# 3. Estadística descriptiva unidimensional

## 3.1. Variables estadísticas

En el capítulo anterior hemos visto cómo seleccionar los individuos de la población que forman la muestra, de forma que podamos tener una gran seguridad de que esta muestra sea representativa. Una vez que hemos determinado los individuos de la muestra, el siguiente paso es realizar el experimento sobre estos individuos para recabar información.

Para cada uno de los individuos seleccionados en el muestreo observaremos una o varias **características** o **variables estadísticas**. Por ejemplo, son variables estadísticas el peso, la altura, el color de ojos, ... Denotaremos las variables estadísticas mediante letras mayúsculas $X, Y, Z, ...$ Lo que se intenta con la estadística descriptiva es describir los datos que se han obtenido. En este capítulo estudiaremos el caso en el que se observa una sola variable, mientras que en el próximo estudiaremos el caso en el que se observan varias variables.

Las variables estadísticas se clasifican según sus posibles valores en varios tipos:

> **Cualitativas.** Son aquellas que no toman valores numéricos. Por ejemplo, el color de ojos es una característica cualitativa. A su vez, las variables cualitativas pueden ser de dos tipos:
>
> - **Nominales.** Aquellas variables que ni siquiera admiten una ordenación en sus valores. Por ejemplo, el color de ojos es una característica cualitativa nominal.

https://dx.doi.org/10.5209/docm.003.03

- **Ordinales.** Son aquellas en las que, a pesar de tomar valores no numéricos, sus valores admiten una ordenación natural. Por ejemplo, la nota obtenida en una asignatura (suspenso, aprobado, notable, sobresaliente, matrícula de honor) es una característica cualitativa ordinal.

- **Cuantitativas.** Son aquellas que se expresan numéricamente. Por ejemplo, el tiempo que tarda un alumno en hacer un programa informático sería una variable cuantitativa. A su vez, las variables cuantitativas se dividen en dos tipos:

  - **Discretas.** Son aquellas que toman valores numéricos aislados. Otra forma alternativa de definir este tipo de variables es como aquellas que se pueden medir con exactitud. Es interesante notar que el conjunto de valores posibles puede ser finito o infinito. Veamos un par de ejemplos:
    - El resultado del lanzamiento de un dado es una variable discreta que toma un número finito de valores (1, 2, 3, 4, 5, 6).
    - El número de veces que hay que lanzar una moneda hasta obtener la primera cara es una característica discreta que toma infinitos valores, pues su número no puede acotarse superiormente. Así, los posibles resultados de esta variable son 1, 2, 3, ...
  - **Continuas.** Son aquellas que toman cualquier valor dentro de un intervalo (finito o infinito). Por ejemplo, la altura de los alumnos de una clase es una variable continua, pues puede tomar todos los valores en el intervalo [1.40, 2.40]. Otra forma de ver una variable continua es tener en cuenta que los valores que se miden nunca son exactos, sino aproximaciones. Así, la altura de los alumnos no es discreta, pues aunque se aproxime la altura a centímetros, la *verdadera* altura será cualquier número dentro del intervalo.

El siguiente esquema representa la clasificación de los distintos tipos de variables.

$$variables \begin{cases} cualitativas & \begin{cases} nominales \\ ordinales \end{cases} \\ cuantitativas & \begin{cases} discretas \\ continuas \end{cases} \end{cases}$$

Esta clasificación también atiende, como veremos más adelante, al tipo de información que se puede obtener de los datos.

# 3.2. Representaciones tabulares

## 3.2.1. Frecuencias y modalidades

Supongamos ahora que ya hemos realizado el experimento y tenemos unos datos.

**Ejemplo 14.**

*Se está estudiando el número de crías de una camada de conejos. Aquí la variable (lo que se va a medir cada vez que se realice el experimento) es el número de crías de cada camada, que es una variable discreta.*

*Se observan* 35 *camadas seleccionados al azar, anotándose el número de crías en cada camada. Estos resultados son los que aparecen en la tabla 3.1.*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 3 | 4 | 6 | 0 | 1 | 2 | 3 | 4 | 6 |
| 1 | 2 | 3 | 4 | 6 | 2 | 3 | 4 | 6 | 2 | 3 | 4 |
| 6 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | |

**Tabla 3.1.** Resultados obtenidos para el ejemplo 14.

El resultado del experimento $i$-ésimo lo denotaremos por $x_i$. De esta manera, $x_1 = 1, x_2 = 2, x_3 = 0$, y así sucesivamente. En nuestro ejemplo, tenemos 35 datos, $x_1, ..., x_{35}$.

Es interesante notar que los resultados del experimento están influidos por el azar. Así, si repetimos el experimento con otras 35 camadas,

los resultados obtenidos serían distintos (por ejemplo, podría obtenerse algún valor 5), *aunque es de esperar que sean similares*, ya que nosotros esperamos que la muestra sea representativa.

Recordemos que el número de veces que se realiza el experimento se llama **tamaño de muestra** y se denota por $n$. En nuestro caso, $n = 35$.

Como se explicó en el primer tema, el objetivo de la E. Descriptiva es obtener información sobre el comportamiento del fenómeno aleatorio a partir de los datos muestrales. En primer lugar vamos a representar los datos muestrales en una tabla para así tener una visión más clara del comportamiento de la variable.

Fijada nuestra variable, el primer concepto importante es el de **modalidad**. Una modalidad es cada uno de los valores *distintos* que aparecen en la muestra. En nuestro ejemplo, las modalidades son 0, 1, 2, 3, 4, 6. Si tenemos $k$ modalidades distintas, se denotan también por $x_1, ..., x_k$. Es importante no confundir modalidad y dato; en nuestro ejemplo tenemos 6 modalidades $x_1, ..., x_6$ y 35 datos $x_1, ..., x_{35}$. Si tenemos datos numéricos u ordinales, supondremos siempre que $x_1 < x_2 < ... < x_k$. De esta manera, la primera modalidad es $x_1 = 0$, la segunda es $x_2 = 1$, la tercera $x_3 = 2$, y así sucesivamente.

El primer valor lógico para dar información sobre la muestra es contar el número de veces que se repite cada modalidad. De esta forma se da una idea de lo importante (por frecuente) que es cada una de ellas. El número de veces que se repite una modalidad $x_i$ se llama **frecuencia absoluta** y se denota por $n_i$. Nótese que la suma de todas las frecuencias absolutas es $n$, es decir,

$$n_1 + ... + n_k = n.$$

**Ejemplo 15.** *(Continuación del ejemplo 14)*
*En nuestro ejemplo se tienen los resultados de la tabla 3.2.*

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ | 2 | 3 | 10 | 10 | 5 | 5 |

**Tabla 3.2.** Frecuencias absolutas correspondientes al ejemplo 14.

La frecuencia absoluta no da una información clara a no ser que se disponga de todas las frecuencias, pues depende del tamaño de muestra. Así, la modalidad 2 tiene una frecuencia de 10, lo que hace a 2 una modalidad muy importante en una muestra de tamaño 35, pero sería un valor poco importante si hubiésemos realizado el experimento 1000 veces. Para evitar este problema se define la **frecuencia relativa**, que no es más que la proporción de individuos de la muestra que toman una modalidad concreta. La frecuencia relativa de una modalidad $x_i$ se denota por $f_i$ y se calcula dividiendo la correspondiente frecuencia absoluta entre el tamaño de muestra:

$$\boxed{f_i = \frac{n_i}{n}.}$$

Nótese que la suma de todas las frecuencias relativas es 1, pues

$$f_1 + ... + f_k = \frac{n_1}{n} + ... + \frac{n_k}{n} = \frac{n}{n} = 1.$$

**Ejemplo 16.** *(Continuación del ejemplo 14)*
*En nuestro ejemplo se tienen los resultados de la tabla 3.3*

| $x_i$ | $f_i$ |
|---|---|
| 0 | $\frac{2}{35} \approx 0{,}057$ |
| 1 | $\frac{3}{35} \approx 0{,}085$ |
| 2 | $\frac{10}{35} \approx 0{,}286$ |
| 3 | $\frac{10}{35} \approx 0{,}286$ |
| 4 | $\frac{5}{35} \approx 0{,}143$ |
| 6 | $\frac{5}{35} \approx 0{,}143$ |

**Tabla 3.3.** Frecuencias relativas correspondientes al ejemplo 14.

*De esta forma la modalidad 2 tiene una frecuencia relativa de 0.286, que es una frecuencia alta sin necesidad de conocer el tamaño muestral o las frecuencias relativas de otras modalidades.*

Existen otros dos tipos de frecuencias, conocidas como frecuencias acumuladas:

Se llama **frecuencia absoluta acumulada** para una modalidad $x_i$ a la suma de las frecuencias absolutas de las modalidades con valor menor o igual a $x_i$. Se denota por $N_i$. Nótese que el valor correspondiente a $x_k$ (la modalidad con mayor valor) ha de ser $n$.

**Ejemplo 17.** *(Continuación del ejemplo 14)*
*En nuestro ejemplo se tienen los resultados de la tabla 3.4.*

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|---|
| $N_i$ | 2 | 5 | 15 | 25 | 30 | 35 |

**Tabla 3.4.** Frecuencias absolutas acumuladas correspondientes al ejemplo 14.

Se llama **frecuencia relativa acumulada** para una modalidad $x_i$ al cociente entre la frecuencia absoluta acumulada y $n$. Se denota por $F_i$. Nótese que el valor correspondiente a $x_k$ (la modalidad con mayor valor) ha de ser 1.

**Ejemplo 18.** *(Continuación del ejemplo 14)*
*En nuestro ejemplo se tienen los resultados de la tabla 3.5.*

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|---|
| $F_i$ | $\frac{2}{35}$ | $\frac{5}{35}$ | $\frac{15}{35}$ | $\frac{25}{35}$ | $\frac{30}{35}$ | $\frac{35}{35}$ |

**Tabla 3.5.** Frecuencias relativas acumuladas correspondientes al ejemplo 14.

Estas dos últimas medidas necesitan un orden en los valores de la variable, por lo que no son válidas para variables nominales (pero sí para variables ordinales aunque no tomen valores numéricos). Estas medidas nos permiten determinar el número de datos o el porcentaje de los mismos que están por debajo de un valor de la variable fijado, de manera que dan una medida de lo grande que es un determinado dato en comparación con los otros datos de la muestra. Serán muy útiles en el cálculo de las medidas de posición que veremos más adelante.

Estas cuatro medidas se suelen representar en una tabla dando lugar a lo que se conoce como **representación tabular de los datos** o **tabla de frecuencias**.

**Ejemplo 19.** *(Continuación del ejemplo 14)*

*En nuestro ejemplo se tiene que la tabla de frecuencias correspondiente uniendo todas las tablas anteriores sería la que aparece en la tabla 3.6.*

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | 2 | $\frac{2}{35}$ | 2 | $\frac{2}{35}$ |
| 1 | 3 | $\frac{3}{35}$ | 5 | $\frac{5}{35}$ |
| 2 | 10 | $\frac{10}{35}$ | 15 | $\frac{15}{35}$ |
| 3 | 10 | $\frac{10}{35}$ | 25 | $\frac{25}{35}$ |
| 4 | 5 | $\frac{5}{35}$ | 30 | $\frac{30}{35}$ |
| 6 | 5 | $\frac{5}{35}$ | 35 | $\frac{35}{35}$ |

**Tabla 3.6.** Tabla de frecuencias correspondiente al ejemplo 14.

*Nótese que los valores de las frecuencias acumuladas se pueden obtener sumando en diagonal con la columna de las frecuencias. Es decir,*

$$F_i = F_{i-1} + f_i,\ N_i = N_{i-1} + n_i.$$

*Así, por ejemplo,*

$$F_3 = F_2 + f_3,\ N_4 = N_3 + n_4.$$

## 3.2.2. Agrupamiento en clases

En muchas ocasiones, cuando tratamos con variables cuantitativas, casi todas las modalidades tienen frecuencia 1. Esto hace que no tenga sentido construir la tabla de frecuencias, pues el conocimiento de las modalidades proporcionaría prácticamente la misma información. Esta situación aparece por ejemplo con las variables continuas, en las que el conjunto de posibles valores de la variable es infinito. Por ejemplo, si consideramos la variable tiempo de vida una bacteria, y observamos el tiempo de vida para una muestra de tamaño $n$, es casi seguro que

obtendremos $n$ valores diferentes. Otra situación en la que es muy posible obtener una proporción grande de modalidades con frecuencia 1 aparece cuando tenemos una variable discreta que toma muchos valores y el tamaño de muestra es pequeño en comparación con este número. Por ejemplo, si estamos considerando como variable la nota obtenida en un examen y necesariamente esta nota es múltiplo de 0.25, entonces tenemos 41 modalidades distintas como máximo. Si ahora tomamos una muestra de tamaño 50, es de esperar que muchas de las modalidades tengan frecuencia 1.

Otro problema relacionado con el anterior que puede aparecer es que al haber muchas modalidades, la tabla de frecuencias no da una idea clara del funcionamiento de la variable.

**Ejemplo 20.**

*Consideremos la variable $X$ dada por el gasto farmacéutico de una persona en un mes. Elegidas al azar 50 personas se obtienen los datos de la tabla 3.7.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 13.73 | 20.93 | 10.49 | 17.82 | 9.48 | 12.21 | 37.51 | 12.88 | 15.04 |
| 24.47 | 6.62 | 10.86 | 11.84 | 81.84 | 27.61 | 12.11 | 11.99 | 13.98 |
| 12.39 | 12.30 | 10.74 | 11.15 | 18.72 | 18.92 | 13.64 | 34.13 | 26.30 |
| 15.10 | 22.85 | 15.89 | 29.83 | 18.83 | 12.16 | 14.89 | 4.83 | 6.11 |
| 5.71 | 9.78 | 6.63 | 17.79 | 33.26 | 18.49 | 12.87 | 9.54 | 10.69 |
| 12.09 | 11.85 | 11.85 | 11.85 | 5.30 | | | | |

**Tabla 3.7.** Resultados obtenidos para el ejemplo 20.

*En este caso, a pesar de que la variable es discreta, tiene muchas modalidades y la tabla de frecuencias no es tan informativa como en el caso anterior, tal y como se puede ver en la tabla 3.8.*

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| *4.83* | *1* | *0.02* | *1* | *0.02* |
| *5.3* | *1* | *0.02* | *2* | *0.04* |
| *5.71* | *1* | *0.02* | *3* | *0.06* |
| *6.11* | *1* | *0.02* | *4* | *0.08* |
| *6.62* | *1* | *0.02* | *5* | *0.10* |

*Sigue en la página siguiente*

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 6.63 | 1 | 0.02 | 6 | 0.12 |
| 9.48 | 1 | 0.02 | 7 | 0.14 |
| 9.54 | 1 | 0.02 | 8 | 0.16 |
| 9.78 | 1 | 0.02 | 9 | 0.18 |
| 10.49 | 1 | 0.02 | 10 | 0.20 |
| 10.69 | 1 | 0.02 | 11 | 0.22 |
| 10.74 | 1 | 0.02 | 12 | 0.24 |
| 10.86 | 1 | 0.02 | 13 | 0.26 |
| 11.15 | 1 | 0.02 | 14 | 0.28 |
| 11.84 | 1 | 0.02 | 15 | 0.30 |
| 11.85 | 3 | 0.06 | 18 | 0.36 |
| 11.99 | 1 | 0.02 | 19 | 0.38 |
| 12.09 | 1 | 0.02 | 20 | 0.40 |
| 12.11 | 1 | 0.02 | 21 | 0.42 |
| 12.16 | 1 | 0.02 | 22 | 0.44 |
| 12.21 | 1 | 0.02 | 23 | 0.46 |
| 12.30 | 1 | 0.02 | 24 | 0.48 |
| 12.39 | 1 | 0.02 | 25 | 0.50 |
| 12.87 | 1 | 0.02 | 26 | 0.52 |
| 12.88 | 1 | 0.02 | 27 | 0.54 |
| 13.64 | 1 | 0.02 | 28 | 0.56 |
| 13.73 | 1 | 0.02 | 29 | 0.58 |
| 13.98 | 1 | 0.02 | 30 | 0.60 |
| 14.89 | 1 | 0.02 | 31 | 0.62 |
| 15.04 | 1 | 0.02 | 32 | 0.64 |
| 15.10 | 1 | 0.02 | 33 | 0.66 |
| 15.89 | 1 | 0.02 | 34 | 0.68 |
| 17.79 | 1 | 0.02 | 35 | 0.70 |
| 17.82 | 1 | 0.02 | 36 | 0.72 |
| 18.49 | 1 | 0.02 | 37 | 0.74 |
| 18.72 | 1 | 0.02 | 38 | 0.76 |
| 18.83 | 1 | 0.02 | 39 | 0.78 |
| 18.92 | 1 | 0.02 | 40 | 0.80 |
| 20.93 | 1 | 0.02 | 41 | 0.82 |
| 22.85 | 1 | 0.02 | 42 | 0.84 |

*Sigue en la página siguiente*

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| *24.47* | *1* | *0.02* | *43* | *0.86* |
| *26.30* | *1* | *0.02* | *44* | *0.88* |
| *27.61* | *1* | *0.02* | *45* | *0.90* |
| *29.83* | *1* | *0.02* | *46* | *0.92* |
| *33.26* | *1* | *0.02* | *47* | *0.94* |
| *34.13* | *1* | *0.02* | *48* | *0.96* |
| *37.51* | *1* | *0.02* | *49* | *0.98* |
| *81.84* | *1* | *0.02* | *50* | *1.00* |

**Tabla 3.8.** Tabla de frecuencias correspondiente al ejemplo 20 con los datos sin agrupar.

Para evitar estos problemas se agrupan las distintas modalidades en intervalos o **clases**. Estos intervalos son consecutivos y el extremo superior coincide con el extremo inferior del siguiente. También suelen tener todos la misma amplitud, aunque en ocasiones esto no es así, especialmente si hay regiones amplias con muy pocos datos y regiones estrechas con muchos datos. Por ejemplo, en datos económicos las clases suelen ser de distinta amplitud. Denotaremos el extremo inferior del intervalo $i$-ésimo por $a_i$ y el extremo superior por $b_i$. Lo que realmente se está haciendo con el agrupamiento en clases es sustituir el valor real de la variable por la clase en la que está, de manera que las modalidades pasan a ser las clases.

**Ejemplo 21.** *(Continuación del ejemplo 20)*
*En el ejemplo anterior, podemos considerar 10 clases:*

$$(4,8),(8,12),(12,16),(16,20),(20,24),$$
$$(24,28),(28,32),(32,36),(36,40),(40,100).$$

*Así, el primer individuo de la muestra (13.77) se asigna a la tercera clase ya que 13.77 está en el intervalo (12-16), y así sucesivamente.*
*Utilizando clases, se obtendría la tabla de frecuencias de la tabla 3.9.*

Nótese que el objetivo que se perseguía con las tablas de frecuencias, es decir, el dar una idea del funcionamiento de la variable, se consigue mejor mediante el agrupamiento en clases que trabajando con los datos

| $(a_i, b_i)$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| $4-8$ | 6 | 0,12 | 6 | 0,12 |
| $8-12$ | 13 | 0,26 | 19 | 0,38 |
| $12-16$ | 15 | 0,30 | 34 | 0,68 |
| $16-20$ | 6 | 0,12 | 40 | 0,80 |
| $20-24$ | 2 | 0,04 | 42 | 0,82 |
| $24-28$ | 3 | 0,06 | 45 | 0,90 |
| $28-32$ | 1 | 0,02 | 46 | 0,92 |
| $32-36$ | 2 | 0,04 | 48 | 0,96 |
| $36-40$ | 1 | 0,02 | 49 | 0,98 |
| $40-100$ | 1 | 0,02 | 50 | 1,00 |

**Tabla 3.9.** Tabla de frecuencias correspondiente al ejemplo 20 con los datos agrupados en clases.

directamente, ya que permite detectar regiones de la recta real que tienen una mayor frecuencia. Así, parece que hay muchos datos entre 8 y 16; esta información es más difícil de extraer de la tabla con datos sin agrupar.

Como contrapartida, al agrupar en clases prescindimos del valor real del dato para quedarnos con el intervalo en el que está; es decir, no usaremos el verdadero valor del dato. Al no utilizar ninguna información sobre el valor real de los datos que están en la clase, supondremos que estos están repartidos uniformemente en la clase, es decir, que no hay tendencia a que los datos estén concentrados cerca de alguno de los extremos del intervalo. Es por ello que al considerar los extremos de un intervalo, es deseable que los datos contenidos en él se repartan de la manera más uniforme posible. Resumiendo, al agrupar los datos en clases se está perdiendo información en aras de la simplicidad y visibilidad.

¿Cuántas clases es adecuado considerar? Si tenemos muchas clases, entonces los intervalos son pequeños y los verdaderos valores de los datos no se alejan mucho del punto medio del intervalo; sin embargo, con esto se pierde en cierto sentido la utilidad del agrupamiento en clases. Por otra parte, si tenemos muy pocas clases se pierde mucha información. Por ello, es importante determinar el número de clases que

se van a tomar; esta es una decisión subjetiva y depende del problema en cuestión, aunque suele funcionar bien un entero cercano a $\sqrt{n}$.

## 3.3. Representaciones gráficas

El objetivo de las representaciones gráficas es dar una idea del comportamiento de la variable mediante una figura. Por ello, nos interesa esencialmente la forma de la gráfica y no la escala de la misma. Existen muchas representaciones gráficas; aquí sólo veremos las más generales dependiendo del tipo de variable utilizada: diagrama de sectores, diagrama de barras e histograma. Estudiaremos también la poligonal de frecuencias acumuladas, que nos será de utilidad posteriormente. Finalmente, cuando estudiemos las medidas de posición presentaremos el diagrama de cajas.

### 3.3.1. Diagrama de sectores

Esta representación consiste en dividir un círculo en tantos sectores como modalidades, de forma que a cada modalidad se le asigna un sector circular de área proporcional a la frecuencia (absoluta o relativa) de la modalidad.

Como dado un círculo de radio $R$ el área del sector circular de $s$ grados es

$$Area = \pi R^2 \frac{s}{360} = cte \cdot s,$$

podemos reformular la definición anterior y asignar a cada modalidad un sector circular con un número de grados proporcional a la frecuencia.

Nótese que debe dividirse completamente el círculo, luego la suma de los grados de los diferentes sectores ha de ser 360. Por ello, no podemos fijar nosotros la constante de proporcionalidad, sino que el número de grados $g_i$ correspondientes a la modalidad $x_i$ se obtiene por una regla de tres:

$$\begin{array}{ccc} n & \longrightarrow & 360 \\ n_i & \longrightarrow & g_i \end{array} \Rightarrow g_i = \frac{n_i \times 360}{n}.$$

| Comunidad Autónoma | Población ($n_i$) | Porcentaje ($100\, f_i$) |
|---|---|---|
| Andalucía | 8 202 220 | 18.20 |
| Cataluña | 7 364 078 | 16.34 |
| Comunidad de Madrid | 6 271 638 | 13.92 |
| Comunidad Valenciana | 5 029 601 | 11.16 |
| Galicia | 2 784 169 | 6.18 |
| Castilla y León | 2 557 330 | 5.68 |
| Otras | 12 851 042 | 28.52 |

**Tabla 3.10.** Comunidad de residencia en 2008.

El diagrama de sectores puede calcularse para cualquier tipo de variable estadística, aunque suele aplicarse a variables cualitativas; por ejemplo, es la representación habitual de los resultados en las encuestas de opinión.

**Ejemplo 22.**

*Consideremos por ejemplo la variable $X \equiv$ Comunidad autónoma de residencia. Los datos de 1 de enero de 2008 proporcionan la información de la tabla 3.10.*

*En este caso el correspondiente diagrama de sectores viene dado en la figura 3.1.*

## 3.3.2. Diagrama de barras

Esta representación se aplica a variables discretas que no estén agrupadas en clases. Es una representación en el plano. En el eje de abscisas se sitúan las modalidades de la variable, y sobre cada uno de estos valores se dibuja una barra de altura proporcional a la frecuencia de la modalidad. Es decir, para la modalidad $x_i$, la altura correspondiente $h_i$ sería

$$h_i = kn_i,$$

donde $k$ es una contante de proporcionalidad (que fijamos nosotros).

**Ejemplo 23.** *(Continuación del ejemplo 14)*

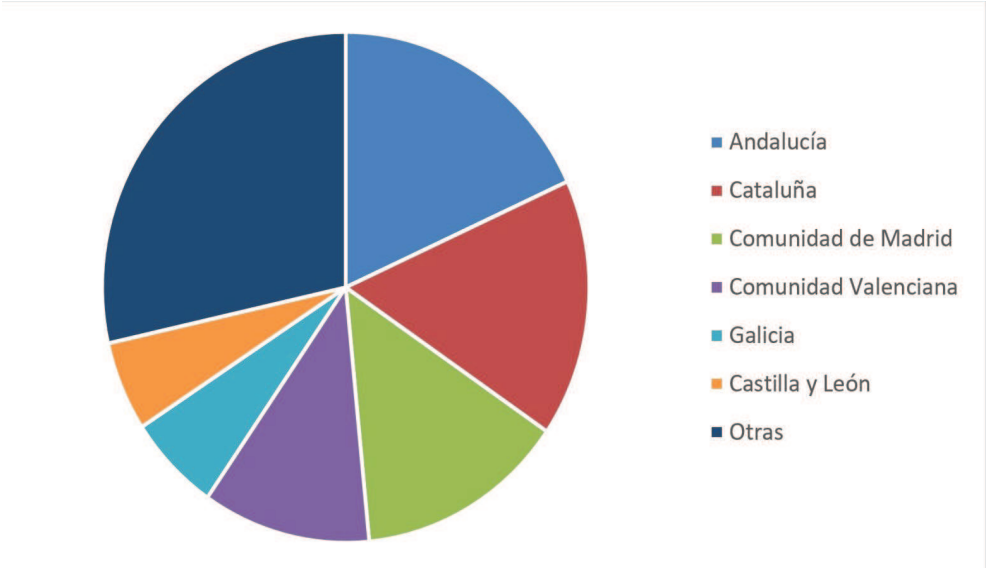


**Figura 3.1.** Diagrama de sectores correspondiente a los datos de la tabla 3.10.

*En el ejemplo 14 el diagrama de barras (con constante de proporcionalidad 1) es el que aparece en la figura 3.2.*

### 3.3.3. Histograma

Esta representación se aplica a variables agrupadas en clases. Como en el caso del diagrama de barras, es una representación en el plano. En el eje de abscisas se sitúan las distintas clases, y sobre cada uno de estos intervalos se dibuja un rectángulo con ÁREA proporcional a la frecuencia de la clase. Es decir, para la clase $(a_i, b_i)$ se tiene que el área $A_i$ del correspondiente rectángulo sería

$$A_i = kn_i,$$

donde $k$ es una constante de proporcionalidad fijada por nosotros. Y como por otra parte $A_i = h_i(b_i - a_i)$, entonces

$$h_i = \frac{kn_i}{b_i - a_i}.$$

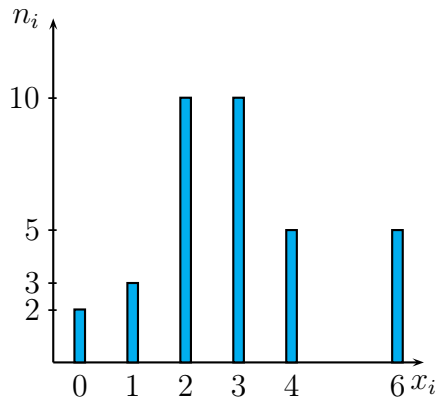


**Figura 3.2.** Diagrama de barras correspondiente a los datos del ejemplo 14.

Si todas las clases tienen la misma amplitud, es decir, $b_i - a_i = b_j - a_j$ para cualesquiera $i$ y $j$, entonces podemos dibujar el histograma dibujando rectángulos con altura proporcional a la frecuencia de clase. Pero si las amplitudes no son iguales para todos los intervalos, no podemos hacer esta simplificación. Esto es lo que ocurre en el siguiente ejemplo.

**Ejemplo 24.** *(Continuación del ejemplo 20)*

*Para el ejemplo 20 el histograma puede verse en la figura 3.3.*

*Es interesante comprender cómo se han calculado las alturas de los rectángulos, especialmente el correspondiente a la clase (40, 100). Hemos fijado como constante de proporcionalidad 4, de manera que el área de cada rectángulo es cuatro veces su frecuencia. Para los otros intervalos esto implica que la altura coincide con la frecuencia absoluta puesto que su amplitud es 4. Para (40, 100), tenemos una frecuencia de 1, con lo que el área debe ser* $4 \times 1 = 4$, *y por otra parte este valor es el producto de la base (60) por la altura (h). De esta forma* $h = 4/60 = 0{,}06$.

## 3.3.4. Polígono de frecuencias acumuladas

Esta representación se aplica también a variables agrupadas en clase. No da una idea tan clara como el histograma, pero nos será muy útil cuando queramos calcular medidas de posición.

El polígono de frecuencias acumuladas es la gráfica de la función

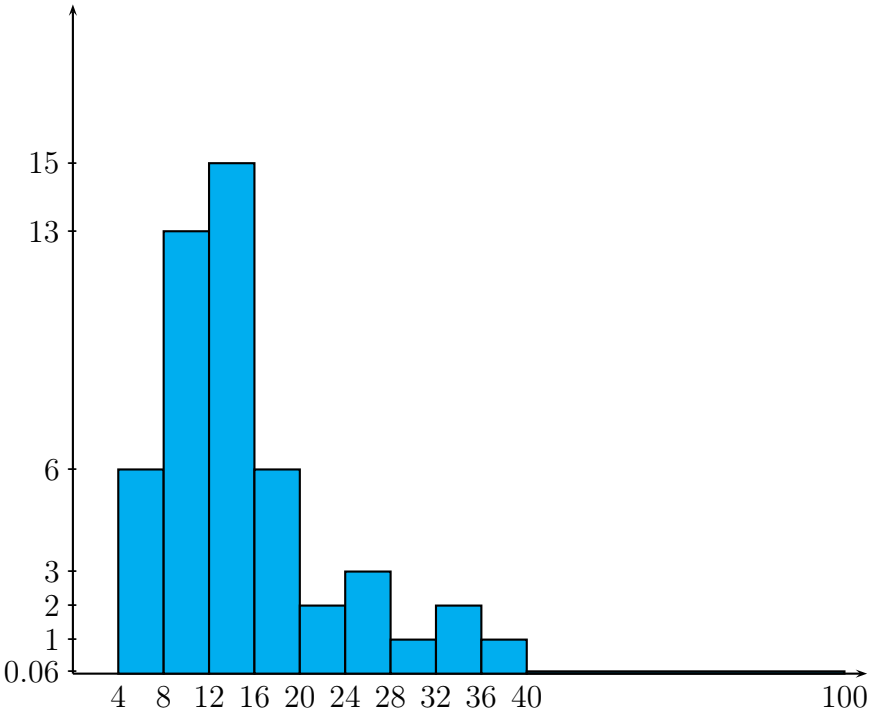


**Figura 3.3.** Histograma correspondiente a los datos del ejemplo 20.

que a cada valor real le asigna la proporción (o también el número) de individuos cuyo valor es menor o igual al considerado, es decir, la frecuencia acumulada para ese valor.

Veamos cómo se construye: Como los valores están agrupados en clases hemos perdido los verdaderos valores; sin embargo, sí es posible calcular los valores del número de datos menores o iguales para los extremos del intervalo. Por ejemplo, para el ejemplo 20, el número de individuos cuyo valor es menor o igual a 8 es 6, el número de individuos cuyo valor es menor o igual a 12 es 19, etc. Nótese que estos valores son los valores de la correspondiente frecuencia absoluta acumulada. También conocemos con exactitud la función para los valores menores que el extremo inferior del primer intervalo (vale 0) y para el extremo superior del último intervalo (vale 50). Esto mismo se puede hacer con las frecuencias relativas acumuladas. Así, la proporción de datos con valor menor o igual que 8 es 0.12, la proporción de datos con valor

menor o igual que 12 es 0.38, etcétera.

Quedan por determinar los otros valores. Para ello usaremos el hecho de que es de esperar que todos los valores dentro de cada clase se distribuyan de manera uniforme. Y la uniformidad en términos de la representación se traduce en una línea recta que una los dos puntos correspondientes al extremo inferior y al extremo superior. De esta manera, podemos aproximar el valor de la función mediante una secuencia de segmentos rectilíneos.

**Ejemplo 25.** *(Continuación del ejemplo 20)*

*En este caso el polígono de frecuencias acumuladas viene dado por la figura 3.4.*

## 3.4. Medidas de centralización

Una vez que tenemos nuestra tabla de frecuencias y la correspondiente representación gráfica, nuestro siguiente objetivo es dar un valor representativo de nuestra variable estadística. Esto vamos a hacerlo mediante las llamadas **medidas de centralización** o **medidas de tendencia central**. Entenderemos por medida de tendencia central un valor numérico obtenido a partir de los valores de la variable estadística que tenga un determinado significado, de modo que de acuerdo con algún criterio, los datos de la variable estadística oscilan frente a él.

Existen muchas medidas de tendencia central. Nosotros estudiaremos tres: la moda, la media (aritmética) y la mediana.

### 3.4.1. La moda

Se define la **moda** como aquel valor con mayor frecuencia relativa. Se denota por $mo$. Es interesante notar que la moda no tiene que ser necesariamente única.

**Ejemplo 26.** *(Continuación del ejemplo 14)*

*En nuestro ejemplo de las crías de conejo se tiene* $mo = 2$ *o* $3$.

La moda no requiere ningún tipo de condición sobre la variable estadística. Es decir, puede calcularse para cualquier tipo de variable, sea cuantitativa o cualitativa.

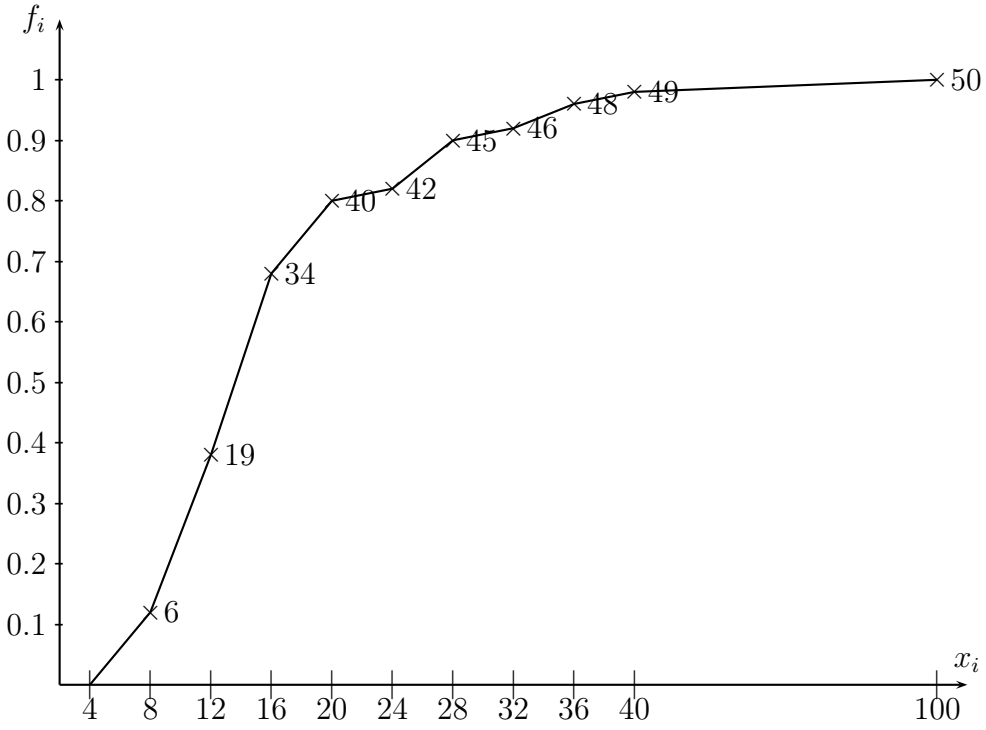


**Figura 3.4.** Poligonal de frecuencias correspondiente al ejemplo 20. En los puntos de la poligonal aparece la frecuencia absoluta acumulada, que nos da una de las formas de hacer esta representación. En el eje de ordenadas se han utilizado las frecuencias relativas acumuladas, que proporcionan otra manera de hacer la representación. Ambas dan lugar a la misma figura.

## 3.4.2. La media aritmética

Consideremos una variable estadística *cuantitativa* $X$, de la que una muestra de tamaño $n$ ha obtenido los datos $x_1, ..., x_n$. Se define la **media aritmética** de la variable $X$, denotada $\overline{x}$, como

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.}$$

**Ejemplo 27.** *(Continuación del ejemplo 14)*
*Para nuestro ejemplo de las crías de conejo se tiene:*

$$\overline{x} = \frac{1+2+0+3+...+2+3}{35} = \frac{103}{35}.$$

La media representa un valor promedio de la variable estadística, es decir, las diferencias por encima y por debajo de la media se compensan. Matemáticamente, esto se escribe

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Por su propia definición, la media es un valor comprendido entre el mínimo valor obtenido y el máximo valor obtenido en la muestra. Finalmente, la media no es necesariamente un valor posible de la variable, como por ejemplo ocurre en el caso anterior, en que la media es 103/35, que no es un número entero.

Supongamos ahora que tenemos la tabla de frecuencias para la variable $X$. En ese caso, tenemos que $X$ toma las modalidades $x_1, ..., x_k$ con frecuencias absolutas $n_1, ..., n_k$. Como hemos visto al tratar las tablas estadísticas, esto significa que $x_1$ aparece $n_1$ veces en la muestra, $x_2$ aparece $n_2$ veces en la muestra y así sucesivamente. Por lo tanto, la suma de todos los valores de la muestra se puede escribir como

$$\sum_{i=1}^{k} x_i n_i,$$

y por lo tanto la media se puede escribir como

$$\boxed{\overline{x} := \frac{\sum_{i=1}^{k} x_i n_i}{n}.}$$

Equivalentemente, si las frecuencias relativas son $f_1, ..., f_k$, y basándonos en que $f_i = \frac{n_i}{n}$, la media aritmética puede calcularse mediante la expresión

$$\boxed{\overline{x} = \sum_{i=1}^{k} x_i f_i.}$$

**Ejemplo 28.** *(Continuación del ejemplo 14)*

*Para nuestro ejemplo de las crías de conejo se tiene:*

$$\overline{x} = \frac{1}{35}[0 \cdot 2 + 1 \cdot 3 + 2 \cdot 10 + 3 \cdot 10 + 4 \cdot 5 + 6 \cdot 5] = \frac{103}{35}.$$

Supongamos ahora que tenemos los datos agrupados en clases. Entonces, las modalidades son intervalos y no valores numéricos. Para poder utilizar la fórmula anterior, tomaremos un representante de clase; este representante será el punto medio del intervalo y se llama la **marca de clase**; denotaremos por $c_i$ la marca de clase del intervalo $(a_i, b_i)$. Entonces,

$$c_i = \frac{a_i + b_i}{2}.$$

La elección de este valor como representante de la clase viene motivada por el hecho de que se ha supuesto que los datos dentro de una clase se reparten uniformemente. Así, si por ejemplo en la clase [0,3] hay dos datos, se supone que serán aproximadamente 1 y 2, o 0.5 y 2.5; más concretamente, esta suposición hace que al sumarlos nos dará el valor 3, o sea, 2 veces la marca de clase, que es 1.5.

De esta manera, la media para una variable de datos agrupados viene dada por:

$$\boxed{\overline{x} := \frac{\sum_{i=1}^{k} c_i n_i}{n} = \sum_{i=1}^{k} c_i f_i.}$$

**Ejemplo 29.** *(Continuación del ejemplo 20)*

*Para el ejemplo de datos agrupados se tienen las marcas de clase de la tabla 3.11.*

*Por lo tanto, la media viene dada por:*

| $a_i - b_i$ | 4-8 | 8-12 | 12-16 | 16-20 | 20-24 |
|---|---|---|---|---|---|
| $c_i$ | 6 | 10 | 14 | 18 | 22 |
| | | | | | |
| $a_i - b_i$ | 24-28 | 28-32 | 32-36 | 36-40 | 40-100 |
| $c_i$ | 26 | 30 | 34 | 38 | 70 |

**Tabla 3.11.** Marcas de clase para la agrupación hecha en el ejemplo 20.

$$\overline{x} = \frac{1}{50}[6 \cdot 6 + 10 \cdot 13 + ... + 38 \cdot 1 + 70 \cdot 1] = 15{,}64.$$

*Si no hubiésemos hecho el agrupamiento en clases, el valor de la media sería*

$$\overline{x} = \frac{13{,}73 + 20{,}93 + ..,11{,}85 + 5{,}30}{50} = 16{,}56,$$

*que es un valor diferente al que se obtuvo al aplicar la fórmula cuando los datos están agrupados en clases. Esta diferencia es debida a que los valores en realidad no se reparten uniformemente dentro de los distintos intervalos, y ese error se traduce en una diferencia entre los dos resultados. Sin embargo, es mucho más rápido hallar la media con los datos agrupados y usando las marcas de clase que utilizando los datos directamente. Esta es una de las razones por las que antes de la aparición de los ordenadores se agrupasen los datos en clases en muchas ocasiones.*

Veamos ahora algunas propiedades de la media. En general, dada una constante $a \in \mathbb{R}$, se tienen las siguientes propiedades:

- Consideremos la aplicación $f : \mathbb{R} \mapsto \mathbb{R}$ definida por $f(x_i) = x_i + a$. Si aplicamos esta transformación a las modalidades de la v. estadística $X$, esto da lugar a una nueva variable, que denotamos por $X + a$, en la que a todos los valores de la muestra se les ha sumado el valor $a$, tal y como se ve en la tabla 3.12.

  Entonces, para la nueva variable $X + a$ se tiene

$$\overline{x + a} = \overline{x} + a.$$

| $X$ | $x_1$ | ... | $x_n$ |
|---|---|---|---|
| $X+a$ | $x_1+a$ | ... | $x_n+a$ |

**Tabla 3.12.** Resultados de la muestra de la variable $X+a$.

Por ejemplo, supongamos que estamos estudiando los sueldos en una empresa y nos sale un sueldo medio de 1500 euros mensuales. Si ahora se decide dar una bonificación extraordinaria a todos los trabajadores de 100 euros, el sueldo medio pasará a ser de 1600 euros, y no será necesario volver a realizar las cuentas con los nuevos ingresos.

- Consideremos la aplicación $f : \mathbb{R} \mapsto \mathbb{R}$ definida por $f(x_i) = a \cdot x_i$. Si aplicamos esta transformación a las modalidades de la v. estadística $X$, esto da lugar a una nueva variable, que denotamos por $aX$, en la que a todos los valores de la muestra se les ha multiplicado por el valor $a$, tal y como se ve en la tabla 3.13.

| $X$ | $x_1$ | ... | $x_n$ |
|---|---|---|---|
| $aX$ | $a \cdot x_1$ | ... | $a \cdot x_n$ |

**Tabla 3.13.** Resultados de la muestra de la variable $aX$.

Entonces, para la $aX$ se tiene

$$\overline{ax} = a\overline{x}.$$

Por ejemplo, supongamos nuevamente que estamos estudiando los sueldos en una empresa y nos sale un sueldo medio de 1500 euros mensuales. Si ahora pasamos el sueldo a dólares y un euro equivale a 1.1 dólares, entonces el sueldo medio pasará a ser de $1{,}1 \times 1500$ dólares, y no será necesario aplicar esta transformación a todos los sueldos para luego volver a calcular la media.

- Consideremos la aplicación $g : \mathbb{R} \mapsto \mathbb{R}$. Si aplicamos esta transformación a las modalidades de la v. estadística $X$, esto da lugar

a una nueva variable, que denotamos por $Y = g(X)$, en la que a cada valor $x_i$ de la muestra ha sido sustituido por $g(x_i)$, tal y como se muestra en la tabla 3.14.

| $X$ | $x_1$ | ... | $x_n$ |
|---|---|---|---|
| $g(X)$ | $g(x_1)$ | ... | $g(x_n)$ |

**Tabla 3.14.** Resultados de la muestra de la variable $Y = g(X)$.

Entonces, para $Y$ se tiene

$$\overline{y} = \sum_{i=1}^{k} g(x_i) f_i.$$

Es interesante notar que $\overline{y} \neq g(\overline{x})$. Solo para casos especiales como los de las dos propiedades anteriores se tendrá la igualdad.

Estas propiedades son también válidas aunque la variable esté agrupada en clases.

## 3.4.3. La mediana

Como hemos visto anteriormente, la media tiene muy buenas propiedades matemáticas. Sin embargo, tiene el inconveniente de que está muy afectada por valores extremos muy grandes o muy pequeños. Por ejemplo, si estamos estudiando los sueldos de una empresa que tiene 99 trabajadores que ganan 1 000 euros al mes y el empresario que gana 101 000 euros al mes, la media aritmética nos dice que el sueldo medio es de 2 000 euros al mes, y este valor no parece muy representativo de lo que está sucediendo realmente en la muestra. Esto es lo que ha ocurrido también con los datos del ejemplo 20, en el que hay un valor que es mucho mayor que los demás y arrastra el valor de la media hacia un valor más alto. Para evitar esta influencia de los valores extremos de la muestra definimos la mediana.

Para ello, tenemos que tener en cuenta que intuitivamente, el valor representativo debería ser un valor que esté más o menos en la mitad

de la muestra si ordenamos los valores. Y el problema que ha ocurrido con la media en los ejemplos del párrafo anterior es que la media deja a casi toda la población por debajo del valor de la media. La idea de la mediana es considerar como valor representativo el valor que está en el medio de los valores de la muestra.

Se define la **mediana**, denotada $me$, como aquel valor que, supuestos los valores $x_1, ..., x_n$ (no las modalidades, puede haber valores repetidos) de la variable ordenados en forma creciente $x_{(1)} \leq ... \leq x_{(n)}$ , deja *al menos* la mitad de los datos de la muestra *por debajo o iguales* que $me$ y *al menos* la mitad de los datos *por encima o iguales* que $me$.

Aunque esta definición parece un tanto extraña, está intentando describir el dato en la posición central de la muestra. Así, si todos los datos son diferentes entre sí, y $n = 7$, entonces la mediana sería el cuarto dato más pequeño. Este dato dejaría con valor menor o igual a 4 datos (más de la mitad, que serían 3.5) y con valor mayor o igual a 4 datos. Y sería el único dato que cumpliese las dos condiciones.

Desde un punto de vista práctico, la mediana se calcula de manera diferente para datos agrupados y para datos sin agrupar, dependiendo en este último caso si tenemos la representación tabular o solo los datos de la muestra.

Comencemos con el caso de datos no agrupados y sin disponer de la representación tabular. Tenemos entonces la muestra de tamaño $n$ dada por $x_1, ..., x_n$. Se calcula la mediana de la siguiente manera:

- En primer lugar, tenemos que ordenar los datos de menor a mayor. Denotaremos por $x_{(1)}$ el menor valor de la muestra, por $x_{(2)}$ el segundo valor más pequeño, y así sucesivamente. Entonces $x_{(n)}$ es el mayor valor de la muestra. Tenemos entonces la muestra ordenada que viene dada por

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

- Buscamos ahora el dato que está en la posición central. Se procede de la siguiente manera:

  - Si $n$ es un número impar, entonces $n = 2p - 1$ y la mediana es el dato que en la muestra ordenada está en la posición $p$.

Por ejemplo, si $n = 7$, entonces $7 = 2 \times 4 - 1$, por lo que $p = 4$ y la mediana es $x_{(4)}$.

- Si $n$ es un número par, entonces $n = 2p$ y la mediana es cualquier valor entre los de los datos que en la muestra ordenada están en las posiciones $p$ y $p + 1$. Cualquier valor en este intervalo sirve como mediana, aunque muchas veces se toma la media de estos dos valores. Por ejemplo, si $n = 8$, entonces $8 = 2 \times 4$, por lo que $p = 4$ y la mediana es cualquier valor en el intervalo $[x_{(4)}, x_{(5)}]$.

Equivalentemente, podemos hallar la mediana siguiendo el procedimiento que se detalla a continuación. En primer lugar, calculamos el valor $n \times 0{,}5$.

- Si este valor es entero, la mediana es cualquier valor en el intervalo $[x_{(n\times 0,5)}, x_{(n\times 0,5)+1}]$.
- Si no es entero, entonces la mediana es el valor del dato en la posición del primer entero que supera a $n \times 0{,}5$, que denotaremos $\lceil n \times 0{,}5 \rceil$. Por ejemplo, si $n = 7$, entonces $n \times 0{,}5 = 3{,}5$ y $\lceil n \times 0{,}5 \rceil = 4$. Así, $me = x_{(\lceil n\times 0,5 \rceil)}$.

Esta última forma de proceder es la que usaremos para calcular las medidas de posición en la próxima sección en el caso de disponer de la muestra pero no de la tabla de frecuencias.

**Ejemplo 30.** *(Continuación del ejemplo 14)*
*En nuestro ejemplo de las crías de conejo la muestra ordenada es*

$$0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3,$$
$$3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 6, 6, 6, 6, 6 \quad .$$

*Como* $n = 35$, *se tiene que* $p = 18$ *y, por tanto,* $me = x_{(18)} = 3$.

Veamos ahora cómo calcular la mediana si tenemos la representación tabular de los datos. En primer lugar buscamos las modalidades que dejan por debajo o con valor igual al menos el 50 % de los datos. Para ello, basta observar la columna de las frecuencias relativas acumuladas y buscar las modalidades cuya frecuencia relativa acumulada sea superior o igual a 0.5. Equivalentemente, podemos buscar las modalidades

cuya frecuencia absoluta acumulada sea superior a $n/2$. Así, buscamos valores con frecuencia acumulada grande. Pasamos ahora a la segunda condición. Entonces debemos tener en cuenta que la proporción de valores mayores o iguales que una modalidad $x_i$ viene dada por $1 - F_{i-1}$, es decir, 1 menos la proporción de valores menores que $x_i$. Entonces, tenemos que buscar las modalidades tales que 1 menos la frecuencia relativa acumulada es mayor o igual que 0.5. Equivalentemente, podemos buscar las modalidades tales que 1 menos la frecuencia absoluta acumulada sea superior a $n/2$. Así, buscamos valores con frecuencia acumulada pequeña.

Tenemos ahora dos situaciones:

- Si el valor 0.5 no aparece en la tabla en la columna de las frecuencias relativas acumuladas, entonces los valores con frecuencia relativa acumulada menores que 0.5 no cumplen la primera condición y se descartan. Para los valores que sí cumplen esta condición, el primero de ellos cumple también la segunda. Para los demás, ninguna de ellos cumple la segunda condición. Así, la mediana es la primera modalidad cuya correspondiente frecuencia relativa acumulada supera el valor 0.5.

- Si el valor 0.5 aparece en la taba en la columna de las frecuencias relativas acumuladas, entonces los valores con frecuencia relativa acumulada menores que 0.5 no cumplen la primera condición y se descartan. Para los valores que sí cumplen esta condición, el primero de ellos cumple también la segunda. El siguiente valor también cumple la segunda condición. Para los demás, ninguno de ellos cumple la segunda condición. Ahora bien, cualquier valor entre esas dos modalidades también cumple las dos condiciones (nótese que en ningún momento se pide que la mediana sea una modalidad). Así, la mediana es cualquier valor entre la modalidad que tiene 0.5 como frecuencia relativa acumulada y el valor de la siguiente modalidad.

Esto mismo puede hacerse con frecuencias absolutas acumuladas, cambiando 0.5 por $n/2$.

**Ejemplo 31.** *(Continuación del ejemplo 14)*

*En nuestro ejemplo ya se había obtenido la representación tabular de los datos, que repetimos en la tabla 3.15.*

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | 2 | $\frac{2}{35}$ | 2 | $\frac{2}{35}$ |
| 1 | 3 | $\frac{3}{35}$ | 5 | $\frac{5}{35}$ |
| 2 | 10 | $\frac{10}{35}$ | 15 | $\frac{15}{35}$ |
| 3 | 10 | $\frac{10}{35}$ | 25 | $\frac{25}{35}$ |
| 4 | 5 | $\frac{5}{35}$ | 30 | $\frac{30}{35}$ |
| 6 | 5 | $\frac{5}{35}$ | 35 | $\frac{35}{35}$ |

**Tabla 3.15.** Representación tabular correspondiente a los datos del ejemplo 14.

*Entonces, puede aplicarse el procedimiento anterior para comprobar que la mediana ha de ser 3.*

En el caso de datos agrupados hemos perdido los valores reales de los datos, por lo que no puede procederse de la misma manera que con los datos sin agrupar. Para el cálculo de la mediana usaremos el polígono de frecuencias acumuladas. Como buscamos un valor que deje el 50 % de los datos por debajo o iguales, una vez dibujado el polígono de frecuencias acumuladas, se dibuja la recta horizontal

$$y = n\frac{50}{100}.$$

Esta recta tiene que cortar a la poligonal en algún momento, pues esta es una gráfica continua y las ordenadas pasan de 0 a 1. Una vez determinado el punto de corte, se mira la abscisa de este punto; esta abscisa es la mediana. Para determinar este valor usaremos semejanza de triángulos.

**Ejemplo 32.** *(Continuación del ejemplo 20)*

*En este caso se tiene la gráfica de la figura 3.5.*

*De esta forma, la mediana es un valor en el intervalo 12-16 y viene dada, aplicando semejanza de triángulos, por:*

$$\frac{BaseTriangGrande}{AlturaTriangGrande} = \frac{BaseTriangPeq}{AlturaTriangPeq}.$$

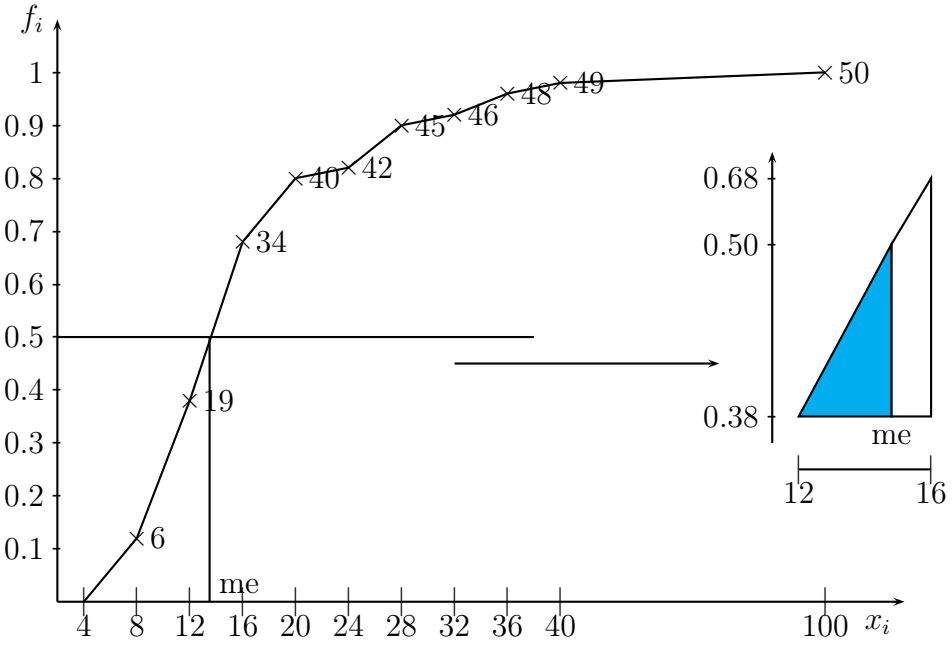


**Figura 3.5.** Procedimiento paso para hallar la mediana en una distribución agrupada en clases.

*O equivalentemente,*

$$\frac{BaseTriangGrande}{BaseTriangPeq} = \frac{AlturaTriangGrande}{AlturaTriangPeq}.$$

*En nuestro caso,*

$$\frac{0{,}68 - 0{,}38}{16 - 12} = \frac{0{,}50 - 0{,}38}{me - 12} \Rightarrow me = 13{,}8.$$

En definitiva, si la mediana está en el intervalo $[a_i, a_{i+1}]$ y la frecuencia de esta clase es $F_i$, la mediana viene dada por

$$\boxed{me = a_i + \frac{0{,}5 - F_{i-1}}{F_i - F_{i-1}}(a_{i+1} - a_i).}$$

Finalmente, al contrario que la media, que siempre necesita variables cuantitativas, la mediana, así como las medidas de posición que veremos a continuación, pueden calcularse para variables ordinales.

# 3.5. Medidas de posición

Análogamente a la mediana, se pueden definir otras medidas conocidas como **medidas de posición**. Las medidas de posición son generalizaciones de la mediana que nos sirven para dar un valor que divida la muestra en dos partes, una parte por encima de dicho valor y otra por debajo, de manera que el número o proporción de datos en cada parte verifique unas condiciones determinadas.

## 3.5.1. Cuartiles, deciles y percentiles

La primera medida de posición que veremos son los cuartiles, denotados $q_i, i = 1, 2, 3$. Se define el **cuartil** $i$ como aquel valor que, supuestos los valores (no las modalidades) de la variable ordenados en forma creciente, deja al menos el $25 \cdot i\,\%$ de los datos de la muestra por debajo o iguales que ese valor y al menos el $(100 - 25 \cdot i)\,\%$ de los datos por encima o iguales que ese valor. En particular, la mediana coincide con $q_2$.

La forma de calcular los cuartiles a partir de una muestra es análoga a la vista para el cálculo de la mediana.

En el caso de tener los datos de la muestra, $x_1, ..., x_n$, tenemos primeramente que ordenar la muestra, de forma que tenemos

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

Ahora, se busca el valor $n \times 0{,}25, n \times 0{,}5$ o $n \times 0{,}75$ para $q_1, q_2$ o $q_3$, respectivamente.

- Si este valor es entero, el primer cuartil es cualquier valor en el intervalo $[x_{(n\times 0{,}25)}, x_{(n\times 0{,}25)+1}]$. Lo mismo se haría para los otros cuartiles.

- Si no es entero, entonces el primer cuartil es el valor del dato en la posición del primer entero que supera a $n \times 0{,}25$, que denotaremos

$\lceil n \times 0{,}25 \rceil$. Por ejemplo, si $n = 7$, entonces $n \times 0{,}25 = 1{,}75$ y $\lceil n \times 0{,}5 \rceil = 2$. Así, $me = x_{(2)}$.

En el caso de variables sin agrupar y si disponemos de la tabla de frecuencias, podemos hallar los cuartiles sustituyendo el valor 0.5 (o $n \times 0{,}5$) por los valores 0.25 ($q_1$), 0.5 ($q_2$) o 0.75 ($q_3$).

**Ejemplo 33.** *(Continuación del ejemplo 14)*

*En este ejemplo ya se había obtenido la representación tabular de los datos. Entonces:*

$$q_1 = 2, q_2 = 3, q_3 = 4.$$

El cálculo para variables agrupadas en intervalos es similar al de la mediana y se calcula a partir del polígono de frecuencias acumuladas, sustituyendo la recta $y = n\frac{50}{100}$ por $y = n\frac{25}{100}$ (para $q_1$), $y = n\frac{50}{100}$ (para $q_2$) o $y = n\frac{75}{100}$ (para $q_3$).

**Ejemplo 34.** *(Continuación del ejemplo 20)*

*En este caso se tienen las gráficas de las figuras 3.6 y 3.7.*

*Luego,*

$$\frac{0{,}38 - 0{,}12}{12 - 8} = \frac{0{,}25 - 0{,}12}{q_1 - 8} \Rightarrow q_1 = 10.$$

*Para hallar $q_3$ se procede de la misma manera.*

$$\frac{0{,}80 - 0{,}68}{20 - 16} = \frac{0{,}75 - 0{,}68}{q_3 - 16} \Rightarrow q_3 = 18{,}33.$$

Se define el **decil** $i$, $i = 1, ..., 9$, denotado $d_i$, como aquel valor que, supuestos los valores de la muestra ordenados de forma creciente, deja al menos el $10 \cdot i\,\%$ de los datos de la muestra por debajo o iguales y al menos el $(100 - 10 \cdot i)\,\%$ de los datos por encima o iguales. En particular, la mediana coincide con $d_5$. Al igual que para los cuartiles, su cálculo es similar al de la mediana.

**Ejemplo 35.** *(Continuación del ejemplo 14)*

*En este caso se tiene por ejemplo $d_4 = 2, d_7 = 3, d_9 = 6$.*

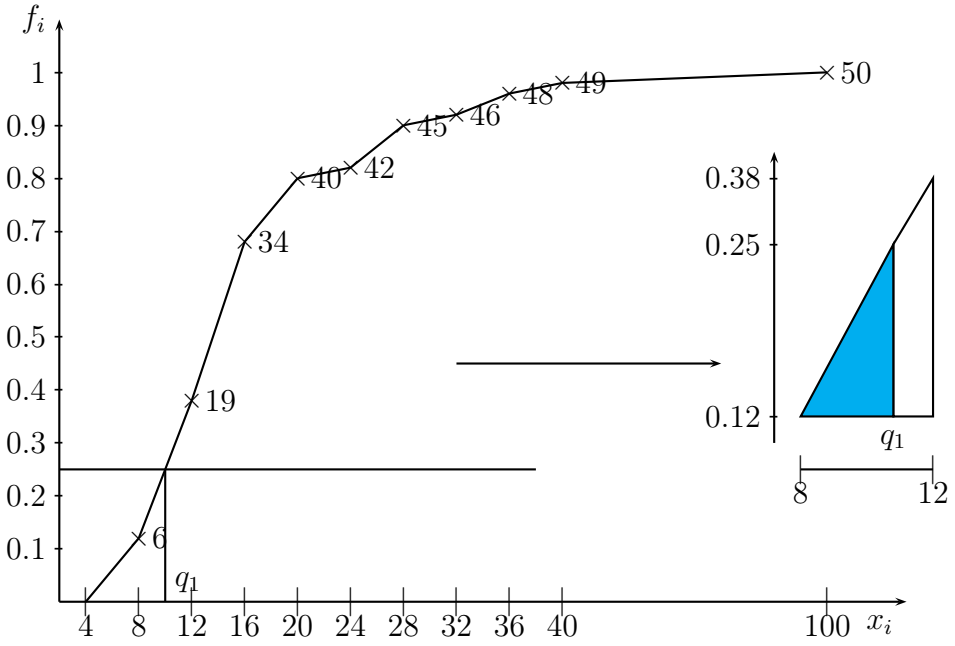


**Figura 3.6.** Ejemplo de cálculo del primer cuartil con datos agrupados en clases.

Finalmente, se define el **percentil** o **cuantil** $i$, $i = 1, ..., 99$, denotado $p_i$, como aquel valor que, supuestos los valores de la muestra ordenados en forma creciente, deja al menos el $i\,\%$ de los datos de la muestra por debajo o iguales y al menos el $(100 - i)\,\%$ de los datos por encima o iguales. En particular, la mediana coincide con $p_{50}$. El cálculo de los percentiles sigue las mismas pautas que el cálculo de la mediana.

**Ejemplo 36.** *(Continuación del ejemplo 14)*
*En este caso se tiene por ejemplo* $p_{18} = 2, p_{94} = 6$.

## 3.5.2. Diagrama de cajas

A partir de las medidas de posición podemos dibujar una nueva representación gráfica llamada **diagrama de cajas**, que está pensada principalmente para distribuciones de datos no agrupadas y sin datos

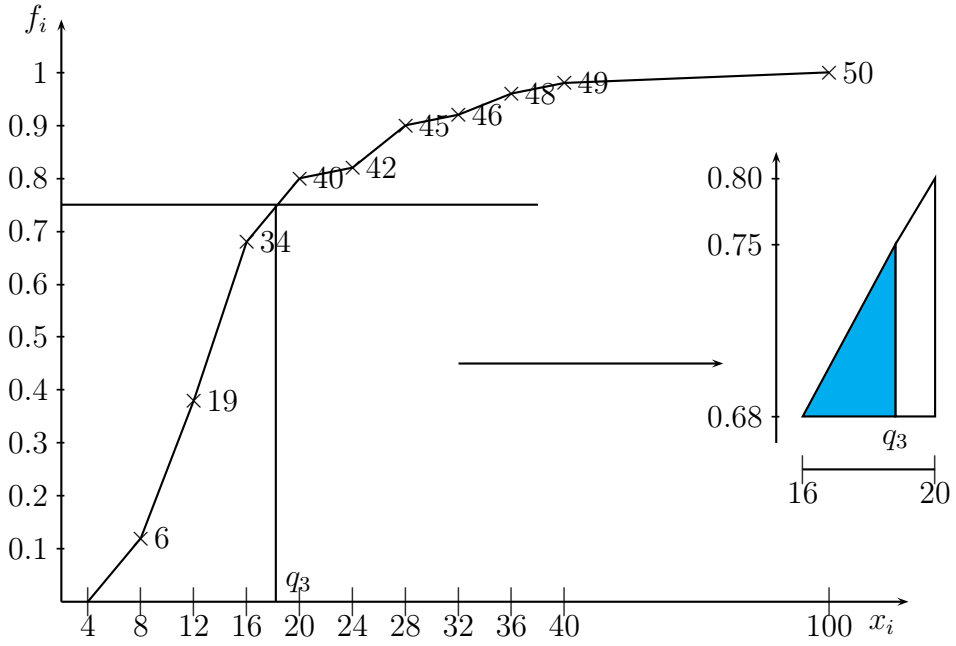


**Figura 3.7.** Ejemplo de cálculo del tercer cuartil con datos agrupados en clases.

repetidos. Esta representación hace hincapié en los valores que se alejan de los valores típicos de la variable. Veamos cómo construir esta gráfica. Tenemos en primer lugar que calcular los tres cuartiles. Los valores de $q_1$ y $q_3$ determinan los límites de una *caja*, en cuyo interior se tiene una línea con el valor de la mediana *me*. Los valores dentro de la caja son los valores considerados típicos para la variable.

De esta caja salen dos líneas (llamadas *bigotes*) que determinan los valores que se consideran normales (no típicos). Los límites de estas líneas son:

$$\text{máx}\{x_{(1)}, q_1 - 1{,}5 \cdot (q_3 - q_1)\} \quad \text{para la línea inferior,}$$

$$\text{máx}\{x_{(n)}, q_3 + 1{,}5 \cdot (q_3 - q_1)\} \quad \text{para la línea inferior.}$$

Si $x_{(1)} < q_1 - 1{,}5 \cdot (q_3 - q_1)$ todavía quedan valores que no están dentro de estas líneas. Estos valores se representan de dos formas distintas:

- Los datos entre $q_1 - 3 \cdot (q_3 - q_1)$ y $q_1 - 1{,}5 \cdot (q_3 - q_1)$ se dibujan con $*$. Estos son los valores considerados raros (por demasiado pequeños) para la muestra.
- Los datos menores que $q_1 - 3 \cdot (q_3 - q_1)$ se dibujan con $\circ$. Estos son valores considerados muy raros para la muestra.

Lo mismo se hace con los valores grandes. Así, si $x_{(n)} > q_3 + 1{,}5 \cdot (q_3 - q_1)$ todavía quedan valores que no están dentro de estas líneas. Estos valores se representan de dos formas distintas:

- Los datos entre $q_3 + 1{,}5 \cdot (q_3 - q_1)$ y $q_3 + 3 \cdot (q_3 - q_1)$ se dibujan con $*$. Estos son los valores considerados raros (por demasiado grandes) para la muestra.
- Los datos mayores que $q_3 + 3 \cdot (q_3 - q_1)$ se dibujan con $\circ$. Estos son valores considerados muy raros para la muestra.

**Ejemplo 37.** *(Continuación del ejemplo 20)*

*Para el caso del gasto farmacéutico del ejemplo 20 sin agrupar en clases se puede ver que* $me = 12{,}63$, *(la media de los valores 25 y 26),* $q_1 = 10{,}86$, *(dato en posición 13) y* $q_3 = 18{,}72$ *(dato en posición 38). Luego* $q_3 - q_1 = 7{,}86$. *Entonces el diagrama de cajas sería el que aparece en la figura 3.8.*

## 3.6. Medidas de dispersión

Las medidas de dispersión son medidas que nos permitirán estudiar lo diferentes que son los datos de la muestra. Esto nos sirve para darnos una idea de hasta qué punto las medidas de tendencia central son representativas o no. Así, si tenemos dos muestras de tamaño 2, una con los valores 0, 100 y otra con los valores 50, 50, se tendría que su media coincide; sin embargo, la media en el segundo caso es muy representativa pues todos los valores de la muestra coinciden con ella, mientras que en el primer caso la media es un valor alejado de todos los datos.

**Figura 3.8.** Ejemplo de diagrama de cajas para los datos del ejemplo 20 sin agrupar en clases. Para los valores pequeños, el límite está en el mínimo dato, mientras que para los valores grandes hay valores raros. Por ello, el bigote para valores pequeños es más corto que el de valores grandes. Hay tres datos que se consideran grandes porque superan el valor $q_3 + 1{,}5(q_3 - q_1) = 30{,}51$, pero no el valor $q_3 + 3(q_3 - q_1) = 42{,}73$ y un valor muy raro (el máximo) que supera este último valor.

Existen muchas medidas de dispersión y todas ellas siguen esta idea intuitiva: medir la separación entre los datos o equivalentemente, la distancia entre los datos y alguna medida de tendencia central. Aquí estudiaremos las más usuales, que son la varianza y la desviación típica.

### 3.6.1. La varianza

Supongamos que tenemos una muestra de valores $x_1, ..., x_n$. La **varianza** de la muestra, denotada $v(x_1, ..., x_n), v$ o $v(X)$, se define como el valor dado por

$$\boxed{v(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.}$$

Si tenemos una variable estadística que toma las modalidades $x_1, ..., x_k$ con frecuencias absolutas $n_1, ..., n_k$ podemos aplicar el mismo razonamiento que se hizo para la media y que nos permite calcular la varianza más rápidamente. La **varianza** de la muestra viene entonces dada por

$$\boxed{v(X) = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{n},}$$

o equivalentemente

$$\boxed{v(X) = \sum_{i=1}^{n} (x_i - \overline{x})^2 f_i}$$

si usamos las frecuencias relativas.

En el caso de tener la variable agrupada en clases, se sustituyen los intervalos por las correspondientes marcas de clase. Así, si tenemos las marcas de clase $c_1, ..., c_k$ con frecuencias absolutas $n_1, ..., n_k$, la varianza viene dada por

$$\boxed{v(X) = \frac{\sum_{i=1}^{k} (c_i - \overline{x})^2 n_i}{n},}$$

o equivalentemente

$$\boxed{v(X) = \sum_{i=1}^{n} (c_i - \overline{x})^2 f_i}$$

si usamos las frecuencias relativas.

Veamos una explicación intuitiva que permita justificar la varianza como medida de dispersión. Empecemos considerando el numerador. En primer lugar, nótese que está basado en las diferencias con la media. Si los datos son poco variables, estarán cerca de la media y por tanto estas diferencias serán pequeñas, mientras que si los datos son muy variables estarán lejos de la media y las diferencias serán grandes. El uso del cuadrado es necesario para que no se produzcan compensaciones entre diferencias positivas y negativas. Finalmente, todo se divide entre el tamaño de muestra; la razón está en conseguir una medida que no esté influida por el número de datos que tenemos. En otro caso, podría pasar que se tuviese una muestra poco variable pero de muchos datos y otra muy variable con pocos datos, ambas con la misma varianza.

La varianza es sin duda la medida de dispersión más utilizada. Para su cálculo, puede demostrarse que

$$\boxed{v(X) = \overline{x^2} - \overline{x}^2,}$$

donde si tenemos la muestra $(x_1, ..., x_n)$, entonces

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n},$$

si tenemos la tabla de frecuencias, entonces

$$\overline{x^2} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} = \sum_{i=1}^{k} x_i^2 f_i,$$

y en el caso de tener la variable agrupada en clases se utiliza la misma expresión, pero sustituyendo las modalidades por las marcas de clase. Es decir

$$\overline{x^2} = \frac{\sum_{i=1}^{k} c_i^2 n_i}{n} = \sum_{i=1}^{k} c_i^2 f_i.$$

Esta expresión para la varianza reduce el número de operaciones necesarias para calcular la varianza.

**Ejemplo 38.** *(Continuación de los ejemplos 14 y 20)*
*En el ejemplo 14:*

$$\overline{x^2} = \frac{0^2 \cdot 2 + 1^2 \cdot 3 + 2^2 \cdot 10 + 3^2 \cdot 10 + 4^2 \cdot 5 + 6^2 \cdot 5}{35} = \frac{393}{35}.$$

*Luego,*

$$v(X) = \frac{393}{35} - (\frac{103}{35})^2 = 2{,}57.$$

*Para el ejemplo 20:*

$$\overline{x^2} = \frac{6^2 \cdot 6 + 10^2 \cdot 13 + ... + 38^2 \cdot 1 + 70^2 \cdot 1}{50} = \frac{20814{,}8877}{50} = 416{,}30.$$

*Luego,*

$$v(X) = 416{,}30 - 15{,}64^2 = 171{,}69.$$

Veamos ahora algunas propiedades de la varianza, que son similares a las vistas para la media.

- Consideremos la aplicación $X + a : \{x_1, ..., x_n\} \mapsto \mathbb{R}$ definida por $(X + a)(x_i) = x_i + a$. Entonces, para la nueva variable $X + a$ se tiene

$$v(X + a) = v(X).$$

  Nótese que este resultado es lógico, pues esta transformación traslada los datos, pero mantiene las diferencias.

- Consideremos la aplicación $aX : \{x_1, ..., x_n\} \mapsto \mathbb{R}$ definida por $(aX)(x_i) = ax_i$. Entonces, para la nueva variable $aX$ se tiene

$$v(aX) = a^2 v(X).$$

  En este caso se está estirando la población, por lo que las diferencias cambian y esto se refleja en la varianza.

- Consideremos la aplicación $g(X) : \{x_1, ..., x_n\} \mapsto \mathbb{R}$ definida por $g(X)(x_i) = g(x_i)$. Es interesante notar que, al igual que sucedía con la media, en general $v(g(X)) \neq g(v(X))$.

- La varianza vale 0 si y sólo si todos los valores de la muestra son iguales.

- Finalmente, nótese que la varianza nunca puede ser negativa, pues es una suma de cuadrados.

### 3.6.2. La desviación típica

La varianza tiene el problema de que sus unidades son las unidades de la variable elevadas al cuadrado. Es por ello que aparece el concepto de **desviación típica**, denotada por $d(x_1, ..., x_n)$, $d(X)$ o $d$, que se define como la raíz cuadrada positiva de la varianza, es decir,

$$\boxed{d(X) := \sqrt{v(X)}.}$$

La desviación típica está medida en las mismas unidades que la variable $X$.

**Ejemplo 39.** *(Continuación de los ejemplos 14 y 20)*
*En el ejemplo 14,* $d(X) = \sqrt{2{,}57} = 1{,}603$. *Para el ejemplo 20,* $d(X) = 13{,}10$.

Veamos ahora la traducción de las propiedades que se vieron para la varianza para el caso de la desviación típica.

- Consideremos la aplicación $X + a : \{x_1, ..., x_n\} \mapsto \mathbb{R}$ definida por $(X + a)(x_i) = x_i + a$. Entonces, para la nueva variable $X + a$ se tiene

$$d(X + a) = d(X).$$

- Consideremos la aplicación $aX : \{x_1, ..., x_n\} \mapsto \mathbb{R}$ definida por $(aX)(x_i) = a \cdot x_i$. Entonces, para la nueva variable $aX$ se tiene

$$d(aX) = |a|d(X).$$

  Es interesante notar que el valor de la constante $a$ pasa a $|a|$.

- La desviación típica vale 0 si y solo si todos los valores de la muestra son iguales.

- La desviación típica no puede ser negativa.

### 3.6.3. El coeficiente de variación de Pearson

Finalmente, consideremos la situación en que queremos comparar las variaciones de dos muestras distintas. En principio, podría pensarse en comparar las varianzas o las desviaciones típicas. Sin embargo, ambas medidas tienen unidades y, por tanto, un cambio de escala llevaría a conclusiones erróneas. Para ilustrar esta situación, consideremos el caso de medir la altura de 15 personas. Supongamos que se ha obtenido una desviación típica de 1cm. Esto implica que la variable $X$ ha sido medida en cm. Sin embargo, si hubiésemos medido $X$ en metros, todos los valores de $X$ aparecerían divididos por 100, con lo que tendríamos

la variable $X/100$, y por las propiedades vistas anteriormente, también la desviación típica aparecería dividida por 100 y valdría 0.01m. Si queremos utilizar la desviación típica para comparar las dispersiones de las dos variables y obviamos las unidades de medida, se tendría que la variable $X$ tendría una variación mayor que $X/100$, mientras que en realidad la dispersión es la misma, pero medida en unidades diferentes. En este caso todavía sería posible hacer una transformación entre las unidades de medida para así realizar una comparación efectiva; sin embargo, en muchos otros casos, esto no es posible.

Para evitar estos problemas sería muy útil tener un coeficiente que no dependiese de unidades (lo que se llama *adimensional*). Esto es lo que se consigue mediante el **coeficiente de variación de Pearson**, que se define como

$$\boxed{cv(X) = \frac{d(X)}{|\overline{x}|}.}$$

Obviamente, este coeficiente no está definido para variables con media nula. Así, una variable se considera más variable que otra si su coeficiente de variación de Pearson es mayor.

**Ejemplo 40.** *(Continuación de los ejemplos 14 y 20)*

*En el ejemplo 14,* $cv(X) = \frac{1{,}603}{2{,}94} = 0{,}545$. *Para el ejemplo 20,* $cv(X) = \frac{13{,}10}{15{,}64} = 0{,}838$.

*Por tanto, los datos están más dispersos en el ejemplo 20.*

## 3.7. Medidas de forma

En las secciones anteriores hemos estudiado las medidas de centralización, posición y dispersión. En esta sección estudiaremos otros dos tipos de medidas conocidas como *medidas de forma*. Como su propio nombre indica, las medidas de forma nos permitirán hacernos una idea de la forma (o la tendencia) que tiene la distribución de frecuencias.

Clasificaremos las medidas de forma en dos tipos: las medidas de asimetría y las medidas de curtosis.

### 3.7.1. Medidas de asimetría

El objetivo de las medidas de asimetría es medir hasta qué punto los datos se reparten de forma especular alrededor de un punto (que será la media) o si hay una tendencia a que los datos con valor superior a la media estén más (o menos) alejados que los datos con valor inferior. En el primer caso diremos que la distribución es *asimétrica por la derecha* y en el segundo caso que es *asimétrica por la izquierda*.

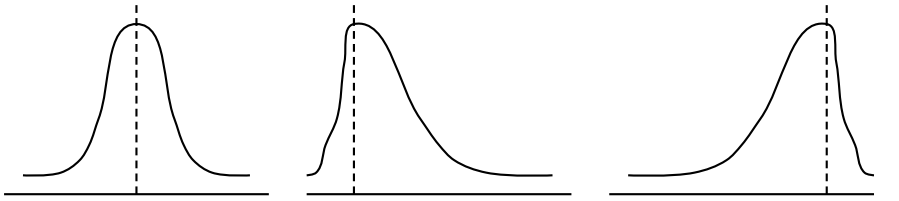


**Figura 3.9.** Tipos de distribuciones en función de su asimetría.

Otra forma un poco más matemática de interpretar la asimetría es la siguiente: Dada una variable estadística, la variable será asimétrica por la derecha si $\overline{x} > me$ y será asimétrica por la izquierda si $\overline{x} < me$.

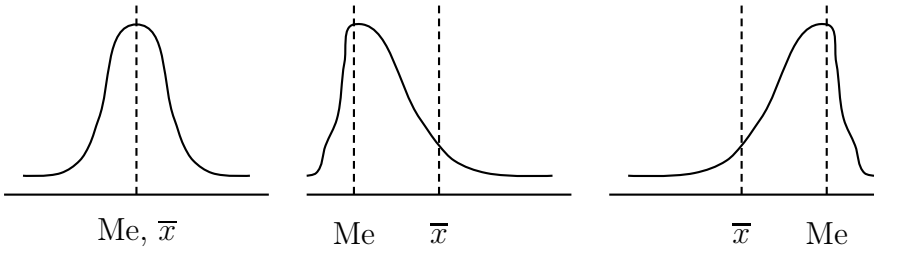


**Figura 3.10.** Comparación media-mediana en función de la asimetría.

Las medidas de asimetría tratan de medir esta tendencia. Al igual que sucedía con las medidas de centralización y de dispersión, existen muchas medidas de asimetría, aunque todas tratan de reflejar el comportamiento anterior. Veremos a continuación una de ellas.

El **coeficiente de asimetría de Fisher** es sin duda la medida de asimetría más utilizada. Dada una variable estadística $X$, este coeficiente se define por

$$\boxed{g_1(X) := \frac{\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})^3}{n}}{d(X)^3} = \frac{\dfrac{\sum_{i=1}^{k}(x_i-\overline{x})^3 n_i}{n}}{d(X)^3} = \frac{\sum_{i=1}^{k}(x_i-\overline{x})^3 f_i}{d(X)^3}.}$$

Hemos puesto la definición en el caso de tener la muestra (primera fórmula) o la tabla de frecuencia con datos no agrupados (segunda fórmula y tercera fórmula). Para distribuciones agrupadas en datos, sustituimos los valores de las modalidades $x_i$ por los correspondientes valores de las marcas de clase. Es decir,

$$\boxed{g_1(X) := \frac{\dfrac{\sum_{i=1}^{k}(c_i-\overline{x})^3 n_i}{n}}{d(X)^3} = \frac{\sum_{i=1}^{k}(c_i-\overline{x})^3 f_i}{d(X)^3}.}$$

La idea de esta medida es la siguiente: Si la distribución es muy simétrica, es de esperar que las diferencias de los datos con la media que sean positivas y negativas se compensen; esto se traduciría en

$$\sum_{i=1}^{n}(x_i-\overline{x}) \approx 0.$$

Sin embargo, esto se tiene siempre (es una de las propiedades de la media). Por ello tenemos que buscar otra manera de representar nuestra idea. Si tomamos las diferencias al cuadrado obtenemos una suma de valores positivos, con lo que no podremos distinguir las diferencias positivas de las negativas. Así, pasamos a elevar al cubo, que ya tiene en cuenta el signo de las diferencias. De esta forma, si hay un dato mucho mayor que la media pero no hay como contrapartida un valor menor en esas condiciones, sino que hay varios valores menores pero cercanos a la media que compensen el dato anterior, entonces tendríamos una

distribución asimétrica por la derecha. Al aplicar la fórmula de la asimetría, se tendría una diferencia muy grande de este dato especial con la media, y cuando se eleva al cubo será todavía más grande. Este valor no se compensa por las varias diferencias pequeñas y negativas cuando se elevan al cubo, con lo que obtendríamos un valor positivo para $g_1(X)$.

El denominador $d(X)^3$ es un término que sirve para relativizar el resultado. Así, si los datos tienen mucha dispersión podría ocurrir que el numerador fuese grande, no debido a diferencias significativas entre valores positivos y negativos, sino a que estos valores son grandes y se están elevando al cubo. Análogamente, si los datos están muy concentrados alrededor de la media, es de esperar que el numerador sea pequeño aunque haya una tendencia a la asimetría. Finalmente, $d(X)$ se eleva al cubo para obtener un coeficiente adimensional.

De esta manera, si $g_1(X) = 0$, concluiremos que la distribución es simétrica, si $g_1(X) > 0$ la distribución será asimétrica por la derecha y si $g_1(X) < 0$ la distribución será asimétrica por la izquierda. Cuanto mayor sea el valor de $g_1(X)$ en valor absoluto más pronunciada será la asimetría.

**Ejemplo 41.**

*Consideremos la distribución $X$ cuyos valores son 0, 1, 2 y 10, todos ellos con frecuencia $\frac{1}{4}$. Esta distribución es claramente asimétrica por la derecha, pues el valor 10 está muy alejado de los otros valores de la distribución. Si hallamos el coeficiente de asimetría de Fisher obtenemos:*

$$\overline{x} = \frac{13}{4} = 3{,}25, v(X) = \frac{267}{16} \Rightarrow d(X) = \frac{\sqrt{267}}{4} = 4{,}08, g_1(X) = 1{,}545.$$

**Ejemplo 42.** *(Continuación de los ejemplos 14 y 20)*

*Si consideramos el ejemplo 14 se obtiene* $\overline{x} = \frac{103}{35} = 2{,}94, d(X) = \sqrt{2{,}57} = 1{,}603$. *El numerador vale*

$$\frac{\sum_{i=1}^{6}(x_i - \overline{x})^3 n_i}{n} = \frac{-31{,}36}{35} = -0{,}896.$$

*Luego $g_1(X) = -0{,}21$, y concluimos que la distribución es un poco asimétrica por la izquierda.*

*Para el ejemplo 20 tenemos* $\overline{x} = 15{,}64, d(X) = 13{,}10$. *El numerador vale*

$$\frac{\sum_{i=1}^{10}(c_i - \overline{x})^3 n_i}{n} = 3664{,}92.$$

*Luego* $g_1(X) = 1{,}63$, *y concluimos que la distribución es bastante asimétrica por la derecha.*

El estudio de la simetría resulta útil cuando ha de establecerse una variable que modele unos datos. En muchas ocasiones se presupone un modelo normal que veremos en el capítulo 6, y esta propiedad no se corresponde con la realidad en algunos casos. Las medidas de simetría, junto con las medidas de curtosis que veremos a continuación, juegan un papel muy importante para validar este modelo normal.

### 3.7.2. Medidas de curtosis o apuntamiento

Las medidas de curtosis miden el «aplastamiento» de la distribución. En otras palabras, miran si hay tendencia a que todas las modalidades tengan la misma frecuencia o si, por el contrario, hay una tendencia a que existan grandes diferencias entre ellas.

El aplastamiento se utiliza casi siempre para comparar los datos de la muestra con los valores teóricos de una normal estándar que estudiaremos en el capítulo 6. Por ello, se estudia la curtosis en distribuciones que son prácticamente simétricas y con una sola moda. Si el aplastamiento de los datos es similar al de la distribución normal estándar se dice que la distribución es *mesocúrtica*. Si es más aplastada diremos que la distribución es *platicúrtica* y si es más apuntada que es *leptocúrtica*.

Al igual que sucedía con otras medidas que hemos visto, existen muchas medidas de curtosis, todas ellas intentando reflejar en un valor este comportamiento, aunque la conocida como medida de curtosis de Pearson es con mucho la más utilizada.

Dada una variable estadística $X$, el **coeficiente de curtosis de Pearson** se define por

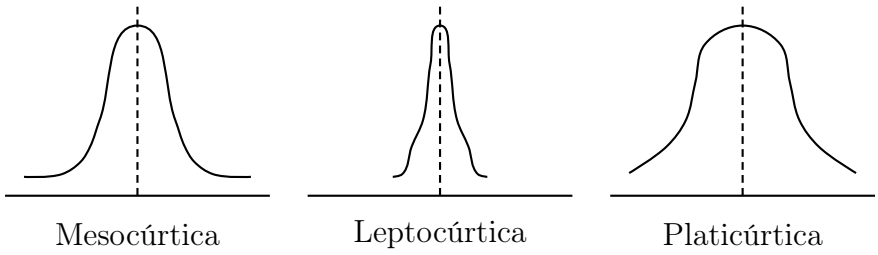


**Figura 3.11.** Interpretación gráfica de las distribuciones según su curtosis.

$$g_2(X) := \frac{\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}}{d(X)^4} - 3.$$

Nótese en primer lugar que, al igual que pasaba con el coeficiente de asimetría de Fisher, este coeficiente es adimensional.

Veamos ahora la idea de esta medida: Si tenemos en mente una distribución simétrica con una sola moda, es fácil ver que cuanto mayor sea el apuntamiento, mayor será la frecuencia de los datos cercanos a la media. Así, el valor del numerador será tanto más pequeño cuanto más apuntada sea la distribución. De la misma manera, el valor de $d(X)$ se hará más pequeño cuanto más apuntada sea la distribución. Sin embargo, puede demostrarse que el cociente entre el numerador y $d(X)^4$ aumenta y que esto sucede en general.

En una normal estándar, este cociente vale 3, lo que explica que se reste 3 en la expresión anterior.

Si la distribución es mesocúrtica, entonces tenemos el mismo apuntamiento que la normal estándar y $g_2(X)$ se anula. Si el apuntamiento es mayor que el de la normal estándar, $g_2(X)$ toma un valor positivo y si el apuntamiento es menor que el de la normal estándar, $g_2(X)$ toma un valor negativo. Cuanto mayor sea el valor de $g_2(X)$, mayor será el apuntamiento.

Para variables de las que tenemos la distribución de frecuencias y no están agrupadas en clases se tiene la expresión

$$g_2(X) := \frac{\dfrac{\sum_{i=1}^{k}(x_i - \overline{x})^4 n_i}{n}}{d(X)^4} - 3 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 f_i}{d(X)^4} - 3.$$

Y para distribuciones agrupadas en clases se tiene que

$$g_2(X) := \frac{\dfrac{\sum_{i=1}^{k}(c_i - \overline{x})^4 n_i}{n}}{d(X)^4} - 3 = \frac{\sum_{i=1}^{k}(c_i - \overline{x})^4 f_i}{d(X)^4} - 3.$$

**Ejemplo 43.**

*Consideremos una variable cuya tabla de frecuencias es la de la tabla 3.16.*

| $x_i$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $f_i$ | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |

**Tabla 3.16.** Distribución del ejemplo 43.

*Esta distribución es simétrica respecto a 2 y es claramente platicúrtica, pues su diagrama de barras da una figura completamente plana. Entonces,*

$$\overline{x} = 2, \quad \sum_{i=1}^{k}(x_i - \overline{x})^4 f_i = \frac{34}{5}, \quad v(X) = 2, \quad d(X)^4 = 4.$$

*Entonces,*

$$g_2(X) = \frac{34}{20} - 3 = -\frac{26}{20} = -1{,}3,$$

*concluyendo que efectivamente es una distribución platicúrtica.*