual a aprobado. En consecuencia, el resto de estudiantes ha obtenido una calificación superior a aprobado, esto es, $90 - 76 = 14$ estudiantes. Otra alternativa consiste en sumar directamente las frecuencias correspondientes $n_4 + n_5 + n_6 = 14$. ■

Ejemplo 1.4 (Distribución de frecuencias de una variable cuantitativa)**.** *Se ha recogido el número de días de ausencia a clase de Estadística de un grupo de 10 estudiantes en un determinado mes, obteniéndose los siguientes resultados: 7, 1, 5, 7, 9, 13, 5, 7, 9, 7. Construir la tabla de frecuencias y contestar las siguientes cuestiones:*

a) *¿Cuántos estudiantes han faltado a clase 5 o menos días en dicho mes?*

b) *¿Qué porcentaje de estudiantes ha faltado a clase 7 días en ese mes?*

c) *¿Qué porcentaje de estudiantes ha faltado menos de 9 días a clase?*

d) *¿Cuántos estudiantes han faltado 1 día a clase?*

e) *¿Qué porcentaje de estudiantes faltan más de 5 días a clase?*

f) *¿Cuántos estudiantes han faltado más de 7 días a clase?*

Solución. Tras un recuento, y teniendo en cuenta la relación entre las distintas frecuencias, es posible construir la siguiente tabla de frecuencias:

x_i	n_i	N_i	f_i	F_i
1	1	1	0,1	0,1
5	2	3	0,2	0,3
7	4	7	0,4	0,7
9	2	9	0,2	0,9
13	1	10	0,1	1
	10		1	

a) La respuesta a esta pregunta nos la proporciona $N_2 = 3$, que también se puede responder sumando los que faltan en una ocasión o en 5: $n_1 + n_2$.

b) Porcentaje implica hablar de frecuencia relativa. Como nos piden el correspondiente a un valor exacto, podemos responder a través de la frecuencia relativa, y concretamente, $f_3 = 0{,}4$ que, en porcentaje, es $40\,\%$.

c) Faltar menos de 9 días implica, de acuerdo con los datos tabulados, ausentarse 7 o menos, por lo que se nos está pidiendo $F_3 \times 100 = 0{,}7 \times 100 = 70\,\%$.

d) La respuesta a esta cuestión la tiene la frecuencia absoluta correspondiente al primer valor, esto es, $n_1 = 1$.

e) Faltarán más de 5 días a clase aquellos que no falten 5 días o menos, así que al porcentaje total, 100 %, debemos restar el que acumulan los que faltan 1 o 5 días, esto es, $F_2 \times 100 = 30\,\%$, y la solución es 70 %. Otra alternativa consiste en sumar, directamente, las frecuencias relativas correspondientes: $(f_3 + f_4 + f_5) \times 100 = 70\,\%$.

f) Los que han faltado más de 7 días son los que han faltado 9 o 13 días, esto es, $n_4 + n_5 = 2 + 1 = 3$ estudiantes, cantidad que también se puede razonar como $N - N_3 = 3$.

■

1.1.2. Tabla de frecuencias con datos agrupados en intervalos

Se utiliza cuando el número de valores distintos que toma la variable de estudio es muy elevado y dichos valores apenas se repiten (variable cuantitativa continua), con lo que parece aconsejable, para mayor comodidad en el tratamiento de la información, agrupar estos valores en intervalos de clase, lo que permitirá reducir la dimensión de la tabla de frecuencias asociada a dicha variable. Los puntos inicial y final de cada intervalo son sus extremos. Así, denotaremos a dichos intervalos de la forma $L_{i-1} - L_i$, donde L_{i-1} representa el extremo inferior de dicho intervalo y L_i el extremo superior.

A la hora de agrupar los datos en intervalos es necesario especificar el tipo de intervalo que se va a utilizar (si son abiertos o cerrados por cada uno de sus extremos). Esta cuestión es importante ya que puede darse el caso de que algunos de los valores de la variable coincidan con algún extremo de los intervalos, planteándose entonces el interrogante de en qué intervalo se debería incluir ese valor. De aquí en adelante, por convenio, consideraremos intervalos de la forma $[L_{i-1}, L_i)$, es decir, cerrados por el extremo inferior y abiertos por el extremo superior. En consecuencia, incluiremos el valor L_{i-1} en el intervalo, pero excluiremos el extremo superior, L_i, salvo que se trate del intervalo que alberga los mayores valores, en cuyo caso queda este incluido.

Cuando hay que agrupar datos en intervalos de clases se debe ponderar entre:

- Realizar pocos intervalos a costa de perder mucha información sobre los datos reales de cada intervalo.

- Agrupar en muchos intervalos, con los que las frecuencias resultantes de cada intervalo de clase pueden ser demasiado pequeñas para que se reconozcan los patrones de forma.

Se recomienda no construir más de 20 intervalos y, aunque lo más habitual suele ser construir entre 5 y 10 intervalos de clase, el número apropiado es una elección subjetiva donde el investigador puede, naturalmente, probar distintos números de intervalos de clases para ver cuál de los gráficos resultantes revela más información sobre los datos. No obstante, existen también ciertos criterios que nos ayudan aproximar el número de clases K a utilizar. Los más utilizados son los dos siguientes:

- Criterio de Norcliffe: K debe ser, aproximadamente, $\sqrt{N}$.
- Fórmula de Sturges, que se emplea cuando N es demasiado grande: $K \approx \frac{\log N}{\log 2} + 1$.

En cuanto a la amplitud de los intervalos, lo más corriente (aunque no esencial) es construir intervalos de la misma amplitud, ya que esto simplifica los gráficos y los cálculos. Para calcular la amplitud de los intervalos se aplica la siguiente fórmula:

$$\text{amplitud} = \frac{x_{\text{máx}} - x_{\text{mín}}}{K},$$

siendo $x_{\text{máx}}$ el valor más grande de la variable estadística analizada y $x_{\text{mín}}$ el más pequeño.

Una vez agrupados los datos en intervalos de clase, solamente hay que contabilizar cuántos valores de la variable se incluyen en cada intervalo para obtener las correspondientes frecuencias absolutas (n_i) y, a partir de ahí, el resto de frecuencias se obtienen de forma análoga al caso en que los valores estén sin agrupar.

De este modo, la tabla de frecuencias asociada a una **variable cuantitativa continua** adquiere el formato siguiente:

Intervalos	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada
$[L_{i-1}, L_i)$	n_i	N_i	f_i	F_i
$[L_0, L_1)$	n_1	N_1	f_1	F_1
$[L_1, L_2)$	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{i-1}, L_i)$	n_i	N_i	f_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[L_{n-1}, L_n]$	n_n	N	f_n	1
	N		1	

Es importante señalar que la agrupación de valores en intervalos de clase tiene una única finalidad visual, que es la de reducir la dimensión de la tabla de frecuencias. Por tanto, para cualquier otro tipo de análisis estadístico de interés, se debe trabajar siempre con los valores sin agrupar de la variable.

Ejemplo 1.5 (Distribución de frecuencias de una variable cuantitativa agrupada en intervalos)**.** *Para un grupo de 20 estudiantes, se recogieron las alturas, en cm, obteniéndose los resultados siguientes:*

174, 185, 166, 176, 145, 166, 191, 177, 164, 171
175, 158, 156, 156, 187, 162, 172, 193, 183, 173

Construir la tabla de frecuencias, agrupando los valores en intervalos de igual amplitud.

Solución. Dado que disponemos de un total de 20 datos, el criterio de Norcliffe aconseja considerar un total de $K \approx \sqrt{20}$ intervalos, esto es, $K = 4$. Ahora bien, como el menor valor de todos los que tenemos a nuestra disposición es 145 y el mayor 193, la amplitud será

$$\text{amplitud} = \frac{193 - 145}{4} = \frac{48}{4} = 12.$$

Por tanto, la tabla de frecuencias agrupada en intervalos de clase es la siguiente:

Intervalos	n_i	N_i	f_i	F_i
$[145, 157)$	3	3	0,15	0,15
$[157, 169)$	5	8	0,25	0,4
$[169, 181)$	7	15	0,35	0,75
$[181, 193]$	5	20	0,25	1
	20		1	

■

1.2. Representaciones gráficas

Aunque la distribución de frecuencias es una representación exhaustiva de la información disponible, es recomendable complementarla con una representación gráfica, pues ésta nos ayuda a visualizar de forma más rápida y clara distintas características de la variable de estudio. Los gráficos expresan con sencillez relaciones y propiedades que son más difíciles de ver en las tablas de frecuencias, permiten descubrir datos anómalos e identificar rápidamente algunos valores característicos (como el máximo, el mínimo o la moda, entre otros).

En cualquier caso, las tablas de frecuencias y las representaciones gráficas son dos maneras equivalentes de presentar la información. Las dos exponen ordenadamente la información recogida en una muestra, por lo que debe existir una concordancia absoluta entre ellas.

A la hora de representar una variable, hay que escoger un gráfico adecuado según el tipo de variable que se trate. De entre la gran diversidad de gráficos que existen, nos vamos a limitar a comentar los más elementales y de uso más habitual en la práctica.

1.2.1. Gráficos para variables cualitativas

1. **Diagrama de sectores**: círculo que se divide en tantos sectores como valores tenga la variable. El área de cada sector es proporcional a la frecuencia absoluta o relativa del valor.

   Son útiles cuando se desea comparar la importancia relativa de los distintos valores de la variable, es decir, para transmitir un sentido de equidad, tamaño relativo o desigualdad entre las categorías. Se utiliza sobre todo en variables de tipo cualitativo pero, debido a que en el gráfico de sectores lo importante es mostrar el porcentaje o proporción que le corresponde a cada categoría y no el orden, **son más adecuadas para representar variables nominales en lugar de ordinales**.
   Cuando se realiza un diagrama de sectores, se deben tener en cuenta las siguientes recomendaciones:

   - No deben usarse para mostrar relaciones entre las categorías. Si nuestro interés es mostrar el orden y la comparación entre las categorías, la gráfica de barras es más adecuada.
   - No se debe utilizar si el número de categorías de la variable es excesivamente grande, ya que el gráfico resultante no sería muy claro desde el punto de vista visual. Por este motivo, se recomienda su uso cuando la variable presenta como máximo 5 modalidades diferentes.
   - Actualmente y debido al uso de software de aplicación general es muy común que se elaboren gráficas en perspectiva simulando tres dimensiones cuando solo se desea representar una o dos, lo que produce distorsión y una mala comunicación. Por tanto, se recomienda no usar diagramas de sectores en tres dimensiones, ya que distorsionan y falsean la información.

2. **Diagrama de barras**: los gráficos de barras son usados para comparar dos o más valores. Se utiliza en el caso de que nuestro interés se limite, únicamente, en representar las frecuencias absolutas o relativas. Consiste en representar los valores de la variable y, sobre ella, se levantan rectángulos de igual base y altura proporcional a la frecuencia (absoluta o relativa). Las barras pueden orientarse horizontal o verticalmente. Se deja un hueco entre las barras para indicar los valores de la variable

que no son posibles. Este tipo de gráfico se utiliza para representar:

- variables cualitativas nominales con un gran número de categorías.
- variables cualitativas ordinales, puesto que los valores de la variable aparecen siempre representados de menor a mayor.

Ejemplo 1.6. *El diagrama de sectores asociado al Ejemplo 1.2 es el de la Figura 1.1.*

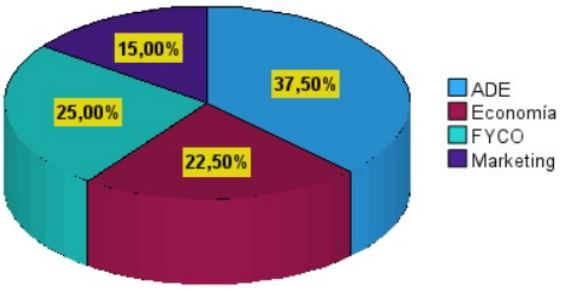

Figura 1.1

Ejemplo 1.7. *El diagrama de barras asociado al Ejemplo 1.3 es el de la Figura 1.2.*

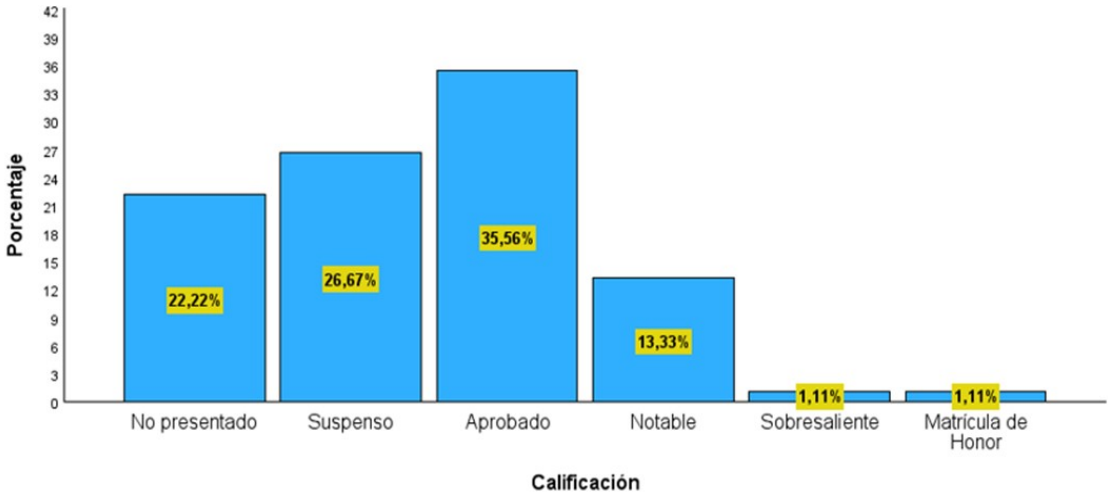

Figura 1.2

1.2.2. Gráficos para variables cuantitativas

1. **Diagrama de barras**: se utiliza para representar variables cuantitativas discretas, ya que este tipo de variables toman pocos valores diferentes, que no admiten valores intermedios entre dos valores consecutivos; además, al ser valores numéricos, se

pueden ordenar siempre de menor a mayor, lo que hace que el gráfico de barras sea una representación adecuada.

Ejemplo 1.8. *El diagrama de barras asociado al Ejemplo 1.4 es el de la Figura 1.3.*

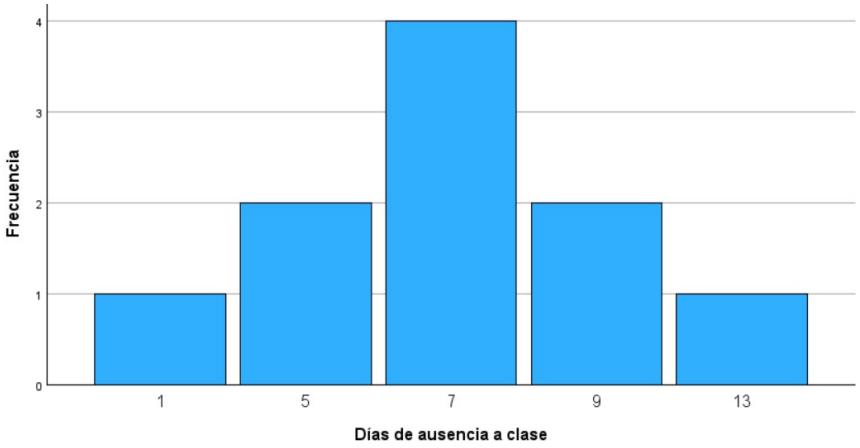

Figura 1.3

2. **Histograma**: se utiliza para representar variables continuas cuyos valores han sido agrupados en intervalos de clase. En el eje de abscisas se representan los extremos de los intervalos $[L_{i-1}, L_i)$ y, sobre cada uno de ellos, se levanta un rectángulo de base la amplitud del intervalo y área proporcional a la frecuencia (absoluta o relativa), en caso de que todos los intervalos tengan la misma amplitud. En caso de que no todos los intervalos tengan la misma amplitud, la altura de cada rectángulo será proporcional a la densidad de frecuencia, d_i, definida como el cociente entre la frecuencia absoluta del intervalo y la amplitud de este, pues de no considerar este valor, podríamos encontrarnos con una interpretación errónea de la distribución de los datos analizados.

   A primera vista, los histogramas parecen ser lo mismo que un gráfico de barras ya que ambas estructuras emplean barras verticales para representar los datos, pero si nos fijamos bien existen claras diferencias entre ambos tipos de gráficos, que encierran conceptos totalmente diferentes:

   - En un histograma no es la altura, sino el área de la barra lo que es proporcional a la frecuencia de ese intervalo. Los intervalos no tienen por qué ser todos iguales (aunque es lo más habitual), pero siempre tendrán un área mayor aquellos intervalos con mayor frecuencia.

- En un histograma se representan todos los valores posibles que existen dentro de los intervalos (por eso las barras están siempre juntas y no separadas como ocurre en el diagrama de barras), aunque no hayamos observado ninguno de forma directa. Permite, así, calcular la probabilidad de que se represente cualquier valor de la distribución, lo que es de gran importancia si queremos hacer inferencia y estimar valores de la población a partir de los resultados de nuestra muestra.

La importancia de un histograma estriba en que nos permite detectar, rápidamente, y de manera visual, características importantes de los datos. Por ejemplo, un histograma puede indicar a menudo: la simetría de los datos; la dispersión de estos; si existen intervalos que tienen un alto nivel de concentración; si algunos valores de datos están muy separados de otros, etc.

Ejemplo 1.9. *El histograma correspondiente al Ejemplo 1.5 es el de la Figura 1.4.*

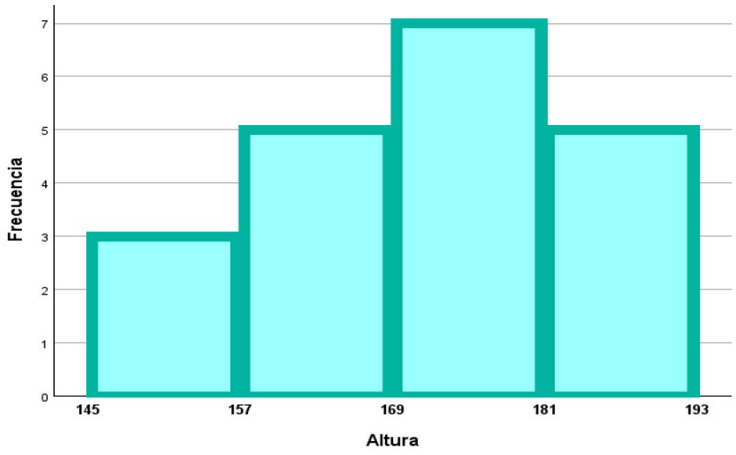

Figura 1.4

1.3. Estadísticos descriptivos

Cuando disponemos de una tabla de frecuencias asociada a una variable estadística, esta puede ser resumida por una serie de medidas que dan una idea global de cómo es la distribución sin tener que recordar todas las frecuencias absolutas o relativas. Las podemos clasificar en tres grupos:

- **Medidas de centralización**: indican valores con respecto a los que los datos parecen agruparse (media, mediana y moda).

- **Medidas de posición**: dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos o elementos (cuartiles, deciles, percentiles,...)
- **Medidas de dispersión**: informan sobre la variación del conjunto de datos. Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.

Aunque existen otro tipo de medidas que informan sobre la forma de la distribución (asimetría y apuntamiento), estas no van a ser abordadas en este manual por no estar incluidas en el contenido de la asignatura al que va destinado.

1.3.1. Medidas de tendencia central

Las medidas de tendencia central sirven para resumir la tabla de frecuencias en un valor (o valores) que permitan hacernos una idea del comportamiento global de la variable estudiada. Son medidas que buscan valores con respecto a los cuales los datos muestran tendencia a agruparse. Estos valores en los que resumimos la distribución facilita la comparación entre distintas distribuciones.

Las medidas de tendencia central más utilizadas en la práctica son: media aritmética (si bien en algunos determinados contextos es más idóneo promediar con media geométrica o armónica), mediana y moda. A continuación, se expone cada una de estas medidas.

- **Media aritmética** ($\bar{x}$): Se calcula dividiendo la suma total de todos los valores de la variable entre el número total de datos. Si tenemos en cuenta el número de veces n_i que se repite cada valor diferente x_i de la variable, la media aritmética se calcula con la siguiente expresión:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \ldots + x_n \cdot n_n}{N}.$$

  Es la medida de centralización más utilizada en **variables cuantitativas** (de su definición se deduce que no tiene sentido calcularse en variables cualitativas), debido a que en su expresión se utilizan todos los valores de la variable, tiene fácil interpretación y es siempre un valor único. El principal inconveniente de la media aritmética es que si la variable presenta valores anormalmente extremos, estos pueden distorsionar el cálculo de la media, haciéndola poco representativa como síntesis de la información. Por tanto, es conveniente usarla cuando los datos se concentran de manera más o menos simétrica con respecto a ese valor.

  Ejemplo 1.10. *Calcular el número medio de días de ausencia a clase de Estadística de los estudiantes del Ejemplo 1.4.*

Solución. El número medio de días de ausencia a clase de Estadística en un mes de ese grupo de 10 se puede hallar añadiendo una columna adicional a la tabla de frecuencias, la cual incluya el producto de cada valor por su frecuencia absoluta, y haciendo el cálculo dividiendo la suma de dichos valores por el total de datos:

x_i	n_i	$x_i \cdot n_i$
1	1	1
5	2	10
7	4	28
9	2	18
13	1	13
	10	70

Así, la media aritmética es $\bar{x} = \frac{70}{10} = 7$ días. ■

Algunas propiedades importantes de la media aritmética son las siguientes:

1. Si tenemos una población de tamaño N, dividida en varias subpoblaciones disjuntas de tamaños $N_1, N_2, \ldots, N_n$, de modo que para cada subpoblación conocemos sus medias respectivas, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$, entonces la media de la población total viene dada por la siguiente expresión:
$$\bar{x} = \frac{N_1 \cdot \bar{x}_1 + N_2 \cdot \bar{x}_2 + \ldots + N_n \cdot \bar{x}_n}{N_1 + N_2 + \ldots + N_n}.$$
2. A la media le afectan los cambios de origen y de escala: si conocemos la media de una variable cuantitativa X, $(\bar{x})$, y se nos define una nueva variable cuantitativa Y de la forma
$$Y = a + b \cdot X \qquad (a, b \in \mathbb{R})$$
entonces la media de la variable Y respeta la relación definida anteriormente, de modo que viene dada por la siguiente expresión: $\bar{y} = a + b \cdot \bar{x}$.

Ejemplo 1.11. *Una entidad bancaria dispone de cuatro sucursales en una determinada localidad. La siguiente tabla recoge el salario medio mensual, en euros, de los empleados de cada sucursal, así como el número de empleados que trabaja en cada una de ellas.*

Sucursal	*Sueldo medio*	*Número de empleados*
1	*1880*	*10*
2	*2000*	*8*
3	*1750*	*12*
4	*1800*	*6*

a) Calcular el sueldo medio de los empleados de dicha entidad bancaria.

b) La entidad bancaria en los últimos años ha incrementado sus ingresos de forma notable, de manera que han decidido subir el sueldo a sus empleados en un 12% ¿Cuál será el sueldo medio de los empleados de dicha entidad bancaria tras el prometido incremento?

Solución. *a)* Dado que son conocidos el sueldo medio y el número de empleados de cada sucursal, el sueldo medio mensual en la entidad bancaria es

$$\bar{x} = \frac{10 \cdot 1880 + 8 \cdot 2000 + 12 \cdot 1750 + 6 \cdot 1800}{10 + 8 + 12 + 6} = \frac{66780}{36} = 1855 \text{ euros.}$$

b) Teniendo en cuenta que la media respeta los cambios de escala, y que la nueva variable estadística es $Y = X + 0{,}12X = 1{,}12X$, se concluye que el nuevo sueldo medio es

$$\bar{y} = 1{,}12\bar{x} = 2077{,}6 \text{ euros mensuales.}$$

■

De aquí en adelante, cuando se haga referencia al término *media* o *promedio*, nos estaremos refiriendo a la media aritmética. Sin embargo, la media aritmética solo puede utilizarse si los datos con los que se trabaja son de naturaleza aditiva, es decir, que al sumar todos los valores, estos representen el total de la población. Esto ocurre, por ejemplo, con variables como los salarios de una empresa, las rentas, etc. Sin embargo, existen otros tipos de variables que no son aditivas, como los tipos de interés, la velocidad o la productividad. Para este tipo de variables deben utilizarse otros tipos de medias:

- **Media geométrica** (G): se utiliza cuando los valores de la variable son de naturaleza acumulativa o con efectos multiplicativos, como los tipos de interés, porcentajes, tasas, números índices, etc. Tiene una amplia aplicación en los negocios y en la economía, debido a que se utiliza para mostrar los cambios porcentuales en una serie de números positivos, como por ejemplo: el cambio porcentual en las ventas, en el producto nacional bruto o en cualquier serie económica. Se calcula mediante la fórmula

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}.$$

  No puede determinarse cuando la variable toma algún valor igual a 0. Además, cuando la variable tome valores positivos y negativos, puede existir o no la raíz.

Ejemplo 1.12. *Las ventas de una empresa, en miles de euros, para el período 2017- 2021 aparecen en la siguiente tabla:*

Año	*2017*	*2018*	*2019*	*2020*	*2021*
Ventas	*90*	*106*	*115*	*129*	*145*

Determinar el incremento medio anual de las ventas para el período considerado.

Solución. Lo primero que debemos hacer es hallar el incremento que se produce de un año para otro:

$$2017-2018: \frac{106-90}{90}\cdot 100 = 17{,}78\,\%$$

$$2018-2019: \frac{115-106}{106}\cdot 100 = 8{,}49\,\%$$

$$2019-2020: \frac{129-115}{115}\cdot 100 = 12{,}17\,\%$$

$$2020-2021: \frac{145-129}{129}\cdot 100 = 12{,}4\,\%$$

De esta forma, el incremento medio anual de las ventas es

$$r = \sqrt[4]{0{,}1778\cdot 0{,}0849\cdot 0{,}1217\cdot 0{,}124} = 12{,}28\,\%,$$

y podemos convencernos de la validez del empleo de esta media teniendo en cuenta que en el período analizado, si queremos promediar razones, habrá que hallar r de modo que

$$90r^4 = 90\cdot 0{,}1778\cdot 0{,}0849\cdot 0{,}1217\cdot 0{,}124.$$

■

- **Media armónica** (H): se suele utilizar para promediar variables que se expresan como el cociente de dos magnitudes, como por ejemplo para promediar productividades, velocidades, tiempos, tipos de cambio y, en general medidas relativas.

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

Al calcular los inversos de los datos, a esta media le influyen mucho los valores pequeños. Además, no puede calcularse en distribuciones que presenten algún valor igual a cero.

Ejemplo 1.13. *La siguiente tabla recoge el número de llamadas atendidas por una sucursal bancaria al cabo de los días laborales de una semana, así como el horario de atención al público de cada día. ¿Cuántas llamadas se atienden por hora en la sucursal?*

Día	*Lunes*	*Martes*	*Miércoles*	*Jueves*	*Viernes*
Número de llamadas	*90*	*100*	*80*	*105*	*60*
Ventas	*8h.-13h.*	*8h.-13h.*	*8h.-13h.*	*8h.-13h.*	*9h.-12h.*

Solución. Nos están pidiendo la «velocidad» con la que se atiende al cliente en la sucursal. N es el total de llamadas atendidas a lo largo de la semana, esto es, $N = 90+100+80+105+60 = 435$ llamadas. Adicionalmente, la velocidad con la que se atienden las llamadas cada día es el cociente entre el número de llamadas y la cantidad de horas que se atiende al público vía telefónica. Por tanto, el valor a hallar será la media armónica que sigue:

$$H = \frac{435}{\frac{90}{5}+\frac{100}{5}+\frac{80}{5}+\frac{105}{5}+\frac{60}{3}} = \frac{87}{19} \approx 4{,}58 \text{ llamadas/hora.}$$

■

- **Mediana** (M_e): es el valor (perteneciente o no a la serie de datos observados) que divide a los valores de la variable, ordenadas de menor a mayor, en dos partes iguales, es decir, es el valor de la variable que se sitúa justo en el centro de las observaciones. Por tanto, por debajo y por encima de la mediana hay un 50 % de la distribución.

  Su propia definición pone de manifiesto que es una medida estadística que tiene sentido calcularse en aquellas variables cuyos valores admitan ordenación, es decir, en variables cualitativas ordinales o en variables cuantitativas. En los casos en los que hay un número impar de observaciones la mediana es única y se corresponde con el valor que ocupa la posición $\frac{N+1}{2}$. En cambio, en el caso de que haya un número par de observaciones, se tendrán dos valores centrales, los que ocupan la posición $\frac{N}{2}$ y $\frac{N}{2}+1$ y, en estos casos, con objeto de establecer la mediana en un único valor, se toma como mediana la media aritmética de los dos valores que ocupan esas posiciones centrales. Evidentemente, si los valores que ocupan esas dos posiciones centrales son en realidad el mismo valor de la variable, la mediana será única y se corresponderá con dicho valor.

  La mediana presenta la ventaja de que, al depender de las observaciones por su orden y no por su valor, no influyen en ella los valores extremos (a diferencia de lo que ocurre con la media aritmética). Por tanto, es una medida de tendencia central apropiada cuando existen valores atípicos (anormalmente bajos o elevados).

Ejemplo 1.14. *Calcular la mediana de los siguientes datos:*

a) 1, 3, 5, 6, 7, 8, 12.

b) 9, 2, 5, 3, 6, 8.

c) 1, 4, 6, 5, 4, 3, 2, 2, 4, 7, 9.

Solución. *a*) Es conveniente notar que los datos están ordenados de menor a mayor en este primer apartado. Dado que hay un total de 7 datos, el central es el número 6, que será el valor de la mediana.

b) Primeramente, debemos ordenar los datos de menor a mayor: 2, 3, 5, 6, 8, 9. Como hay un número par de datos, se considera que la mediana es la media aritmética de los dos más centrales, esto es, $\frac{5+6}{2} = 5{,}5$.

c) Al ordenar los datos de menor a mayor, se obtiene 1, 2, 2, 3, 4, 4, 4, 5, 6, 7, 9. Como hay un total de 11 datos, la mediana será el valor que ocupa la posición central, es decir, 4.

■

- **Moda** (M_o): es el valor de la variable que más veces se repite, es decir, el que presenta mayor frecuencia. Por tanto, se suele referirse a él como el valor más frecuente o valor más habitual. De la propia definición se deduce que **la moda no tiene por qué ser única**, pues en una distribución de frecuencias **puede darse el caso de que haya más de un valor con máxima frecuencia**. Según el número de modas que presente una variable, hablaremos de distribuciones unimodales, bimodales, trimodales,...

  En el caso de que la distribución presente valores extremos, la moda es una buena elección ya que no es sensible a dichos valores. Sin embargo, presenta el inconveniente de que sí es sensible a la fluctuación de las observaciones: un cambio en un único valor puede hacer variar la moda de manera importante.

  Observación 1.2. *La moda es la única medida descriptiva que tiene sentido calcularse para cualquier tipo de variables (cualitativas y cuantitativas).*

  Ejemplo 1.15. *Obtener la moda de la distribución del Ejemplo 1.4.*

  La tabla de frecuencias asociada a esta variable estadística es

x_i	n_i	N_i	f_i	F_i
1	1	1	0,1	0,1
5	2	3	0,2	0,3
7	4	7	0,4	0,7
9	2	9	0,2	0,9
13	1	10	0,1	1
	10		1	

Por tanto, es sencillo concluir que la moda es 7, pues es el valor con mayor frecuencia absoluta (o relativa).

Ejemplo 1.16. *Calcular la moda de la distribución estadística dada por la siguiente tabla:*

x_i	n_i
1	1
5	2
7	3
9	3
13	1

Estamos ante una distribución bimodal, puesto que hay dos valores cuya frecuencia absoluta es máxima (3). Luego, las dos modas son 7 y 9.

1.3.2. Medidas de posición: cuantiles

En ocasiones nos interesa resumir los valores de la variables en un valor que indique una determinada posición, que no necesariamente sea la posición central. Los cuantiles son valores de la distribución que la dividen en partes iguales, es decir, en intervalos que comprenden el mismo número de valores, cuando previamente se han ordenado los datos de menor a mayor. Se clasifican en distintos tipos dependiendo del número de intervalos en que dividen a la población. Los cuantiles más frecuentes son cuartiles, deciles y percentiles.

- **Cuartiles**: son 3 valores (que denotaremos por Q_1, Q_2 y Q_3) que dividen la distribución en cuatro partes iguales, es decir, en 4 intervalos de modo que en cada uno de esos intervalos hay un 25 % de distribución.

  En consecuencia:

  - Q_1 es el valor de la variable por debajo del cual hay un 25 % de distribución

y por encima deja el 75 % restante.

- Q_2 es el valor por debajo del cual hay un 50 % de distribución y por encima deja el 50 % restante.

- Q_3 es el valor por debajo del cual hay un 75 % de distribución y por encima deja el 25 % restante.

Observación 1.3. $Q_2 = M_e$.

- **Deciles**: son 9 valores (que denotaremos por $D_1, D_2, \ldots, D_9$) que dividen la distribución en 10 partes iguales, es decir, en 10 intervalos de modo que en cada uno de esos intervalos hay un 10 % de distribución.

  En consecuencia:

  - D_1 es el valor por debajo del cual hay un 10 % de distribución y por encima deja el 90 % restante.

  - D_2 es el valor por debajo del cual hay un 20 % de distribución y por encima deja el 80 % restante.

  $\vdots$

  - D_9 es el valor por debajo del cual hay un 90 % de distribución y por encima deja el 10 % restante.

  Observación 1.4. $D_5 = Q_2 = M_e$.

- **Percentiles**: Son 99 valores (que denotaremos por $P_1, P_2, \ldots, P_{99}$) que dividen la distribución en 100 partes iguales, es decir, en 100 intervalos de modo que en cada uno de esos intervalos hay un 1 % de distribución.

  En consecuencia:

  - P_1 es el valor por debajo del cual hay un 1 % de distribución y por encima deja el 99 % restante.

  - P_2 es el valor por debajo del cual hay un 2 % de distribución y por encima deja el 98 % restante.

  $\vdots$

  - P_{99} es el valor por debajo del cual hay un 99 % de distribución y por encima deja el 1 % restante.

  Observación 1.5. ∘ $P_{25} = Q_1$, $P_{50} = Q_2 = M_e$ *y* $P_{75} = Q_3$.

  ∘ $P_{10} = D_1, P_{20} = D_2, \ldots, P_{90} = D_9$.

Al igual que sucede con la mediana, en los casos en los que hay un número impar de observaciones el cuantil es un único valor, mientras que en el caso de disponer de número par de observaciones el cuantil no será único, se corresponderá con dos valores, tomándose como cuantil la media aritmética de dichos valores.

Ejemplo 1.17. *A continuación se muestran algunos cuantiles calculados para cada conjunto de datos:*

a) *Primer cuartil* (Q_1)*:* $0, 0, \boldsymbol{1}, 2, 3, 3, 3, 3, 4, 5, 5, \Rightarrow Q_1 = 1$.

b) *Primer cuartil* (Q_1)*:* $0, 0, \boldsymbol{1,2}, 3, 4, 4, 5, 6, 6, 7, 8 \Rightarrow Q_1 = 1{,}5$.

c) *Sexto decil* (D_6)*:* $0, 0, 1, 2, 3, 3, \boldsymbol{3}, 3, 4, 5, 5, \Rightarrow D_6 = 3$.

d) *Sexto decil* (D_6)*:* $0, 0, 3, 4, 4, \boldsymbol{5,6}, 6, 7, 8, \Rightarrow D_6 = 5{,}5$.

e) *Percentil 80* (P_{80})*:* $0, 0, 1, 2, 3, 3, 3, 3, \boldsymbol{4}, 5, 5, \Rightarrow P_{80} = 4$.

f) *Percentil 80* (P_{80})*:* $0, 0, 3, 4, 4, 5, 6, \boldsymbol{6,7}, 8, \Rightarrow P_{80} = 6{,}5$.

Ejemplo 1.18. *Si consideramos la variable «edad» de los estudiantes de un grupo de Estadística, nos podría interesar, por ejemplo, determinar:*

- *la edad máxima del 25 % de los estudiantes más jóvenes: cuartil 1 (o percentil 25).*
- *la edad mínima del 25 % de los estudiantes más mayores: cuartil 3 (o percentil 75).*
- *la edad máxima del 40 % de los estudiantes más jóvenes: decil 4 (o percentil 40).*
- *la edad mínima del 15 % de los estudiantes más mayores: percentil 85.*

1.3.3. Medidas de dispersión

Con las medidas de centralización (media, mediana y moda) resumíamos la tabla de frecuencias en un único valor, para hacernos una idea global de entre qué valores se mueve nuestra variable de estudio. Ahora bien, el siguiente paso será estudiar hasta qué punto esas medidas son representativas o no como síntesis de la información. Si todos los valores que toma la variable están cercanos a la medida de centralización calculada, entonces dicha medida será representativa.

Las medidas de dispersión miden el grado de variabilidad de los datos, independientemente de su causa. Indican la mayor o menor concentración de los datos con respecto a la medida de tendencia central calculada. Las medidas de dispersión se clasifican en:

absolutas (vienen afectadas por unidad de medida) o relativas (no viene afectadas por unidad de medida). Existen numerosas medidas de dispersión, aunque las más utilizadas son las siguientes:

Medidas de dispersión absolutas

1. **Recorrido intercuartílico** (R_I): es la diferencia entre el tercer y el primer cuartil, es decir, $R_I = P_{75} - P_{25}$. Nos indica que en un intervalo de longitud R_I se encuentra el 50 % de los valores centrales. Por tanto, no es una medida de dispersión respecto a una medida de tendencia central concreta, sino que proporciona información global sobre la homogeneidad o heterogeneidad de los valores de la variable.

   Atendiendo a su definición, si R_I es pequeño podemos intuir una pequeña dispersión entre los valores de la variable; por el contrario, si R_I es grande, significa que los datos tienden a distribuirse ampliamente.

   Su principal ventaja es que es una medida de dispersión que no se ve afectada por la existencia de valores atípicos, ya que se fija exclusivamente en la amplitud del 50 % central de la distribución.

   El recorrido intercuartílico suele representarse gráficamente a través del **diagrama de caja y bigotes** (**boxplot**):

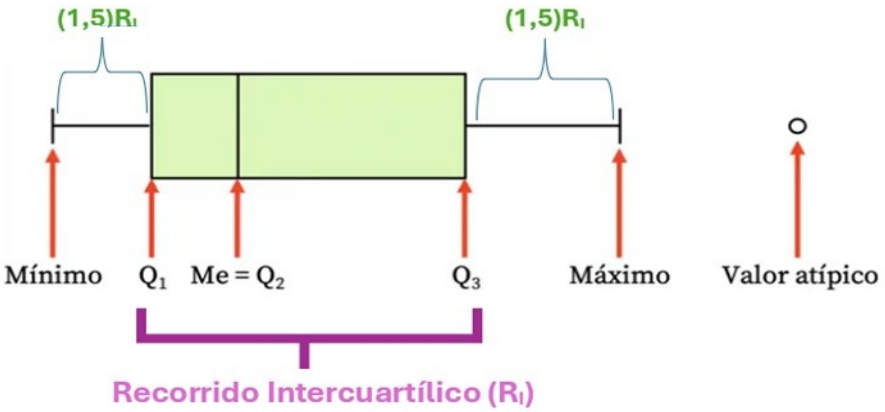

Figura 1.5

   - El dibujo de la caja empieza en el Q_1 y termina en el Q_3.
   - La longitud de la caja es el recorrido intercuartílico ($R_I = Q_3 - Q_1$).

- La línea que divide la caja se corresponde con la mediana (Q_2).
- El bigote de la izquierda se extiende desde el Q_1 hasta el valor menor observado que no supere la longitud $1{,}5(Q_3 - Q_1)$.
- El bigote de la derecha se extiende desde Q_3 hasta el valor mayor observado que no supere la longitud $1{,}5(Q_3 - Q_1)$.
- Los valores que quedan fuera de los bigotes son valores atípicos.

2. **Varianza** (S_X^2): se define como la media de los cuadrados de las de las diferencias de los N valores que toma la variable respecto a su media aritmética.

$$S_X^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{N}.$$

Atendiendo a la definición anterior, podemos observar que la varianza **nunca puede ser negativa** y solo valdrá cero cuando todos los valores de la variable sean iguales y, por lo tanto, la media es ese valor.
Se utiliza para medir la mayor o menor dispersión de los valores respecto de la media $\bar{x}$, de modo que:

- la varianza será pequeña si la mayor parte de las observaciones caen cerca de la media. Por tanto, un valor pequeño indica que la media es representativa como síntesis de la información.
- cuanto más grande sea el valor de la varianza, más dispersión hay de los datos entorno a la media. Por tanto, un valor grande de la varianza indica que la media no es representativa como síntesis de la información.

Es importante aclarar que no hay ningún criterio que nos permita decir si el valor obtenido de la varianza se considera grande o pequeño. Según lo que mida la variable, razonaremos si dicho valor se considera excesivo o no en comparación con el rango de valores de la variable. **Inconveniente**: su problema es que no viene expresada en las mismas unidades de medida que la variable, ya que vendrá dada en las unidades correspondientes pero elevadas al cuadrado. Esto dificulta su interpretación y hace necesario definir otra medida más sencilla para ver si la media es o no representativa.

3. **Desviación típica** (S_X): es la raíz cuadrada, **con signo positivo**, de la varianza.

$$S_X = +\sqrt{S_X^2}.$$

De este modo, conseguimos una medida de dispersión para la media, que sí viene expresada en la misma unidad de medida de la variable de estudio. Por tanto, **la**

desviación típica se convierte en la mejor medida para ver si la media es o no representativa como síntesis de la información. La interpretación es similar a la ya mencionada para la varianza:

- Un valor pequeño de la desviación típica significa que la media es representativa.
- Un valor grande de la desviación típica indica que la media no es representativa y, por tanto, es conveniente resumir la información con otra medida de tendencia central más apropiada.

Propiedad de la desviación típica: Si conocemos la desviación típica de una variable cuantitativa X, (S_X), y se nos define una nueva variable cuantitativa Y de la forma $Y = a + b \cdot X$, con $a, b \in \mathbb{R}$, entonces la desviación típica de la variable Y viene dada por $S_Y = |b| \cdot S_X$. A la desviación típica le afectan, por tanto, solo los cambios de escala.

Ejemplo 1.19. *Calcular la desviación típica y el recorrido intercuartílico de los datos del Ejemplo 1.4 e interpretar los resultados.*

Solución. Recordemos que los estudiantes faltaban, por término medio, 7 días a clase ($\bar{x} = 7$). Para ver si dicha media es o no representativa como síntesis de la información, lo idóneo es calcular la desviación típica, que se puede hallar como la raíz cuadrada de la varianza, que es

$$S_X^2 = \frac{(1-7)^2 + 2 \cdot (5-7)^2 + 4 \cdot (7-7)^2 + 2 \cdot (9-7)^2 + (13-7)^2}{10} = \frac{88}{10} = 8{,}8.$$

En consecuencia,

$$S_X = +\sqrt{8{,}8} \approx 2{,}97 \text{ días de ausencia.}$$

Por otro lado, se calcula el recorrido intercuartílico como $R_I = P_{75} - P_{25}$. Nótese que los datos ordenados de menor a mayor son 1, 5, 5, 7, 7, 7, 7, 9, 9, 13, así que si calculamos la cuarta parte del total de datos, obtenemos $\frac{10}{4} = 2{,}5$, posición entre el primer y segundo 5, así que $P_{25} = 5$. Un razonamiento similar nos conduce a que $P_{75} = 9$, así que $R_I = 9 - 5 = 4$. Esto significa que en un intervalo de longitud 4 se encuentra el 50 % central de los valores. ■

Medidas de dispersión relativas

Hay casos en los que tenemos que comparar poblaciones en las que las unidades de medida son distintas o que, aún teniendo la misma unidad de medida, difieren en sus magnitudes. Esta situación se nos presenta cuando tenemos que comparar, por ejemplo, la dispersión del peso y la altura en los estudiantes de un centro educativo o si queremos comparar la dispersión en las alturas de una población de caballos y otra de ratones. Para los casos anteriores necesitamos una medida de la dispersión en la que no influyan las unidades, sería conveniente tener una medida adimensional. Como tanto la varianza como la desviación típica dependen de la unidad de medida de la variable, no podemos utilizarlas para comparar la homogeneidad de dos o más variables diferentes o para comparar subgrupos de una variable en los que las medias son bastante diferentes. Para esto utilizaremos medidas de dispersión relativas, que no poseen unidades. La medida de dispersión relativa más utilizada es el *coeficiente de variación de Pearson:*

1. **Coeficiente de variación de Pearson** (V_X): se define como el cociente entre la desviación típica y el valor absoluto de la media aritmética.

$$V_X = \frac{S_X}{|\bar{x}|}.$$

   También puede expresarse en porcentaje, multiplicando por 100 la expresión anterior. De la definición se puede deducir lo siguiente:

   - Como las unidades de medida de la desviación típica y la media aritmética son las mismas, este coeficiente es adimensional y, por tanto, útil para comparar varias distribuciones.
   - Representa el número de veces que la desviación típica contiene a la media. En consecuencia, cuanto mayor sea este coeficiente, peor será la representatividad de la media como síntesis de la información y, por tanto, menor será también la homogeneidad de los valores de la distribución.
   - No podrá calcularse cuando la media aritmética valga cero.

   En consecuencia, cuando queramos comparar entre un grupo de variables cuál de ellas tiene media más representativa, debemos calcular para cada una de ellas su coeficiente de variación de Pearson, de modo que aquella que tenga menor valor de dicho coeficiente, será la variable cuya media es más representativa.

Ejemplo 1.20. *Para comparar los rendimientos entre empresas españolas y extranjeras de un mismo sector, se seleccionaron 50 empresas semejantes, obteniéndose*

	Beneficio medio	*Desviación típica*
Españolas	*166000 €*	*120000 €*
Extranjeras	*13000 €*	*1732,05 €*

¿En qué grupo el rendimiento medio es más representativo?

Solución. Para comparar la representatividad de la media aritmética de distintas distribuciones resulta ideal el empleo del coeficiente de variación de Pearson. La tabla ya nos proporciona los valores de las medias y las correspondientes desviaciones típicas, así que los coeficientes de variación de Pearson referidos a empresas españolas y extranjeras son, respectivamente,

$$CV_{\text{esp}} = \frac{120000}{1166000} = \frac{60}{583} \approx 0{,}1029 \quad \text{y} \quad CV_{\text{ext}} = \frac{1732{,}05}{13000} \approx 0{,}1332.$$

Como $CV_{\text{ext}} > CV_{\text{esp}}$, podemos afirmar que el beneficio medio es más representativo en la distribución del rendimiento de las empresas españolas. ■

1.4. Variable tipificada

Relacionado con los conceptos que se han definido en los apartados anteriores, aparece la definición de tipificación (estandarización o normalización) de la variable. La tipificación no es una medida de dispersión, sino que se trata de un procedimiento estadístico que permite comparar datos procedentes de muestras o poblaciones diferentes. Para cualquier variable cuantitativa, conociendo su media y su desviación típica, se pueden transformar sus valores en una nueva escala de medida completamente estandarizada o tipificada. Esta nueva escala se basa en medir la posición relativa que ocupa cada valor dentro de su distribución, entendida como la distancia a la media en unidades de desviación típica.

Supongamos que x es un valor de una variable X procedente de una muestra (o población) con media $\bar{x}$ y desviación típica S_X. En tal caso, el valor de x en unidades tipificadas (que se suele representar por z), se define de la siguiente manera:

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

Las unidades tipificadas muestran el número de desviaciones típicas en que un valor dado se sitúa por encima o debajo de la media de su muestra o población. Una variable tipificada tiene una media nula y una desviación típica igual a uno. Así obtenemos:

- datos independientes de la unidad, o de la escala escogida,
- variables que tienen misma dispersión y misma media.

De este modo, la variable tipificada permite realizar comparaciones entre valores de distintas distribuciones cuando estas tienen medias y varianzas diferentes.

Ejemplo 1.21. *Una persona tiene que escoger entre dos ofertas de trabajo: una propuesta por una empresa española y otra propuesta por una empresa americana. La empresa española le ofrece un sueldo anual bruto de 53000 euros, mientras que la oferta de la americana es de 50000 dolares. Por otro lado, esta persona tiene información sobre el sueldo medio y la desviación típica salarial de las distribuciones de ambas empresas:*

$$\overline{x} = 40000 \text{ euros}, \ S_X = 3500 \text{ euros}, \ \overline{y} = 36500 \text{ dólares}, \ S_Y = 1725 \text{ dólares}.$$

¿En cuál de las dos empresas la posición relativa de esta persona es mejor respecto a los demás trabajadores?

Solución. La respuesta a esta pregunta se puede confeccionar usando la tipificación. La idea es analizar si los 53000 euros mensuales que ganaría esta persona en la empresa española tienen más valor que los 50000 dólares que conseguiría en la americana. Los resultados del procedimiento son

$$z_x = \frac{53000 - 40000}{3500} = \frac{26}{7} \approx 3{,}71 \quad \text{y} \quad z_y = \frac{50000 - 36500}{1725} = \frac{180}{23} \approx 7{,}83,$$

así que la posición de esta persona será mejor con respecto al resto de trabajadores si opta por la oferta de trabajo en la empresa americana. ■