

## 1 DEFINICIÓN DE ESTADÍSTICA

Ciencia que estudia el conjunto de métodos y procedimientos utilizados para:

- La sistematización, recogida, ordenación y presentación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico.
- Deducir las leyes que rigen esos fenómenos.
- Hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones.
- **Estadística descriptiva:** Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.
- **Estadística inferencial:** Apoyándose en el cálculo de probabilidades y a partir de datos muestrales; efectúa estimaciones, decisiones, predicciones u otras generalizaciones sobre un conjunto mayor de datos.

## 2 CONCEPTOS BÁSICOS

- **Individuo:** unidad experimental o estadística.
- **Carácter o variable:** característica de interés sobre cada individuo o elemento individual de una población o muestra.

Se representa por un símbolo (X,Y,A,B,... ) y puede tomar cualquier valor (modalidad) de un conjunto determinado, que llamaremos *dominio de la variable* o *rango*. Cuando la variable puede tomar un único valor se denomina constante.

- **Población:** conjunto de individuos o elementos de la misma naturaleza que presentan uno o varios caracteres comunes a todos ellos susceptibles de medida o clasificación (ej: alumnos de la UAX).

Una población puede ser finita, cuando está constituida por un número limitado de elementos, o bien infinita, cuando el número de elementos que la integran no es cuantificable.

- **Muestra:** subconjunto *representativo* de una población.
- **Dato:** valor de la variable asociada a un elemento de una población o muestra. Este valor puede ser un número, una palabra o un símbolo.
- **Modalidad:** cada uno de los valores que puede tomar una variable. Cada elemento puede tener una, y solo una, modalidad.
- **Clase:** Cada uno de los conjuntos en los que se agrupan las diferentes modalidades de una variable para evitar tener un número muy grande de observaciones diferentes. Cada modalidad ha de pertenecer a una y solo a una de las clases (sistema exhaustivo y excluyente).

Para cada clase es posible definir un valor único que la represente, a ese valor le denominaremos *marca de clase* y lo representamos habitualmente por la letra *m*.

### 3 TIPOS DE VARIABLES ESTADÍSTICAS

Podemos distinguir entre:

- **Variable cualitativa:** Sus valores (modalidades) no se pueden asociar naturalmente a un número (no se pueden hacer operaciones algebraicas con ellos). Podemos distinguir dos tipos:
  - **Nominales:** sus valores no se pueden ordenar. Por ejemplo, el grupo sanguíneo tiene por modalidades: A, B, AB, O.
  - **Ordinales (o cuasicuantitativos):** sus valores son de tipo nominal pero es posible establecer un orden entre ellas. Por ejemplo, si estudiamos el grado de recuperación de un paciente al aplicarle un tratamiento, podemos tener como modalidades: Nada, Poco, Moderado, Bueno, Muy Bueno. A veces los valores a tomar por estas variables se basan en escalas numéricas, puntuar el dolor en una escala de 1 a 10 por ejemplo. En estos casos debemos evitar realizar operaciones algebraicas con estas cantidades.
- **Variables cuantitativas o numéricas:** sus valores posibles son cantidades numéricas con las que podemos hacer operaciones aritméticas. Dentro de este tipo de variables podemos distinguir dos tipos:
  - **Discretas:** solo admiten algunos valores determinados entre su valor máximo y mínimo. Un ejemplo es la variable “número de hijos por mujer”; evidentemente solo puede tomar valores enteros: 0, 1, 2, 3, 4, 5, ...
  - **Continuas:** admiten cualquier valor posible entre dos dados. Por ejemplo, el peso de un niño al nacer.

Cualquier variable cuantitativa puede transformarse, de alguna manera, en cualitativa pero no así al revés. El tratamiento estadístico de datos cuantitativos es distinto del tratamiento de datos cualitativos. Si en alguna ocasión queremos tratar conjuntamente una variable cuantitativa y otra cualitativa habrá que transformar, irremediablemente, una en otra.

### 4 TABLAS ESTADÍSTICAS

En primer lugar es necesario definir una serie de magnitudes que nos ayudarán a describir nuestros datos.

- **Frecuencia absoluta:** número de individuos o elementos que pertenecen a una determinada clase de una variable. Se suele representar por la letra  $f$  acompañada de un subíndice que indica la clase. Así pues representaremos la frecuencia absoluta por  $f_i$ .
- **Frecuencia relativa:** cociente entre las frecuencias absolutas de una clase y el número total de datos. El número total de datos lo representaremos por la letra  $N$  y la frecuencia relativa de cada clase por  $h_i$ .
- **Frecuencia absoluta acumulada:** suma de las frecuencias absolutas de todas las clases inferiores a la clase que estamos calculando. Se representa por  $F_i$ . Si tenemos  $k$  clases distintas se cumplirá, lógicamente, que  $N_k = N$ . Sólo tienen sentido para variables ordinales y numéricas.

- **Frecuencia relativa acumulada:** tanto por uno de datos que pertenecen a la clase en estudio o a alguna de las inferiores. Se representa por  $H_i$ . Si tenemos  $k$  clases distintas se cumplirá que  $H_k = 1$ . Al igual que el caso anterior sólo tienen sentido para variables ordinales y numéricas.

Una tabla estadística no es más que una tabla organizada de tal manera que en la sucesivas filas irán apareciendo cada una de las clases de forma ordenada y en cada una de las sucesivas columnas iremos colocando datos como el intervalo de clase, la marca de clase y las frecuencias.

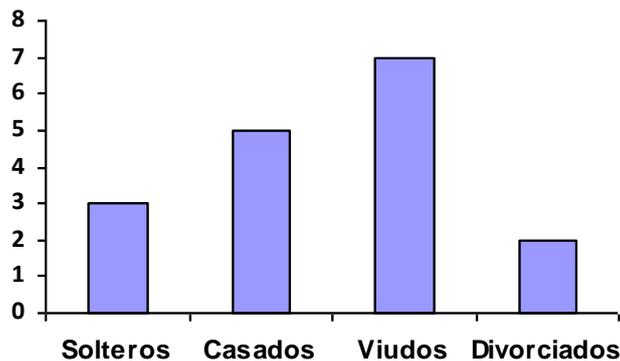
Modalidad	$f_i$	$h_i$	$F_i$	$H_i$
$C_1$	$f_1$	$h_1 = \frac{f_1}{N}$	$F_1 = f_1$	$H_1 = \frac{F_1}{N}$
...	...	...	...	...
$C_j$	$f_j$	$h_j = \frac{f_j}{N}$	$F_j = f_1 + \dots + f_j$	$H_j = \frac{F_j}{N}$
...	...	...	...	...
$C_k$	$f_k$	$h_k = \frac{f_k}{N}$	$F_k = N$	$H_k = 1$

Si se trata de una variable continua y tenemos los datos agrupados en intervalos, la tabla estadística quedaría como se muestra a continuación.

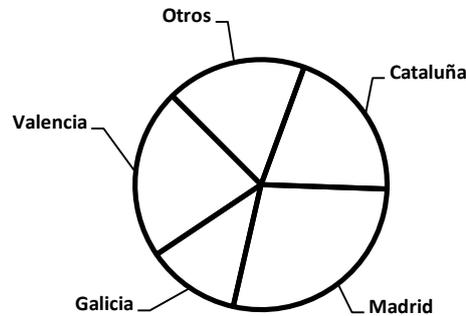
$L_{i-1} - L_i$	$m_i$	$f_i$	$h_i$	$F_i$	$H_i$
$L_0 - L_1$	$m_1 = \frac{L_1 + L_0}{2}$	$f_1$	$h_1 = \frac{f_1}{N}$	$F_1 = f_1$	$H_1 = \frac{F_1}{N}$
...	...	...	...	...	...
$L_{j-1} - L_j$	$m_j = \frac{L_j + L_{j-1}}{2}$	$f_j$	$h_j = \frac{f_j}{N}$	$F_j = f_1 + \dots + f_j$	$H_j = \frac{F_j}{N}$
...	...	...	...	...	...
$L_{k-1} - L_k$	$m_k = \frac{L_k + L_{k-1}}{2}$	$f_k$	$h_k = \frac{f_k}{N}$	$F_k = N$	$H_k = 1$

## 5 GRÁFICOS

- Diagramas de barras: representaremos en el eje de ordenadas los valores posibles de la variable y en el eje de abscisas las frecuencias absolutas o bien, las frecuencias relativas. Si, mediante el gráfico, se intenta comparar varias poblaciones entre sí es conveniente utilizar las frecuencias relativas, ya que en otro caso podrían resultar engañosas.



- Diagramas de sectores: se divide un círculo en tantas porciones como clases existan, de modo que a cada clase le corresponde un arco de círculo proporcional a su frecuencia absoluta o relativa. El arco de cada porción se calcula usando la regla de tres:

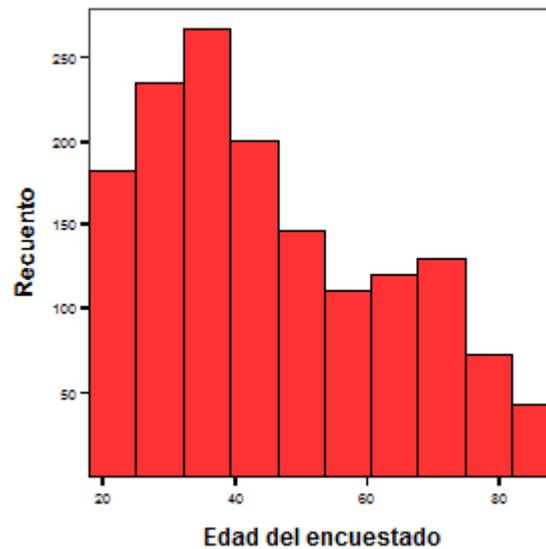


$$N \rightarrow 360^\circ$$

$$f_i \rightarrow x_i = \frac{360 \cdot f_i}{N}$$

- Histograma: es un diagrama de barras en el que el área de la barra que representa a cada clase coincide con su frecuencia absoluta. Se utiliza para representar variables cuantitativas. Para calcular la altura de cada barra, a la que nos referiremos por *densidad*, basta tener en cuenta que:

$$f_i = \text{Amplitud del intervalo} \cdot \text{densidad}$$



- Polígonos de frecuencias: este tipo de gráficos está indicado para datos cuantitativos, especialmente cuando se quiere observar la evolución de las frecuencias conforme aumenta el valor de los datos. A cada clase se le asigna un punto en el plano, con una abscisa que es el valor del dato (o la marca de clase si el dato es continuo) y una ordenada que corresponde a su frecuencia. Los puntos así obtenidos se conectan entre sí mediante segmentos de recta obteniendo así la poligonal buscada. En el caso de haber construido un histograma para construir el polígono de frecuencia no hay más que unir los puntos medios de las bases superiores de los rectángulos del histograma mediante segmentos rectos.

## 6 PARÁMETROS Y ESTADÍSTICOS

- **Parámetro:** Es una cantidad numérica calculada sobre una población.
- **Estadístico:** Es una cantidad numérica calculada sobre una muestra. Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

Normalmente nos interesa conocer un parámetro, pero esto es raramente posible debido a la dificultad que conlleva estudiar a “*TODA*” la población (pensemos en poblaciones infinitas o con un número muy grande de individuos), calculamos un estimador sobre una muestra y “*confiamos*” en que sean próximos. Más adelante veremos cómo elegir muestras para que el error sea “*confiablemente*” pequeño.

Podemos agrupar los estadísticos en:

- Estadísticos de posición: dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos (cuantiles, percentiles, cuartiles, deciles,...).
- Estadísticos de centralización: indican valores con respecto a los que los datos parecen agruparse (Media, mediana y moda). También nos referimos a ellos como *medidas de tendencia central o de centralización*.
- Estadísticos de dispersión: indican la mayor o menor concentración de los datos con respecto a las medidas de centralización (Desviación típica, coeficiente de variación, rango, varianza).
- Estadísticos de forma: asimetría y apuntamiento o curtosis

## 7 MEDIDAS DE CENTRALIZACIÓN

Las medidas de centralización se refieren exclusivamente a variables cuantitativas, ya que no es posible realizar cálculos con las variables cualitativas. El objeto de las medidas de centralización es buscar un buen representante para todos los datos, es decir resumir las observaciones realizadas en un solo valor. Las medidas de centralización son valores situados siempre entre los extremos de los datos (supuestos estos ordenados). No tendría ningún sentido que la media fuera menor o mayor que todos los datos.

### 7.1 MEDIA ARITMÉTICA

Es la medida de tendencia central que se utiliza más a menudo. Suele representarse por  $\bar{X}$  y se calcula sumando todas las observaciones de un conjunto de datos y dividiendo entre el total de las mediciones, esto es:

Con los datos si agrupar

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

Con los datos y sus frecuencias absolutas

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot f_i}{N}$$

Si los datos vienen agrupados en intervalos

$$\bar{x} = \frac{\sum_{i=1}^k m_i \cdot f_i}{N} = \sum_{i=1}^k m_i \cdot h_i \quad m_i: \text{marca de clase}$$

La media presenta algunos inconvenientes de los cuales, los más relevantes son los que se comentan a continuación:

- Es muy sensible a los valores extremos de la variable, ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección. En consecuencia, no es recomendable usar la media como medida central en las distribuciones muy asimétricas.
- Si consideramos una variable discreta, por ejemplo, el número de hijos en las familias españolas el valor de la media puede no pertenecer al conjunto de valores de la variable. Por ejemplo  $\bar{X} = 1,2$  hijos.

## 7.2 MEDIAS GENERALIZADAS

- Media geométrica: se utiliza con variables que evolucionan exponencialmente respecto al tiempo: número de microorganismos de un cultivo.

$$\bar{X}_g = \sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N} \qquad \bar{X}_g = \sqrt[N]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_k^{f_k}}$$

- Media armónica: se utiliza para calcular velocidades medias de sedimentación.

$$\bar{X}_a = \frac{N}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

- Media cuadrática: se utiliza cuando la variable toma valores positivos y negativos (por ejemplo, en los errores de medida) e interesa obtener un promedio que no recoja los efectos del signo.

$$\bar{X}_c = \sqrt{\frac{x_1^2 + \dots + x_n^2}{N}}$$

## 7.3 MEDIANA

Si ordenamos de menor a mayor las observaciones de una variable discreta, llamaremos mediana, y la denotaremos por Me, al primer valor de la variable que deja por debajo de sí al 50% de las observaciones. La mediana se define como el percentil 50 de un conjunto de mediciones.

Si el número de datos, N, que tiene un conjunto es impar, la mediana es el valor de la observación que está en medio de la lista ordenada, es decir, el dato que ocupa la posición  $(n + 1)/2$ .

Si N es par, la mediana, normalmente, se toma como el promedio de los dos valores centrales, es decir, el punto medio de los datos que ocupan las posiciones  $n/2$  y  $\frac{n}{2} + 1$ .

En el caso de variables continuas, las clases vienen dadas por intervalos, y aquí la fórmula de la mediana se complica un poco más. Sea  $(l_{i-1}, l_i]$  el intervalo donde hemos encontrado que por debajo están el 50% de las observaciones y que llamaremos intervalo mediano. Para encontrarlo basta dividir N entre 2 y buscar el primer intervalo de la tabla cuya frecuencia absoluta acumulada iguale o supere el valor resultante, ese será el intervalo mediano. Entonces se obtiene la mediana mediante interpolación lineal según la fórmula:

$$Me = l_{i-1} + \frac{\frac{N}{2} - F_{i-1}}{f_i} \cdot (l_i - l_{i-1})$$

Esto equivale a decir que la mediana divide al histograma en dos partes de áreas iguales.

Entre las ventajas de la mediana como medida descriptiva se cuentan:

- Tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.
- Es de cálculo rápido y de interpretación sencilla.
- A diferencia de la media, la mediana de una variable discreta es siempre un valor del dominio de la variable que estudiamos con la única excepción del caso de un número par de datos referentes a una variable discreta.

#### 7.4 MODA

Es el máximo de la función de distribución. Es decir, es el valor que más se repite, el más frecuente. Puede no ser única, esto es, podemos tener distribuciones con unimodales (una moda), bimodales (dos modas), trimodales (tres modas), etcétera. La moda o modas, caso de no ser única, la denotaremos por  $M_o$ .

Cuando se trata de datos agrupados en intervalos, llamaremos intervalo modal al que contenga la mayor cantidad de datos, es decir al de mayor frecuencia absoluta, y tomaremos su marca de clase como la moda de la distribución.

#### 7.5 RELACIÓN ENTRE MEDIA, MEDIANA Y MODA

La mejor medida de tendencia central de un conjunto de datos a menudo depende de la forma en que se distribuyan los valores. Si son simétricos y unimodales (su histograma sólo tiene un pico), la media, la mediana y la moda deberían tener valores muy parecidos. Si la distribución es simétrica pero bimodal (dos picos) la media y la mediana tendrían valores parecidos, pero no tendría por qué ser así con la moda, que podría registrar dos valores.

Cuando los datos no son simétricos, la mediana es a menudo la mejor medida de tendencia central. Debido a que la media es sensible a las observaciones extremas, se carga en dirección de los datos alejados del centro y, como consecuencia, podría obtener un valor excesivamente grande o pequeño. Cuando los datos están sesgados a la derecha, la media se sitúa a la derecha de la mediana, y si están sesgados a la izquierda, entonces se localiza a la izquierda de la mediana. Muestras con la misma media, mediana y moda pueden ser en realidad muy diferentes.

## 8 ESTADÍSTICOS DE POSICIÓN

Son valores de la variable que se caracterizan por superar a un cierto porcentaje de los valores de la muestra.

- Percentiles: dividen la muestra en cien partes iguales. El percentil de orden  $k$  se define como la observación de deja por debajo de si el  $k\%$  de las observaciones y se representa por  $P_k$ .

En el caso de que la variable sea continua, una vez localizado el intervalo  $[l_{i-1}, l_i)$  que deja por debajo de si al  $k\%$  de las observaciones, el  $P_k$  se obtiene:

$$P_k = l_{i-1} + \frac{N \cdot \frac{k}{100} - F_{i-1}}{f_i} \cdot (l_i - l_{i-1})$$

- Cuartiles: dividen a los datos en cuatro partes iguales, luego hay tres cuartiles y son un caso particular de los percentiles,  $P_{25}$ ,  $P_{50}$  y  $P_{75}$ . El segundo cuartil,  $P_{50}$ , coincide con la mediana.
- Deciles: dividen a los datos en diez partes iguales

La siguiente fórmula sirve para encontrar el intervalo percentílico  $K(x)$  en que se halla un valor  $x$ :

$$K(x) = E^+ \left( \frac{\text{lugar que ocupa } x \text{ en la lista ordenada}}{\text{número de valores en la lista}} \cdot 100 \right)$$

$E^+$  significa entero positivo siguiente al resultado de la operación. Sustituyendo 100 por 4 y por 10 se calculan los intervalos cuartílicos y decílico, respectivamente.

## 9 MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Nos indican si las observaciones de una muestra están próximas entre sí o si por el contrario están muy dispersas.

### 9.1 RANGO

Se define como la diferencia entre la mayor observación y la menor observación. Se denota por  $R$  o  $R_g$ . Su interés es muy relativo, puesto que utiliza para su cálculo únicamente dos valores de la serie: toma en cuenta los valores extremos de un conjunto de datos en lugar de considerar la mayoría de las observaciones.

### 9.2 VARIANZA

Es una forma de cuantificar la cantidad de variabilidad, o dispersión, alrededor de la media de las observaciones. La varianza se representa por  $s^2$  y se calcula con las fórmulas según los datos se presenten agrupados en intervalos, de forma independiente o conocida su frecuencia:

$$s^2 = \frac{1}{N} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{N} \cdot \sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{N} \cdot \sum_{i=1}^n f_i \cdot (m_i - \bar{x})^2$$

Es muy habitual (por razones que explicaremos más adelante en el curso) calcular la varianza con denominador  $N-1$ , es decir:

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^n f_i \cdot (m_i - \bar{x})^2$$

### 9.3 DESVIACIÓN TÍPICA O ESTÁNDAR

La varianza no tiene la misma magnitud que los datos (ej. si las observaciones se miden en metros, la varianza lo hace en metros cuadrados) lo que hace más difícil su interpretación. En la práctica se prefiere trabajar con su raíz cuadrada que se denomina desviación típica y se representa por  $s$ . Es decir, la desviación típica tendrá las mismas unidades que los datos, asegurando de esta manera una más fácil interpretación.

$$s = \sqrt{s^2}$$

La desviación típica tiene la propiedad de que en el intervalo  $(\bar{x} - 2 \cdot s, \bar{x} + 2 \cdot s)$  se encuentra, al menos, el 75% de las observaciones.

No es recomendable su uso cuando tampoco lo sea el de la media como medida de tendencia central.

## 10 COEFICIENTE DE VARIACIÓN

Cuando queremos comparar la variabilidad de características de dos muestras o poblaciones no podemos hacerlo directamente comparando los valores de las desviaciones típicas ya que podría llevarnos a errores. Necesitamos eliminar la dimensionalidad de las magnitudes a comparar. Para ello calcularemos el coeficiente de variación:

$$CV = \frac{s}{\bar{x}}$$

Solo debe calcularse para variables con todos sus valores positivos ya que no tiene sentido hablar de índices de variabilidad negativos.

## 11 ASIMETRÍA Y APUNTAMIENTO

### 11.1 ESTADÍSTICOS DE ASIMETRÍA

Para saber si una distribución de frecuencias es simétrica hemos de precisar respecto de qué. En nuestro caso mediremos la simetría de la distribución respecto de la media y para ello calcularemos el momento central de tercer orden.

Se denomina momento central de orden  $p$  a la cantidad:

$$m_p = \frac{1}{N} \cdot \sum_{i=1}^n (x_i - \bar{x})^p$$

Si la distribución fuese perfectamente simétrica se cumpliría que  $m_3 = 0$ . Si el valor de  $m_3$  es mayor que cero diremos que hay asimetría positiva y si es menor que cero diremos que hay asimetría negativa.

### 11.2 ÍNDICE DE ASIMETRÍA BASADO EN LOS CUARTILES

El índice de asimetría basado en los cuartiles es muy fácil y rápido de calcular y además es invariante ante cambios de origen o de escala. Se calcula con la fórmula siguiente:

$$\mathcal{A}_s = \frac{(Q_1 - Q_2) - (Q_2 - Q_3)}{(Q_3 - Q_1)}$$

Si  $\mathcal{A}_s$  vale cero, la distribución es simétrica. Si el valor de  $\mathcal{A}_s$  es positivo la distribución será asimétrica positiva y si es negativo la distribución será asimétrica negativa.

### 11.3 ESTADÍSTICOS DE APUNTAMIENTO

Se define el coeficiente de aplastamiento de Fisher (curtosis) como:

$$\gamma_2 = \frac{m_4}{s^4} - 3$$

Donde  $m_4$  es el momento empírico de cuarto orden. El patrón de referencia es la distribución normal para la que  $\gamma_2 = 0$ . Así se clasifican las distribuciones en:

- Leptocúrtica: Si  $\gamma_2 > 0$  y la distribución es más apuntada que la normal.
- Mesocúrtica: Si  $\gamma_2 = 0$  es igual de apuntada que la normal.
- Platicúrtica: Si  $\gamma_2 < 0$  y la distribución es menos apuntada que la normal.

## 12 TIPIFICACIÓN

Se conoce por tipificación al proceso de restar la media y dividir por su desviación típica a una variable  $X$ . De este modo se obtiene una nueva variable  $Z$  de media  $\bar{z} = 0$  y desviación típica  $s = 1$ , que denominamos variable tipificada. El valor de la nueva variable,  $z_i$  expresa la separación del valor  $x_i$  de la media usando como unidad de longitud la desviación típica.

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

La variable tipificada carece de unidades y permite hacer comparables dos medidas que en un principio no lo son. Así por ejemplo nos podemos preguntar si un elefante es más grueso que una hormiga determinada, cada uno en relación a su población.

Conviene no confundir el coeficiente de variación que sirve para comparar la variabilidad de dos conjuntos de datos diferentes y la tipificación que nos permitirá comparar dos individuos concretos de dos conjuntos diferentes.

**EJERCICIO 1.** Se han recogido datos sobre la cantidad diaria de contagios de una enfermedad en una explotación ganadera resultando la siguiente tabla:

Contagios	0	1	2	3	4	5	6	7	8
Días	2	10	24	5	4	15	15	14	2

- Construye una tabla estadística con todas las frecuencias.
- Representa los datos en un diagrama de barras datos-frecuencias absolutas.
- Representa los datos en un diagrama de barras datos-frecuencias acumuladas.
- Representa los datos en un polígono de frecuencias datos-frecuencias relativas.

**EJERCICIO 2.** Se han clasificado cien familias según el número de hijos resultando los datos que se muestran en la siguiente tabla:

Nº hijos	0	1	2	3	4	5	6	7	8
Nº familias	11	13	20	25	14	10	4	2	1

- Representa los datos en un gráfico de datos-frecuencias absolutas.
- Representa los datos en un gráfico de datos-frecuencias relativas.
- Representa los datos en un gráfico de datos-frecuencias absolutas acumuladas.
- Representa los datos en un gráfico de datos-frecuencias relativas acumuladas.
- Calcula la media, la mediana y la moda
- Calcula la varianza y la desviación típica.

**EJERCICIO 3.** Se presentan a continuación los datos de un estudio sobre los niveles de colesterol en la sangre de varones estadounidenses en los años 1976-1980 en distintas franjas de edad:

Nivel de colesterol (mg / 100 ml)	Edades de 25-34 Número de varones	Edades 55-64 Número de varones
88-119	13	5
120-159	150	48
160-199	442	265
200-239	299	458
240-279	115	281
280-319	34	128
320-359	9	35
360-399	5	7

- Crea una tabla con las frecuencias relativa y relativa acumulada.
- Representa el histograma de los datos.
- Calcula la media agrupada, la varianza agrupada y la desviación estándar agrupada de los niveles de colesterol en sangre para cada franja de edades.
- ¿Cuál de las dos distribuciones presenta mayor dispersión?
- ¿Cuál de las dos distribuciones es más simétrica?

**EJERCICIO 4.** Las puntuaciones obtenidas por un grupo de alumnos de veterinaria en un examen fueron las que siguen:

Puntuaciones	0	1	2	3	4	5	6	7	8	9
Nº alumnos	2	3	3	4	7	7	6	5	2	1

- Representación gráfica de las frecuencias
- Calcula la media, moda, mediana
- Calcula la varianza y desviación típica.

**EJERCICIO 5.** Los pesos de los niños observados en una consulta, al cumplir el primer mes de vida, varían de acuerdo con la siguiente tabla:

Peso (gr.)	Nº de niños
3000-3700	2
3700-4000	5
4000-4200	18
4200-4500	91
4500-4700	15
4700-5000	6
5000-5400	3

- Completa una tabla de frecuencias de los datos
- Representa el histograma de los datos
- Halla el peso medio de los niños revisados y la desviación típica.

**EJERCICIO 6.** Se preguntó a 250 jóvenes de edades comprendidas entre los 14 y 25 años por el número de mensajes de texto que enviaban al día y los resultados se representaron en la siguiente tabla:

Nº de mensajes	Nº de jóvenes
0	12
1	10
2	17
3	45
4	18
5	17
6	22
7	13
8	22
9	20
10	15
11	19
12	20

- Calcula la media, la moda, la mediana, la varianza y la desviación típica de esta variable estadística.
- Calcula los tres cuartiles.
- Si cada mensaje de texto cuesta 15 céntimos de euro, ¿cuánto es el gasto medio?

**EJERCICIO 7.** En un estudio sobre el crecimiento de los varones, se obtuvieron estas observaciones sobre el perímetro craneal en centímetros de un niño al nacer.

Perímetro Craneal (cm)	Nº varones
33,10 - 33,78	5
33,78 - 34,46	12
34,46 - 35,14	11
35,14 - 35,82	5
35,82 - 36,50	3

- Dibuja el histograma de los datos.
- Crea una tabla de frecuencias completa.
- Calcula la media, la moda, la mediana, la varianza y la desviación típica de la muestra.
- ¿Cómo clasificarías esta distribución por su apuntamiento?

**EJERCICIO 8.** En un estudio sobre el cangrejo Xanthidae se recogieron el número de huevos puestos por individuo.

Nº de huevos	Individuos
727 – 2153	6
2154 – 3580	4
3581 – 5007	9
5008 – 6434	7
6435 – 7861	11
7862 – 9288	5
9289 – 10715	2
10716 – 12142	1

- Crea una tabla de frecuencias completa.
- Representa el histograma de la distribución.
- Calcula media, mediana, moda, varianza y desviación típica.
- Otra muestra recogida un tiempo después tiene la una media de 600 y una desviación típica de 49. ¿Es su dispersión mayor que la de la tabla o no?

**EJERCICIO 9.** En una explotación ganadera se ha pesado a un grupo de cerdos y los datos obtenidos se han agrupado en la tabla siguiente:

Peso (kg)	Nº animales
504,0 - 541,7	16
541,7 - 579,4	10
579,4 - 617,1	10
617,1 - 654,8	7
654,8 - 692,5	7
692,5 - 730,2	2
730,2 - 767,9	3
767,9 - 805,6	1

- Crea una tabla de frecuencias completa.
- Representa el histograma de la distribución.
- Calcula media, mediana, moda, varianza y desviación típica.
- Estudia la simetría de la distribución.

**EJERCICIO 10.** Ocho individuos que sufrían intoxicación por vitamina D y una de las hipótesis de la causa era una dieta con ingestión excesiva de productos lácteos. Los niveles de calcio y albúmina que presentaban en el momento de ingresar en el hospital eran:

Calcio (amol/l)	Albúmina (g/l)	Calcio (amol/l)	Albúmina (g/l)
2,92	43	2,67	42
3,84	42	3,17	38
2,37	42	3,74	34
2,99	40	3,44	42

- Determinar la media, la mediana, la desviación estándar y el rango de los niveles de calcio y albúmina registrados.

**EJERCICIO 11.** A continuación se presentan un par de distribuciones de frecuencias que contienen los niveles de cotinina en la sangre de un grupo de fumadores y un grupo de no fumadores.

Nivel de cotinina (ngr/ml)	Fumadores	No fumadores
0-13	78	3300
14-49	133	72
50-99	142	23
100-149	206	15
150-199	197	7
200-249	220	8
250-299	151	9
300-349	412	11

- Calcula, para ambas distribuciones, la media, la mediana, la moda y la desviación típica.
- ¿Cuál de las distribuciones presenta menos dispersión?
- Estudia la simetría de ambas distribuciones.
- Clasifícalas por su apuntamiento.
- ¿Quién presenta un nivel más alto de cotinina en relación a su grupo, un fumador que tiene 102 ngr/ml o un no fumador al que se han medido 61 ngr/ml.

**EJERCICIO 12.** Se ha venido realizando un estudio sobre la capacidad curativa de un cierto fármaco y se ha tomado, entre otros, datos sobre la duración del tratamiento hasta la total desaparición de los síntomas. Los datos obtenidos son los que figuran en la siguiente tabla:

Días de tratamiento	Individuos
0 - 2	214
2 - 4	326
4 - 7	857
7 - 10	1532
10 - 15	1651
15 - 20	1233
20 - 25	703
25 - 30	425
30 - 40	87
40 - 50	8

- Completar la tabla con todas las frecuencias y representar el histograma de los datos.

- Calcular todas las medidas de tendencia central y dispersión.
- Calcular los cuartiles y los deciles.

### SOLUCIONES A LOS EJERCICIOS

Se dan a continuación algunos resultados, a fin de que podáis comprobar si habéis resuelto bien el ejercicio.

<b>EJERCICIO 2:</b>	Media: 2,8	Des. Típ.: 1,772	
<b>EJERCICIO 3:</b>	Media 1: 198,89	Des. Típ. 1: 43,81	Mediana 1: 193,03
	Media 2: 229,1	Des. Típ. 2: 46,36	Mediana 2: 225,31
<b>EJERCICIO 4:</b>	Des. Típ.: 2,225	Moda: 4 y 5	
<b>EJERCICIO 5:</b>	Des. Típ.: 367,823		
<b>EJERCICIO 6:</b>	Mediana: 6	Des. Típ.: 3,557	Q3: 9
<b>EJERCICIO 7:</b>	Mediana: 34,522	Des. Típ.: 0,766	Platicúrtica
<b>EJERCICIO 8:</b>	Mediana: 5721	Des. Típ.: 2549,542	
<b>EJERCICIO 9:</b>	Mediana: 574,01	Des. Típ.: 71,234	
<b>EJERCICIO 10:</b>	Des. Típ. 1: 0,477		
	Des. Típ. 2: 2,826		
<b>EJERCICIO 11:</b>	Des. Típ. 1: 106,79		
	Des. Típ. 2: 27,683		
<b>EJERCICIO 12:</b>	Mediana: 11,784	Des. Típ.: 7,262	Q3: 17,826