Estadísticos de Posición

Sesión 4

Los estadísticos de posición van a ser valores de la variable caracterizados por superar a cierto porcentaje de observaciones en la población (o muestra). Tenemos fundamentalmente a los *percentiles* como medidas de posición, y asociados a ellos veremos también los *cuartiles, deciles y cuartiles.*

Percentiles

Para una variable discreta, se define el **percentil de orden** k, como la observación, P_k, que deja por debajo de si el $k\,\%$ de la población. Véase la figura 2.4. Esta definición nos recuerda a la mediana, pues como consecuencia de la definición es evidente que

$$M_{ed} = P_{50}$$

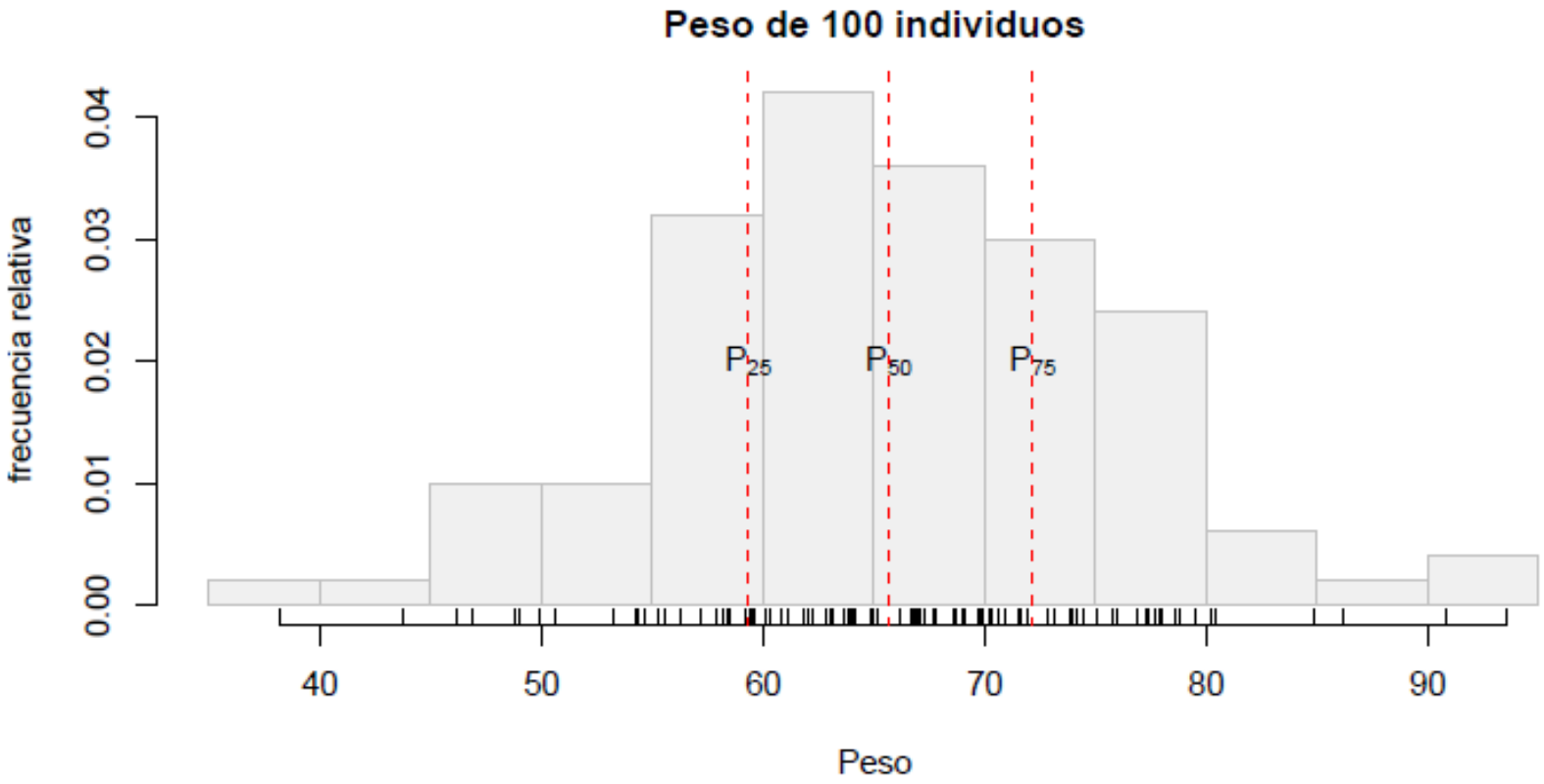


Figura 2.4: Percentiles 25, 50 y 75 de una variable. Los que se muestran dividen a la muestra en cuatro intervalos con similar número de individuos y reciben también el nombre de cuartiles.

En el caso de una variable continua, el intervalo donde se encuentra $P_k \in (l_{i-1}, l_i]$, se calcula buscando el que deja debajo de si al $k\,\%$ de las observaciones. Dentro de él, P_k se obtiene según la relación:

$$\boxed{P_k = l_{i-1} + \frac{n\,\dfrac{k}{100} - N_{i-1}}{n_i} \cdot a_i} \tag{2.3}$$

Cuartiles

Los **cuartiles**, Q_l, son un caso particular de los percentiles. Hay 3, y se definen como:

$$Q_1 = P_{25} \tag{2.4}$$
$$Q_2 = P_{50} = M_{ed} \tag{2.5}$$
$$Q_3 = P_{75} \tag{2.6}$$

Deciles

Se definen los **deciles** como los valores de la variable que dividen a las observaciones en 10 grupos de igual tamaño. Más precisamente, definimos D_1,D_2, ..., D_9 como:

$$D_i = P_{10\,i} \qquad i = 1, \ldots, 9$$

Ejemplo de cálculo de cuartiles con una variable discreta

Dada la siguiente distribución en el número de hijos de cien familias, calcular sus cuartiles.

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
n=100		

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
n=100		

Solución:

1. Primer cuartil:

$$\frac{n}{4} = 25;\ \text{Primera } N_i > n/4 = 39;\ \text{luego } Q_1 = 2.$$

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
n=100		

2. Segundo cuartil:

$$\frac{2\,n}{4} = 50;\ \text{Primera } N_i > 2\,n/4 = 65;\ \text{luego } Q_2 = 3.$$

x_i	n_i	N_i
0	14	14
1	10	24
2	15	39
3	26	65
4	20	85
5	15	100
n=100		

3. Tercer cuartil:

$$\frac{3\,n}{4} = 75;\ \text{Primera } N_i > 3\,n/4 = 85;\ \text{luego } Q_3 = 4.$$

Ejemplo

Calcular los cuartiles en la siguiente distribución de una variable continua:

$l_{i-1} - l_i$	n_i	N_i
0 – 1	10	10
1 – 2	12	22
2 – 3	12	34
3 – 4	10	44
4 – 5	7	51
	$n = 51$	

$l_{i-1} - l_i$	n_i	N_i
$0 - 1$	10	10
$1 - 2$	12	22
$2 - 3$	12	34
$3 - 4$	10	44
$4 - 5$	7	51
	$n = 51$	

Solución:

1. Primer cuartil

$$\frac{N}{4} = 12,75;\ \text{Primera } N_i > n/4 = 22;\ \text{La línea } i \text{ es la del intervalo } [1;2)$$

$$Q_1 = l_{i-1} + \frac{\frac{n}{4} - N_{i-1}}{n_i} a_i = 1 + \frac{12,75 - 10}{12} \times 1 = 1,23$$

$l_{i-1} - l_i$	n_i	N_i
$0 - 1$	10	10
$1 - 2$	12	22
$2 - 3$	12	34
$3 - 4$	10	44
$4 - 5$	7	51
	$n = 51$	

2. Segundo cuartil:

$$\frac{2\,n}{4} = 25,5;\ \text{Primera } N_i > 2\,n/4 = 34;\ \text{La línea } i \text{ es la del intervalo } [2;3)$$

$$Q_2 = l_{i-1} + \frac{\frac{2\,n}{4} - N_{i-1}}{n_i}\,a_i = 2 + \frac{25,5 - 22}{12} \times 1 = 2,29$$

$l_{i-1} - l_i$	n_i	N_i
$0 - 1$	10	10
$1 - 2$	12	22
$2 - 3$	12	34
$3 - 4$	10	44
$4 - 5$	7	51
	$n = 51$	

3. Tercer cuartil

$\frac{3\,n}{4} = 38,25$; Primera $N_i > 3\,n/4 = 44$; La línea i es la del intervalo $[3; 4)$

$$Q_3 = l_{i-1} + \frac{\frac{3\,n}{4} - N_{i-1}}{n_i}\,a_i = 3 + \frac{38,25 - 34}{10} \times 1 = 3,445$$

Ejemplo de cálculo de cuartiles con una variable continua

Han sido ordenados los pesos de 21 personas en la siguiente tabla:

Intervalos $l_{i-1} - l_i$	f.a. n_i
38 — 45	3
45 — 52	2
52 — 59	7
59 — 66	3
66 — 73	6
	21

Intervalos $l_{i-1} - l_i$	f.a. n_i
$38 - 45$	3
$45 - 52$	2
$52 - 59$	7
$59 - 66$	3
$66 - 73$	6
	21

Encontrar aquellos valores que dividen a los datos en 4 partes con el mismo número de observaciones.

Solución: Las cantidades que buscamos son los tres cuartiles: Q_1, Q_2 y Q_3. Para calcularlos, le añadimos a la tabla las columnas con las frecuencias acumuladas, para localizar qué intervalos son los que contienen a los cuartiles buscados:

l_{i-1} — l_i	n_i	N_i	
38 — 45	3	3	
45 — 52	2	5	
52 — 59	7	12	$\ni Q_1, Q_2$
59 — 66	3	15	
66 — 73	6	21	$\ni Q_3$
	21		

Q_1 y Q_2 se encuentran en el intervalo 52—59, ya que $N_3 = 12$ es la primera f.a.a. que supera a $21 \cdot 1/4$ y $21 \cdot 2/4$. Q_3 está en 66—73, pues $N_5 = 21$ es el primer N_i mayor que $21 \cdot 3/4$.

Intervalos $l_{i-1} - l_i$	f.a. n_i
$38 - 45$	3
$45 - 52$	2
$52 - 59$	7
$59 - 66$	3
$66 - 73$	6
	21

Así se tiene que:

$$\frac{1}{4} \cdot 21 = 5,25 \Rightarrow \ i = 3 \Rightarrow Q_1 = l_{i-1} \frac{\frac{1}{4} n - N_{i-1}}{n_i} \cdot a_i$$
$$= 52 + \frac{5,25 - 5}{7} \cdot 7 = 52,25$$

$$\frac{2}{4} \cdot 21 = 10,5 \Rightarrow \ i = 3 \Rightarrow Q_2 = l_{i-1} + \frac{\frac{2}{4} n - N_{i-1}}{n_i} \cdot a_i$$
$$= 52 + \frac{10,5 - 5}{7} \cdot 7 = 57,5$$

Intervalos $l_{i-1} - l_i$	f.a. n_i
38 — 45	3
45 — 52	2
52 — 59	7
59 — 66	3
66 — 73	6
	21

$$\frac{3}{4} \cdot 21 = 15,75 \Rightarrow \ i = 5 \Rightarrow Q_3 \ = \ l_{i-1} + \frac{\frac{3}{4} n - N_{i-1}}{n_i} \cdot a_i$$

$$= \ 66 + \frac{15,75 - 15}{6} \cdot 7 = 66,875$$

Obsérvese que $Q_2 = M_{ed}$. Esto es lógico, ya que la mediana divide a la distribución en dos partes con el mismo número de observaciones, y Q_2, hace lo mismo, pues es deja a dos cuartos de los datos por arriba y otros dos cuartos por abajo.

Ejemplo

La distribución de una variable tiene por polígono acumulativo de frecuencias el de la figura 2.5. Si el número total de observaciones es 50:

1. Elaborar una tabla estadística con los siguientes elementos: intervalos, marcas de clase, frecuencia absoluta, frecuencia absoluta acumulada, frecuencias relativa y frecuencias relativa acumulada.

2. Cuántas observaciones tuvieron un valor inferior a 10, cuántas inferior a 8 y cuántas fueron superior a 11.

3. Determine los cuartiles.

Solución:

1. En la siguiente tabla se proporciona la información pedida y algunos cálculos auxiliares que nos permitirán responder a otras cuestiones.

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
0 – 5	10	10	0,2	0,3	2,5	5	2
5 – 7	25	35	0,5	0,7	6	2	12,5
7 – 12	5	40	0,1	0,8	9,5	5	1
12 – 15	10	50	0,2	1	13,5	7	3,33

2. Calculemos el número de observaciones pedido:

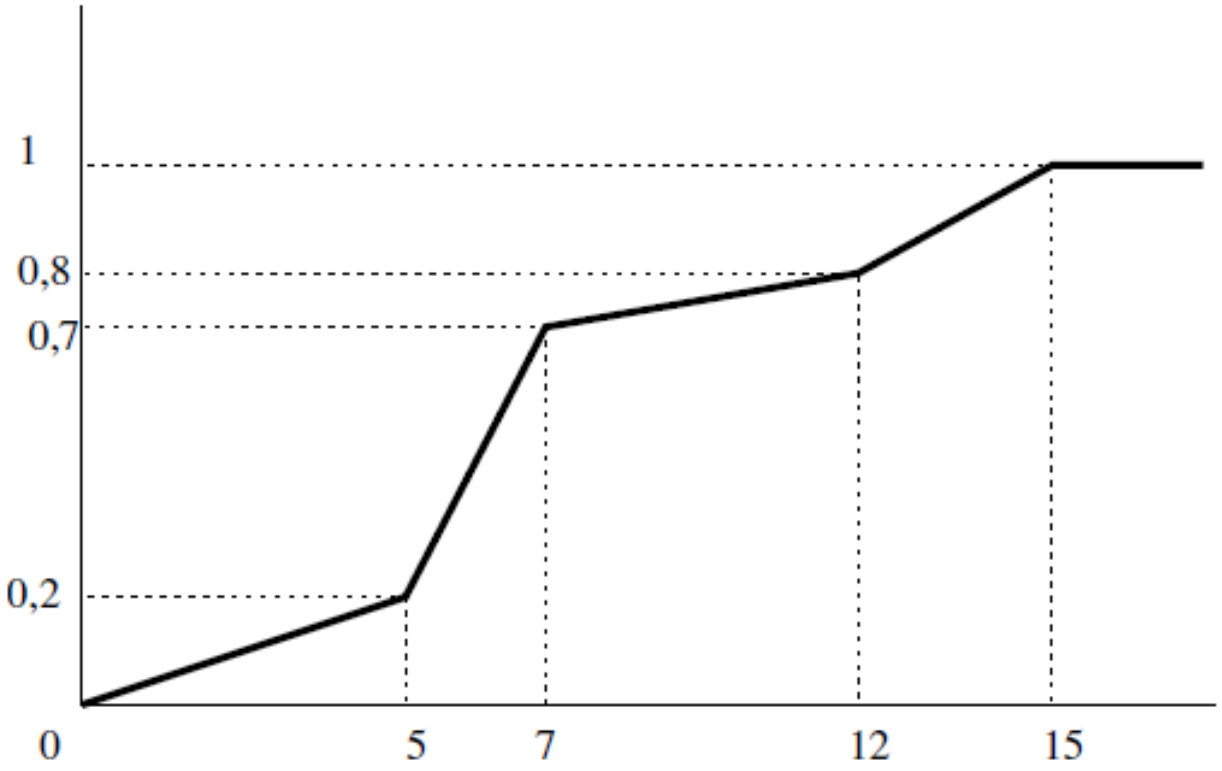


Figura 2.5: Diagrama acumulado de frecuencias relativas.

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
$0-5$	10	10	0,2	0,3	2,5	5	2
$5-7$	25	35	0,5	0,7	6	2	12,5
$7-12$	5	40	0,1	0,8	9,5	5	1
$12-15$	10	50	0,2	1	13,5	7	3,33

$$\begin{array}{ll} 7 \text{ a } 12 & 5 \\ 7 \text{ a } 10 & x \end{array} \Leftrightarrow \begin{array}{ll} 5 & 5 \\ 3 & x \end{array} \Rightarrow x = \frac{3 \times 5}{5} = 3$$

$10 + 25+3 = 38$ observaciones tomaron un valor inferior a 10

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
$0-5$	10	10	0,2	0,3	2,5	5	2
$5-7$	25	35	0,5	0,7	6	2	12,5
$7-12$	5	40	0,1	0,8	9,5	5	1
$12-15$	10	50	0,2	1	13,5	7	3,33

$$\begin{array}{ll} 7 \text{ a } 12 & 5 \\ 7 \text{ a } 8 & x \end{array} \Leftrightarrow \begin{array}{ll} 5 & 5 \\ 1 & x \end{array} \Rightarrow x = \frac{1 \times 5}{5} = 1$$

$10 + 25+1 = 36$ observaciones tomaron un valor inferior a 8

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
$0-5$	10	10	0,2	0,3	2,5	5	2
$5-7$	25	35	0,5	0,7	6	2	12,5
$7-12$	5	40	0,1	0,8	9,5	5	1
$12-15$	10	50	0,2	1	13,5	7	3,33

$$\begin{array}{ll} 7 \text{ a } 12 & 5 \\ 7 \text{ a } 11 & x \end{array} \Leftrightarrow \begin{array}{ll} 5 & 5 \\ 4 & x \end{array} \Rightarrow x = \frac{4 \times 5}{5} = 4$$

50 -(10 + 25+4) = 50-39=11 observaciones tomaron un valor superior a 11

Intervalos	n_i	N_i	f_i	F_i	x_i	a_i	n_i'
$0-5$	10	10	0,2	0,3	2,5	5	2
$5-7$	25	35	0,5	0,7	6	2	12,5
$7-12$	5	40	0,1	0,8	9,5	5	1
$12-15$	10	50	0,2	1	13,5	7	3,33

3. Cuartiles:

$$Q_1 = l_{i-1} + \frac{n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{12,5 - 10}{25} \cdot 2 = 5,2$$

$$Q_2 = l_{i-1} + \frac{2\,n/4 - N_{i-1}}{n_i} \cdot a_i = 5 + \frac{25 - 10}{25} \cdot 2 = 6,2$$

$$Q_3 = l_{i-1} + \frac{3\,n/4 - N_{i-1}}{n_i} \cdot a_i = 7 + \frac{37,5 - 35}{5} \cdot 5 = 9,5$$

Medidas de variabilidad o dispersión

Los estadísticos de *tendencia central* o *posición* nos indican donde se sitúa un grupo de puntuaciones. Los de *variabilidad* o *dispersión* nos indican si esas puntuaciones o valores están próximas entre sí o si por el contrario están o muy dispersas.

Rango

Una medida razonable de la variabilidad podría ser la **amplitud** o **rango**, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto.

2.4.1. Rango

Una medida razonable de la variabilidad podría ser la **amplitud** o **rango**, que se obtiene restando el valor más bajo de un conjunto de observaciones del valor más alto.

Propiedades del rango

- Es fácil de calcular y sus unidades son las mismas que las de la variable.
- No utiliza todas las observaciones (sólo dos de ellas);
- Se puede ver muy afectada por alguna observación extrema;
- El rango aumenta con el número de observaciones, o bien se queda igual. En cualquier caso nunca disminuye.

Varianza

La **varianza**, $\mathcal{S}^2$, se define como la media de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética, es decir

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Esta medida es siempre una cantidad positiva, con propiedades interesante para la realización de inferencia estadística. Como sus unidades son las del cuadrado de la variable, es más sencillo usar su raíz cuadrada, que es la que vemos en la siguiente sección.

Desviación típica o estándar

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en metros cuadrados. Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la **desviación típica**, $\mathcal{S}$, como

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

Ejemplo de cálculo de medidas de dispersión

Calcular el rango, varianza y desviación típica de las siguientes cantidades medidas en metros:

$$3, 3, 4, 4, 5$$

Solución: El rango de esas observaciones es la diferencia entre la mayor y menor de ellas, es decir, $5 - 3 = 2$. Para calcular las restantes medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a medir las diferencias. Éste es la media:

$$\overline{x} = (3 + 3 + 4 + 4 + 5)/5 = 3,8 \text{ metros}$$

$$3, 3, 4, 4, 5$$

La varianza es:

$$\mathcal{S}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{5}\left(3^2 + 3^2 + 4^2 + 4^2 + 5^2\right) - 3,8^2 = 0,56 \text{ metros}^2$$

siendo la desviación típica su raíz cuadrada:

$$\mathcal{S} = \sqrt{\mathcal{S}^2} = \sqrt{0,56} = 0,748 \text{ metros}$$

Propiedades de la varianza y desviacion típica

- Ambas son sensibles a la variación de cada una de las puntuaciones, es decir, si una puntuación cambia, cambia con ella la varianza. La razón es que si miramos su definición, la varianza es función de *cada una de las puntuaciones.*

- *La desviación típica tiene la propiedad de que en el intervalo*

$$(\overline{x} - 2\,\mathcal{S}, \overline{x} + 2\,\mathcal{S}) \overset{\text{def}}{\sim} \overline{x} \pm 2\,\mathcal{S}$$

  se encuentra, al menos, el 75 % de las observaciones Incluso si tenemos muchos datos y estos provienen de una distribución normal *(se definirá este concepto más adelante), podremos llegar al 95 %.*

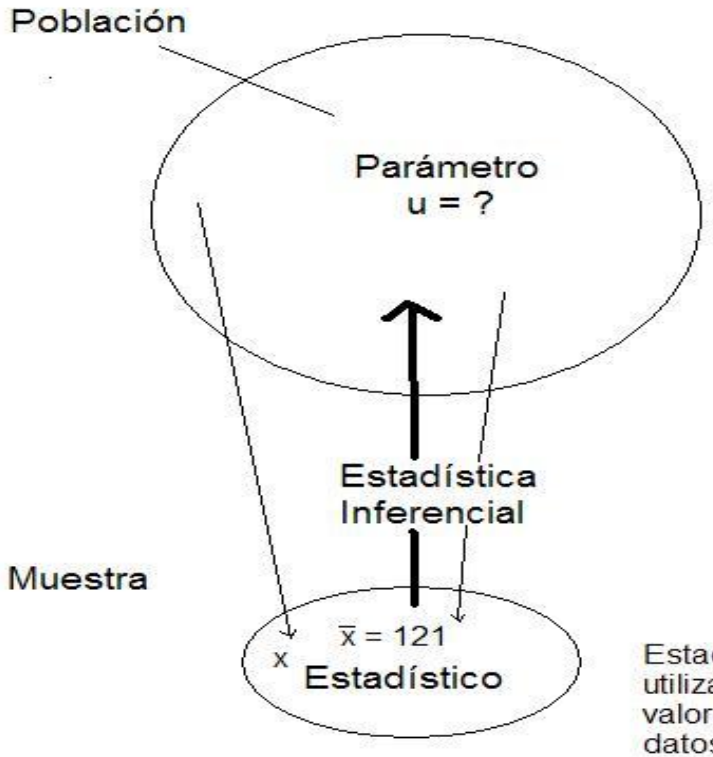
- *No es recomendable el uso de ellas, cuando tampoco lo sea el de la media como medida de tendencia central.*

Coeficiente de variación

Hemos visto que las medidas de centralización y dispersión nos dan información sobre una muestra. Nos podemos preguntar si tiene sentido usar estas magnitudes para comparar dos poblaciones. Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de elefantes de dos circos diferentes, $\mathcal{S}$ nos dará información útil.

¿Pero qué ocurre si lo que comparamos es la altura de unos elefantes con respecto a su peso? Tanto la media como la desviación típica, $\overline{x}$ y $\mathcal{S}$, se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo de que una de las medidas sea de longitud y la otra sea de masa. El mismo problema se plantea si medimos cierta cantidad, por ejemplo la masa, de dos poblaciones, pero con distintas unidades. Este es el caso en que comparamos el peso en *toneladas* de una población de 100 elefantes con el correspondiente en *miligramos* de una población de 50 hormigas.



Estadística descriptiva utilizada para obtener el valor de $\bar{x}$ a partir de los datos

El problema no se resuelve tomando las mismas escalas para ambas poblaciones. Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que los elefantes (toneladas). Si la ingeriería genética no nos sorprende con alguna barbaridad, lo lógico es que la dispersión de la variable *peso de las hormigas* sea practicamente nula (¡Aunque haya algunas que sean 1.000 veces mayores que otras!)

En los dos primeros casos mencionados anteriormente, el problema viene de la *dimensionalidad* de las variables, y en el tercero de la diferencia enorme entre las medias de ambas poblaciones. El *coeficiente de variación* es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica. Se define del siguiente modo:

Coeficiente de variación

$$CV = \frac{S_X}{\overline{x}}$$

Propiedades del coeficiente de variación

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que $\overline{x} > 0$.

- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, entonces $\mathcal{CV}_Y < \mathcal{CV}_X$.

- Es invariante a cambios de escala. Así por ejemplo el coeficiente de variación de una variable medida en metros es una cantidad adimensional que no cambia si la medición se realiza en centímetros.

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

$$\boxed{Z = \frac{X - \overline{x}}{\mathcal{S}}}$$

de media $\overline{z} = 0$ y desviación típica $\mathcal{S}_Z = 1$, que denominamos **variable tipificada**.

Esta nueva variable 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. También es aplicable al caso en que se quieran comparar individuos semejantes de poblaciones diferentes. Por ejemplo si deseamos comparar el nivel académico de dos estudiantes de diferentes Universidades para la concesión de una beca de estudios, en principio sería injusto concederla directamente al que posea una nota media más elevada, ya que la dificultad para conseguir una buena calificación puede ser mucho mayor en un centro que en el otro, lo que limita las posibilidades de uno de los estudiante y favorece al otro. En este caso, lo más correcto es comparar las calificaciones de ambos estudiantes, pero tipificadas cada una de ellas por las medias y desviaciones típicas respectivas de las notas de los alumnos de cada Universidad.

No confundir coeficiente de variación y tipificación

Los *coefientes de variación* sirven para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones), mientras que si deseamos comparar a dos *individuos* de cada uno de esos conjuntos, es necesario usar los *valores tipificados*. Ninguno de ellos posee unidades y es un error frecuente entre estudiantes de bioestadística confundirlos.

Asimetría y apuntamiento

Sabemos cómo calcular valores alrededor de los cuales se distribuyen las observaciones de una variable sobre una muestra y sabemos cómo calcular la dispersión que ofrecen los mismos con respecto al valor de central. Nos

proponemos dar un paso más allá en el análisis de la variable. En primer lugar, nos vamos a plantear el saber si los datos se distribuyen de forma simétrica con respecto a un valor central, o si bien la gráfica que representa la distribución de frecuencias es *de una forma diferente del lado derecho que del lado izquierdo.*

Si la simetría ha sido determinada, podemos preguntarnos si la curva es más o menos *apuntada* (larga y estrecha). Este apuntamiento habrá que medirlo comparado a cierta distribución de frecuencias que consideramos *normal* (no por casualidad es éste el nombre que recibe la distribución de referencia).

Estadísticos de asimetría

Para saber si una distribución de frecuencias es simétrica, hay que precisar con respecto a qué. Un buen candidato es la mediana, ya que para variables continuas, divide al histograma de frecuencias en dos partes de igual área. Podemos basarnos en ella para, de forma natural, decir que **una distribución de frecuencias es simétrica** si el lado derecho de la gráfica (a partir de la mediana) es la imagen por un espejo del lado izquierdo(figura).

Estadísticos de asimetría

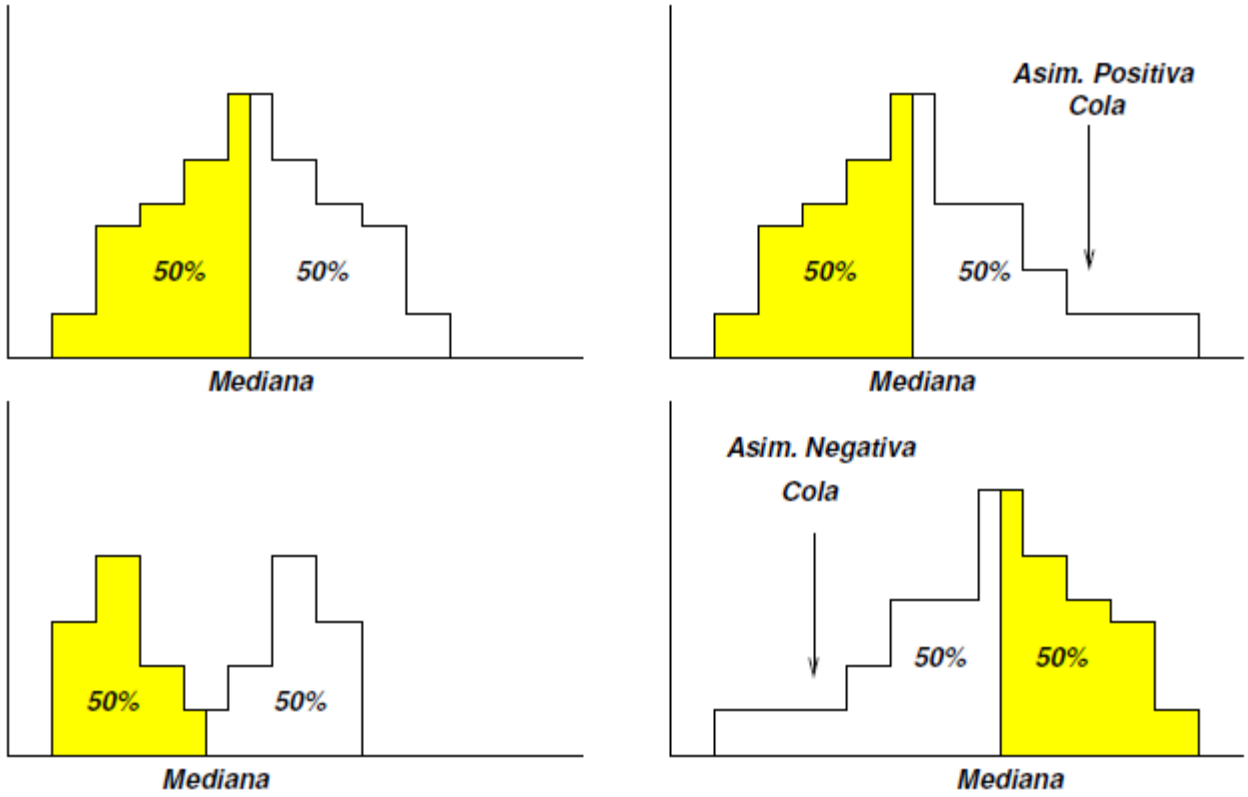
Cuando la variable es discreta, decimos que es simétrica, si lo es con respecto a la media.

Dentro de los tipos de asimetría posible, vamos a destacar los dos fundamentales:

Asimetría positiva: Si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en derecho hay frecuencias más pequeñas (*cola*).

Asimetría negativa: Cuando la cola está en el lado izquierdo.

Cuando realizamos un estudio descriptivo es altamente improbable que la distribución de frecuencias sea totalmente simétrica. En la práctica diremos que la distribución de frecuencias es simétrica si lo es de un modo



Distribuciones de frecuencias simétricas y asimétricas

Por otro lado, aún observando cuidadosamente la gráfica, podemos no ver claro de qué lado están las frecuencias más altas. Se definen entonces toda una familia de estadísticos que ayuden a interpretar la asimetría, denominados **índices de asimetría**. El principal de ellos es el *momento central de tercer orden* que definimos a continuación.

Momento central de tercer orden

Sea X una variable cuantitativa y $p \in I\!N$. Llamamos **momento** de orden p a:

$$\mu_p = \frac{1}{n}\sum_{i=1}^{n} x_i^p$$

Se denomina **momento central** de orden p a la cantidad

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

Los momentos de orden p impar, son siempre nulos en el caso de variables simétricas, ya que para cada i que esté a un lado de la media, con $(x_i - \overline{x}) < 0$, le corresponde una observación j del otro lado de la media tal que $(x_j - \overline{x}) = -(x_i - \overline{x})$. Elevando cada una de esas cantidades a p impar, y sumando se tiene que

$$m_p = 0 \qquad \text{si la distribución es simétrica.}$$

Si la distribución fuese asimétrica positiva, las cantidades $(x_i - \overline{x})^p$, con $p \geq 3$ impar positivas estarían muy aumentadas al elevarse a p. Esta propiedad nos indica que un índice de asimetría posible consiste en tomar $p = 3$ y elegir como estadístico de asimetría al momento central de tercer orden.

Apoyandonos en este índice, diremos que hay asimetría positiva si $a_3 > 0$, y que la asimetría es negativa si $a_3 < 0$.

Índice basado en los tres cuartiles (Yule–Bowley)

Si una distribución es simétrica, es claro que deben haber tantas observaciones entre la que deja por debajo de sí las tres cuartas partes de la distribución y la mediana, como entre la mediana y la que deja por debajo de sí un quarto de todas las observaciones. De forma abreviada esto es,

$$Q_3 - Q_2 = Q_2 - Q_1$$

Una pista para saber si una distribución de frecuencias es asimétrica positiva la descubrimos observando

$$Q_3 - Q_2 > Q_2 - Q_1$$

Por analogía, si es asimétrica negativa, se tendrá

$$Q_3 - Q_2 < Q_2 - Q_1$$

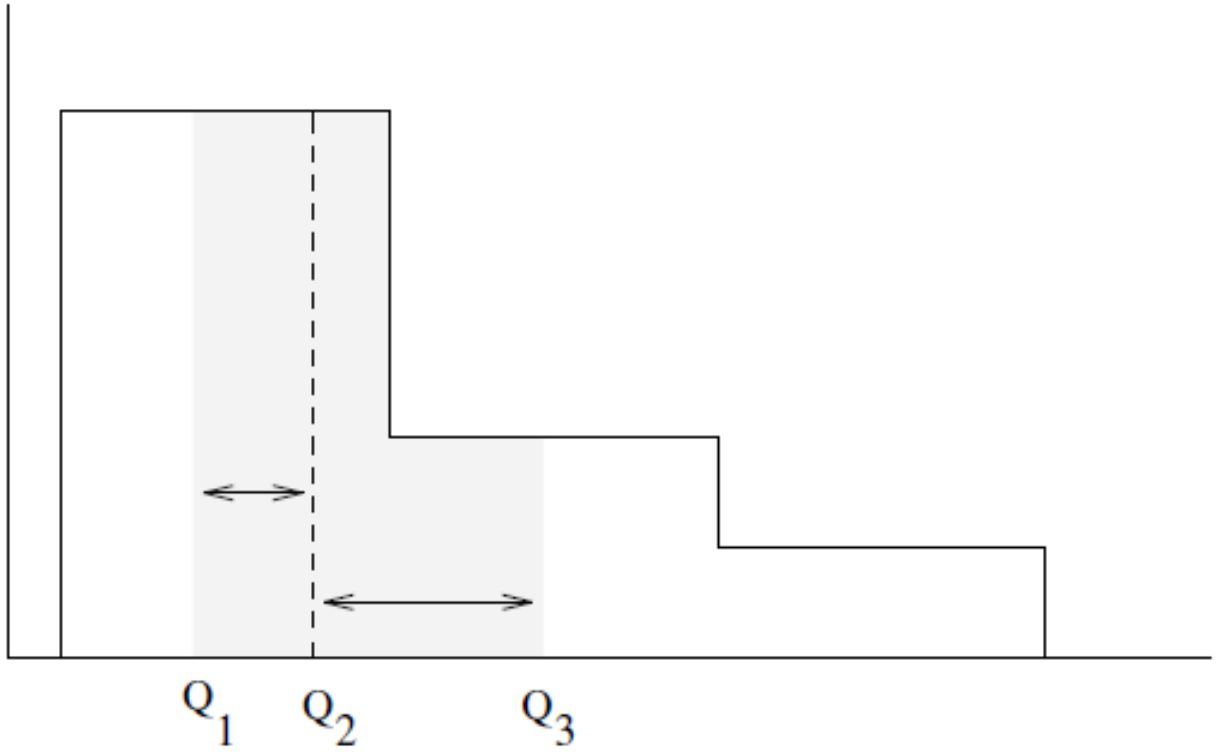
Para quitar dimensionalidad al problema, utilizamos como *índice de asimetría* la cantidad:

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

Es claro que

$$-1 \leq \mathcal{A}_s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \leq 1$$

El número obtenido, $\mathcal{A}_s$, es invariante ante cambios de origen de referencia y de escala.



Uso de los cuartiles para medir la asimetría

Otros índices de asimetría

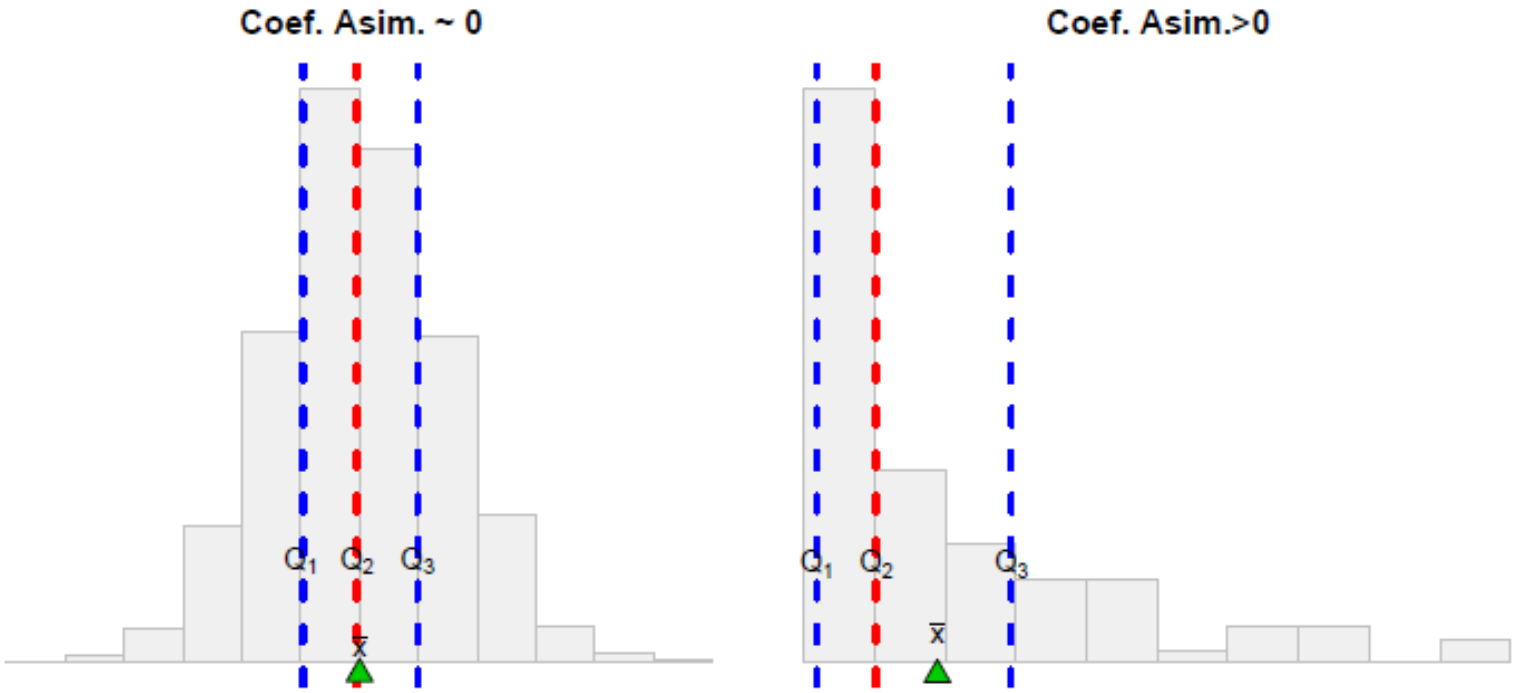
Basándonos en que si una distribución de frecuencias es simétrica y unimodal, entonces la media, la mediana y la moda coinciden, podemos definir otras medidas de asimetría, como son:

$$\boxed{\mathcal{A}_s = \frac{\overline{x} - M_{oda}}{\mathcal{S}}}$$

o bien,

$$\mathcal{A}_s = \frac{3(\overline{x} - M_{ed})}{\mathcal{S}}$$

Diremos que hay asimetría positiva si $\mathcal{A}_s > 0$ y negativa si $\mathcal{A}_s < 0$



Diferencias entre las medidas de tendencia central, o bien entre las distancias entre cuartiles consecutivos indican asimetría.

Ejemplo

Las edades de un grupo de personas se reflejan en la tabla siguiente:

Intervalos	n_i
7 — 9	4
9 — 11	18
11 — 12	14
12 — 13	27
13 — 14	42
14 — 15	31
15 — 17	20
17 — 19	1

Determinar la variabilidad de la edad mediante los estadísticos varianza, desviación típica, coeficiente de variación y rango intercuartílico. Estudie la simetría de la variable.

Solución:

En primer lugar realizamos los cálculos necesarios a partir de la tabla de frecuencias:

Intervalos	n_i	x_i	N_i	$x_i n_i$	$x_i^2 n_i$
7 — 9	4	8	4	32	256
9 — 11	18	10	22	180	1.800
11 — 12	14	11,5	36	161	1.851,5
12 — 13	27	12,5	63	337,5	4.218,75
13 — 14	42	13,5	105	567	7.654,5
14 — 15	31	14,5	136	449,5	6.517,75
15 — 17	20	16	156	320	5.120
17 — 19	1	18	157	18	324
	157			2.065	27.742,25

La media es $\overline{x} = 2{,}065/157 = 13,15$ años. La varianza la calculamos a partir de la columna de la $x_i^2 n_i$ como sigue:

$$\mathcal{S}^2 = 27{,}742,25/157 - 13,15^2 = 3,78 \text{ años}^2 \quad \Rightarrow \quad \mathcal{S} = \sqrt{3,78} = 1,94 \text{ años}$$

Intervalos	n_i	x_i	N_i	$x_i n_i$	$x_i^2 n_i$
7 — 9	4	8	4	32	256
9 — 11	18	10	22	180	1.800
11 — 12	14	11,5	36	161	1.851,5
12 — 13	27	12,5	63	337,5	4.218,75
13 — 14	42	13,5	105	567	7.654,5
14 — 15	31	14,5	136	449,5	6.517,75
15 — 17	20	16	156	320	5.120
17 — 19	1	18	157	18	324
	157			2.065	27.742,25

El coeficiente de variación no posee unidades y es:

$$\mathcal{CV} = \frac{1,94}{13,15} = 0,15 = 15\,\% \text{ de variabilidad.}$$

En lo que concierne a la simetría podemos utilizar el coeficiente de asimetría de Yule–Bowley, para el cual es preciso el cálculo de los cuartiles:

$$\mathcal{Q}_1 = 12 + \frac{39,25 - 36}{27} \times 1 = 12,12$$

$$M_{ed} = \mathcal{Q}_2 = 13 + \frac{78,5 - 63}{42} \times 1 = 13,37$$

Intervalos	n_i	x_i	N_i	$x_i n_i$	$x_i^2 n_i$
7 — 9	4	8	4	32	256
9 — 11	18	10	22	180	1.800
11 — 12	14	11,5	36	161	1.851,5
12 — 13	27	12,5	63	337,5	4.218,75
13 — 14	42	13,5	105	567	7.654,5
14 — 15	31	14,5	136	449,5	6.517,75
15 — 17	20	16	156	320	5.120
17 — 19	1	18	157	18	324
	157			2.065	27.742,25

$$Q_3 = 14 + \frac{117,75 - 105}{31} \times 1 = 14,41$$

Lo que nos dice que aproximadamente en un rango de $Q_3 - Q_1 = 2,29$ años se encuentra el 50 % central del total de observaciones[1] Además:

$$= \mathcal{A}_s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(14,41 - 13,37) - (13,37 - 12,12)}{14,41 - 12,12} = -0,09$$

Este resultado nos indica que existe una ligera asimetría a la izquierda (negativa). Un resultado similar se obtiene si observamos (Figura 2.9) que la distribución de frecuencias es unimodal, siendo la moda:

$$M_{oda} == 13 + \frac{42 - 27}{(42 - 27) + (42 - 31)} \times 1 = 13,57$$

en cuyo caso podemos usar como medida del sesgo:

$$\mathcal{A}_s = \frac{\overline{x} - M_{oda}}{\mathcal{S}} = \frac{13,15 - 13,57}{1,94} = -0,21$$

Estadísticos de apuntamiento

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

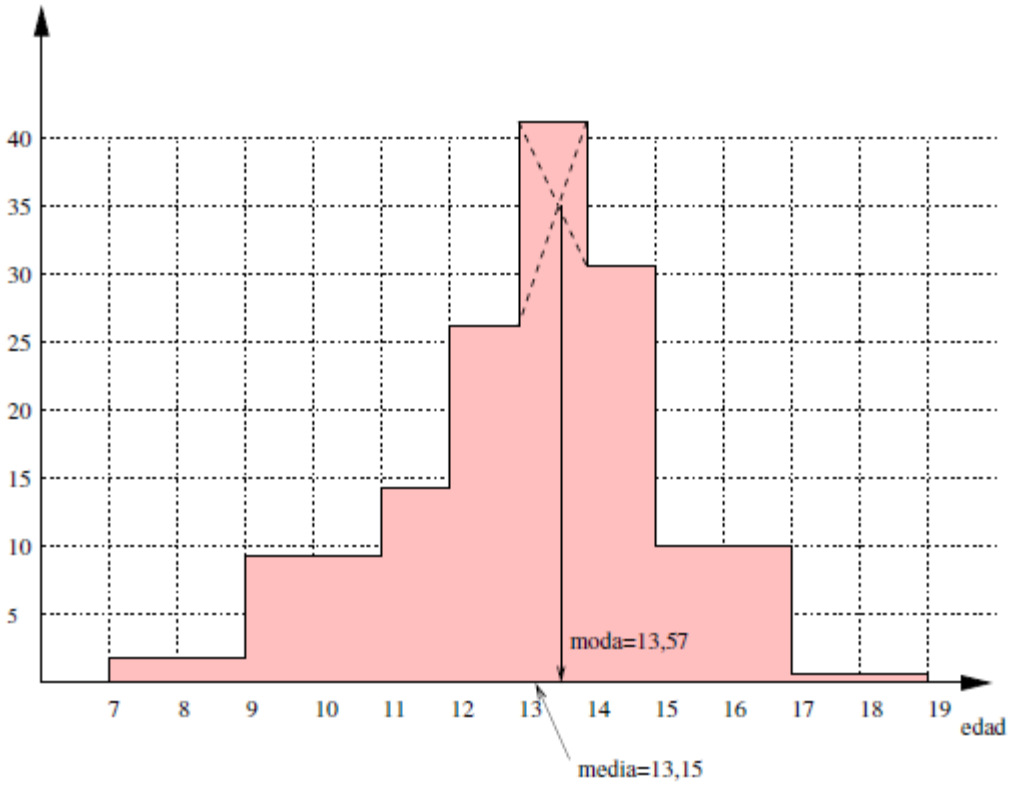
$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

donde m_4 es el momento empírico de cuarto orden. Es éste un coeficiente adimensional, invariante ante cámbios de escala y de origen. Sirve para medir si una distribución de frecuencias es muy apuntada o no. Para decir si la distribución es larga y estrecha, hay que tener un patrón de referencia. El patrón de referencia es la *distribución normal o gaussiana*[2] para la que se tiene

---

[1] Eso hace que dicha cantidad sea usada como medida de dispersión, denominándose **rango intercuartílico**.

[2] Será introducida posteriormente.



La distribución de frecuencias de la edad presenta una ligera asimetría negativa.

coeficiente de aplastamiento de Fisher (curtosis)

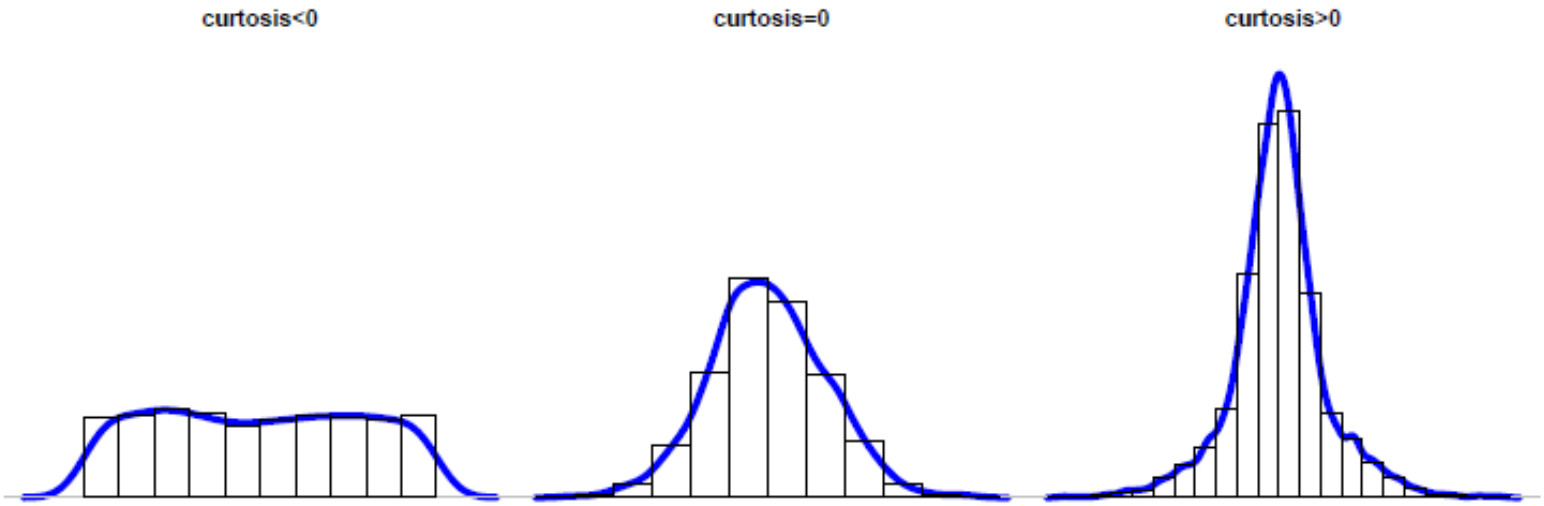
$$\frac{m_4}{\sigma^4} = 3 \Longrightarrow \gamma_2 = 0$$

De este modo, atendiendo a γ_2, se clasifican las distribuciones de frecuencias en

Leptocúrtica: Cuando $\gamma_2 > 0$, o sea, si la distribución de frecuencias es más apuntada que la normal;

Mesocúrtica: Cuando $\gamma_2 = 0$, es decir, cuando la distribución de frecuencias es tan apuntada como la normal;

Platicúrtica: Cuando $\gamma_2 < 0$, o sea, si la distribución de frecuencias es menos apuntada que la normal;



Apuntamiento de distribuciones de frecuencias

Problemas

Ejercicio 2.1. En el siguiente conjunto de números, se proporcionan los pesos (redondeados a la libra más próxima) de los bebés nacidos durante un cierto intervalo de tiempo en un hospital:

4, 8, 4, 6, 8, 6, 7, 7, 7, 8, 10, 9, 7, 6, 10, 8, 5, 9, 6, 3, 7, 6, 4, 7, 6, 9, 7, 4, 7, 6, 8, 8, 9, 11, 8, 7, 10, 8, 5, 7, 7, 6, 5, 10, 8, 9, 7, 5, 6, 5.

1. Construir una distribución de frecuencias de estos pesos.
2. Encontrar las frecuencias relativas.
3. Encontrar las frecuencias acumuladas.
4. Encontrar las frecuencias relativas acumuladas.
5. Dibujar un histograma con los datos de la parte **a**.
6. ¿Por qué se ha utilizado un histograma para representar estos datos, en lugar de una gráfica de barras?
7. Calcular las medidas de tendencia central.

8. Calcular las medidas de dispersión.

9. Calcular las medidas de forma.

10. ¿Es esta una distribución sesgada? De ser así, ¿en qué dirección?

11. Encontrar el percentil 24.

Ejercicio 2.2. A continuación se dan los resultados obtenidos con una muestra de 50 universitarios. la característica es el tiempo de reacción ante un estímulo auditivo:

0,110	0,110	0,126	0,112	0,117	0,113	0,135	0,107	0,122
0,113	0,098	0,122	0,105	0,103	0,119	0,100	0,117	0,113
0,124	0,118	0,132	0,108	0,115	0,120	0,107	0,123	0,109
0,117	0,111	0,112	0,101	0,112	0,111	0,119	0,103	0,100
0,108	0,120	0,099	0,102	0,129	0,115	0,121	0,130	0,134
0,118	0,106	0,128	0,094	0,1114				

1. ¿Cuál es la amplitud total de la distribución de los datos?

2. Obtenga la distribución de frecuencias absolutas y relativas.

3. Obtenga la distribución de frecuencias acumuladas, absolutas y relativas, con los intervalos anteriores.

4. Calcular la media y la varianza con los intervalos del apartado **b** y después calculense las mismas magnitudes sin ordenar los datos en una tabla estadística.¿Con qué método se obtiene mayor precisión?

5. Dibuje el polígono de frecuencias relativas.

6. Dibuje el polígono de frecuencias relativas acumuladas.

Ejercicio 2.3. Con el fin de observar la relación entre la inteligencia y el nivel socioeconómico (medido por el salario mensual familiar) se tomaron dos grupos, uno formado con sujetos de cociente intelectual inferior a 95 y otro formado por los demás; De cada sujeto se anotó el salario mensual familiar. Teniendo en cuenta los resultados que se indican en la tabla:

Nivel socioeconómico	Sujetos con $CI < 95$	Sujetos con $CI \geq 95$
Intervalos	Frecuencia	Frecuencia
10 o menos $\equiv(4,10]$	75	19
10 – 16	35	26
16 – 22	20	25
22 – 28	30	30
28 – 34	25	54
más de 34 $\equiv(34,40]$	15	46

1. Dibuje un gráfico que permita comparar ambos grupos.
2. Calcule las medidas de tendencia central para aquellos sujetos con $CI < 95$.
3. Calcular las medidas de dispersión para aquellos sujetos con $CI \geq 95$.

Ejercicio 2.4. Un estudio consistió en anotar el número de palabras leídas en 15 segundos por un grupo de 120 sujetos disléxicos y 120 individuos normales. Teniendo en cuenta los resultados de la tabla

$N°$ de palabras leídas	Disléxicos n_D	Normales n_N
25 o menos $\equiv 25$	56	1
26	24	9
27	16	21
28	12	29
29	10	28
30 o más $\equiv 30$	2	32

calcule:

1. Las medias aritméticas de ambos grupos.
2. Las medianas de ambos grupos.
3. El porcentaje de sujetos disléxicos que superaron la mediana de los normales.
4. Compare la variabilidad relativa de ambos grupos.

Ejercicio 2.5. La tabla siguiente muestra la composición por edad, sexo y trabajo de un grupo de personas con tuberculosis pulmonar en la provincia de Vizcaya en el año 1979:

Edad	Trabajadores			No trabajadores			Totales		
	Varón	Mujer	Total	Varón	Mujer	Total	Varón	Mujer	Total
14–19	2	1	3	25	40	65	27	41	68
19–24	10	4	14	20	36	56	30	40	70
24–29	32	10	42	15	50	65	47	60	107
29–34	47	12	59	13	34	47	60	46	106
34–39	38	8	46	10	25	35	48	33	81
39–44	22	4	26	7	18	25	29	22	51

1. Representar gráficamente la distribución de frecuencias de aquellas personas trabajadoras que padecen tuberculosis.

2. Representar gráficamente la distribución de frecuencias de los varones no trabajadores que padecen tuberculosis.

3. Representar gráficamente la distribución de frecuencias del número total de mujeres que padecen tuberculosis.

4. ¿Cuál es la edad en la que se observa con mayor frecuencia que no trabajan los varones? ¿Y las mujeres? Determinar asímismo la edad más frecuente (sin distinción de sexos ni ocupación).

5. ¿Por debajo de qué edad está el 50 % de los varones?

6. ¿Por encima de qué edad se encuentra el 80 % de las mujeres?

7. Obtener la media, mediana y desviación típica de la distribución de las edades de la muestra total.

8. Estudiar la asimetría de las tres distribuciones.

Ejercicio 2.6. En una epidemia de escarlatina, se ha recogido el número de muertos en 40 ciudades de un país, obteniéndose la siguiente tabla:

N° de muertos	0	1	2	3	4	5	6	7
Ciudades	7	11	10	7	1	2	1	1

1. Representar gráficamente estos datos.
2. Obtener la distribución acumulada y representarla.
3. Calcular media, mediana y moda.
4. Calcular la varianza y la desviación típica.
5. Porcentaje de ciudades con al menos 2 muertos.
6. Porcentaje de ciudades con más de 3 muertos.
7. Porcentaje de ciudades con a lo sumo 5 muertos.