



Estadística: Medidas estadísticas

Matemáticas I

1.º Bachillerato

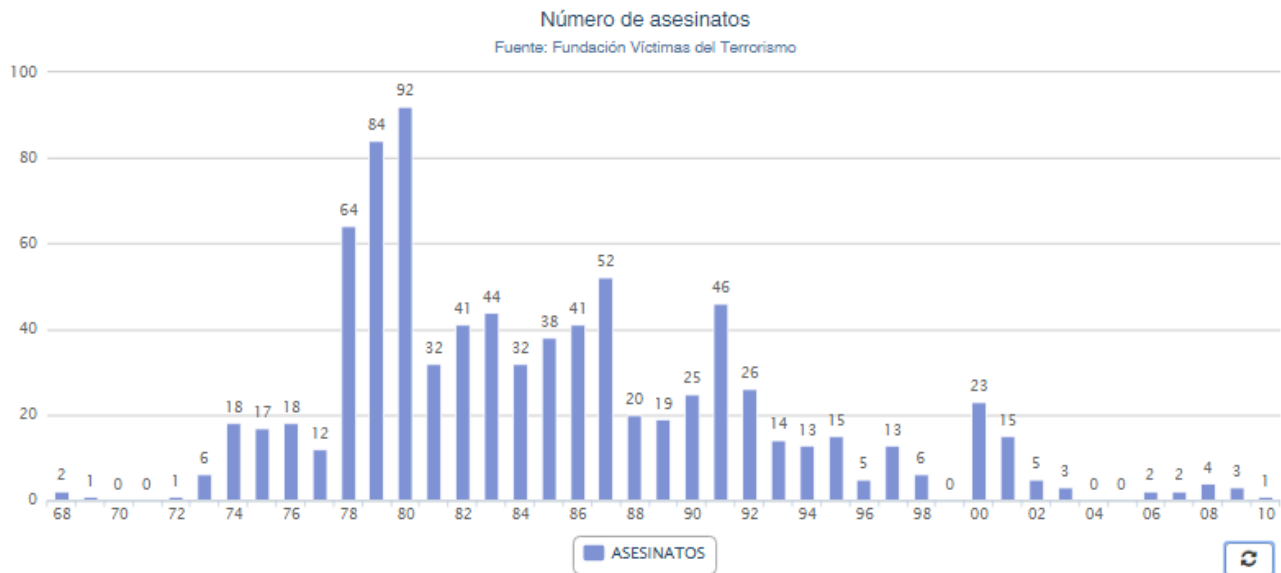
Contenidos

Estadística
Medidas estadísticas

1. Introducción

Normalmente en los estudios estadísticos se manejan gran cantidad de datos, por eso es imprescindible resumirlos de manera que se conserve la mayor información posible y que representen el comportamiento global de la población.

Veamos como ejemplo una infografía del diario [20minutos.es](https://www.20minutos.es) sobre las víctimas de ETA.



Observa que entre 1968 y 2010 la banda asesinó a 855 personas y que 1980 fue el año más sangriento con 92 víctimas. Esta información tan concisa, tan representativa de la barbarie, es gracias a los parámetros o medidas estadísticas.



Importante

Los **parámetros estadísticos** son datos globales que nos informan sobre características de la población estudiada.

2. Medidas de centralización

Para hacernos una idea en tono de humor de los que son los primeros parametros que estudiaremos, aquí tenéis el siguiente video, que es un clip de la fantástica serie *The Big Bang Theory* (haz clic sobre la imagen para ir al vídeo).



Imagen de ChrisJohn23 en [Flickr](#). Licencia [CC by-nc-sa 2.0](#)



Importante

Las **medidas de centralización** son datos que representan de forma global a toda la población, y entorno a los cuales están agrupados todos los valores.

Las más importantes son la **media**, la **mediana** y la **moda**.

Estas medidas de centralización no son conceptos desconocidos, muy al contrario, aparecen continuamente en nuestra vida.



A los niños pequeños, uno de los primeros conceptos que se les enseña son los tamaños. A muy corta edad aprenden la diferencia entre grande, **mediano** y pequeño.

Imagen de Rufus
Gefangen en [Flickr](#).
Licencia [CC by-nc-nd 2.0](#)

Hoy en día ir a la moda es una de las preocupaciones de mucha gente.



Para no quedarte desfasado tienes que estar a la **moda**, es decir, en este caso, vestirse de la misma forma que mucha otra gente.

Imagen de Sarfires en [Flickr](#).
Licencia [CC by-nc-sa 2.0](#)



¿Sabías que van a situar dos radares seguidos y a cierta distancia, para de esa forma calcular la velocidad **media** a la que circulamos, y así poder comprobar si infringimos la ley?

Imagen de photoAtlas en [Flickr](#). Licencia [CC by-nc 2.0](#)



Importante

Decimos que la **media** se corresponde con la idea de repartir todo lo que hay en partes iguales para todos.

La media es el promedio aritmético de las observaciones, es decir, el cociente entre la suma de todos los datos y el número total de ellos. Se representa por \bar{x} . Se calcula:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_i \cdot n_i}{N}$$

Sólo tiene sentido en variables cuantitativas, pues las x_i no son números en las variables cualitativas.



Comprueba lo aprendido

Las medias o promedios son valores que se toman muy en cuenta en muchos deportes, en particular en baloncesto.

Uno de los mejores quintetos de la selección española podría estar compuesto por Ricky Rubio, Juan Carlos Navarro, Rudy Fernández, Felipe Reyes y Marc Gasol.

Los datos del último partido amistoso fueron:

	Puntos	Rebotes	Asistencias
Ricky Rubio	7	6	3
J. Carlos Navarro	20	2	6
Rudy Fernández	5	3	1
Felipe Reyes	16	2	0
Marc Gasol	17	6	2

¿Sabrías decirme cuáles son los promedios del quinteto español?

La media de puntos por partido es puntos.

La media de rebotes por partido es de rebotes.

La media de asistencias por partido es de asistencias.

Es fácil, sólo es cuestión de sumar y dividir.



Importante

Si ordenamos los datos de menor a mayor, la **mediana**, es el valor que está en medio, es decir, tiene tantos valores a la izquierda como a la derecha. Se representa por **M_e** .

Esta medida tiene sentido calcularla para cualquier tipo de variable cuantitativa. Destacar que cuando los datos se agrupan en intervalos se calcula el **intervalo mediano**.

El **intervalo mediano** es aquel que verifica que la frecuencia absoluta acumulada en su extremo inferior (N_{i-1}) es menor que $N/2$ y la frecuencia absoluta acumulada en su extremo superior (N_i) es mayor que $N/2$. De manera equivalente el intervalo mediano verifica que $N_{i-1}/N < 1/2$ y $N_i/N > 1/2$.



Comprueba lo aprendido



Importante

Decimos que la **moda** es el valor que más se repite, es decir el de mayor frecuencia absoluta. Se representa por **M_o** .

Esta medida tiene sentido calcularla para cualquier tipo de variable.

Claramente cuando la variable es cuantitativa continua no tiene sentido hablar de valor más frecuente, si no, de **intervalo modal**.

En caso de que varias modalidades tengan la frecuencia máxima, se habla de **distribución multimoda** (con 2 o más modas).



Caso práctico

En el tema anterior te comentamos que el Instituto Nacional de Estadística realizó en 2008 una encuesta para conocer el número de personas mayores de 25 años que a lo largo de 2007 participó en algún tipo de actividad formativa. Para saber más sobre dicha encuesta pulsa [aquí](#).

De los datos más generales de dicha encuesta hemos calculado las medidas de centralización. En la siguiente presentación puedes ver los pasos que hemos dado.

http://www.slideshare.net/slideshow/embed_code/5070896

[Medidas de centralización](#) from [Jesús Fernández](#)



Reflexiona

El estudio realizado por el INE sobre la participación de la población adulta en actividades de aprendizaje a lo largo de 2007, distingue entre actividades que pertenezcan a una **enseñanza formal** (las que conllevan la obtención de un título oficial), o **no formal** (las que no persiguen la obtención de titulación oficial).

En la tabla siguiente aparecen desglosados los datos.

Tramos de edad	Personas que cursaron enseñanza formal	Personas que cursaron enseñanza no formal
[25, 35)	896270	2475505
[35, 45)	334895	2224578
[45, 55)	200511	1496676
[55, 65)	85918	742246
[65, 75)	26912	277585

Siguiendo los pasos que se dan en la presentación anterior, halla la media y los intervalos modales y medianos de cada tipo de enseñanza. Una vez calculados dichos parámetros, ¿existen algunas diferencias entre los valores obtenidos para cada una de las modalidades de enseñanza? ¿qué justificación darías a estas diferencias?

Para la enseñanza formal la media es 37 años, el intervalo modal y mediano [25, 35).

Para la no formal, la media es 42 años, el intervalo modal [25, 35), y el mediano [35, 45).

La diferencia que se aprecia es una cierta tendencia a medidas de centralización menores para la enseñanza formal. Lo que implicaría que es más joven la población que se inclina por este tipo de enseñanza.



Curiosidad

El símbolo Σ se utiliza para indicar de forma abreviada la suma de varios números.

Por ejemplo, para acortar la siguiente suma $x_1 + x_2 + x_3 + x_4$, escribimos: $\sum_{i=1}^4 x_i$.

Si queremos abreviar la suma $x_1 \cdot f_1 + x_2 \cdot f_2 + x_3 \cdot f_3 + \dots + x_n \cdot f_n$, escribimos: $\sum_{i=1}^n x_i \cdot f_i$.

3. Medidas de dispersión

Tan solamente tres ciudadanos estadounidenses (Bill Gates, Paul Allen y Warren Buffet) poseen, juntos, una fortuna superior al PIB de 42 naciones pobres, en las cuales viven 600 millones de habitantes. Las 356 personas más ricas del mundo disfrutan una riqueza que excede a la renta anual del 40% de la humanidad.

El siguiente vídeo trata de concienciarnos de esta situación.

[Enlace a recurso reproducible >> https://www.youtube.com/embed/FNSeSyTljw](https://www.youtube.com/embed/FNSeSyTljw)

Vídeo de Lord Campamocha alojado en [Youtube](#)

Al igual que pasa en el mundo, las variables estadísticas no siempre están bien repartidas, de tal forma que la media puede no ser representativa si hay valores dispares que le afecten. Para eso tenemos las **medidas de dispersión**, que nos informan de la concentración de esos datos dentro de la variable.



Importante

Las **medidas de dispersión** nos informan de hasta qué punto las medidas de centralización son representativas como síntesis de la información.

Las **medidas de dispersión** cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central.

Estudiaremos a continuación el **recorrido**, la **varianza**, la **desviación típica** y el **coeficiente de variación**.

En el capítulo anterior calculamos los promedios de la selección española de baloncesto. Si hubiéramos realizado un estudio de los datos de los jugadores en todos los partidos amistosos jugados antes del mundial, podríamos preguntarnos qué jugador es el más regular.

Una medida que nos da la regularidad es el **recorrido**, es decir, la diferencia entre los datos del peor partido y el mejor partido de un jugador.



Imagen de Berts @idar en [Flickr](#).
Licencia [CC](#)

El jugador que *menor recorrido* tenga será el más regular.



Importante

El **recorrido** de una variable es la diferencia entre el mayor y menor valor que toma esa variable. Se representa por **R**.



Imagen de artemuestra en [Flickr](#)

.Licencia [cc by-sa/2.0](#)

Imaginemos que estamos interesados en comprar en bolsa acciones de una determinada empresa. Para saber si es una inversión segura, investigamos la evolución del precio de las acciones en el último año. Por supuesto, hemos calculado previamente cuál ha sido el precio medio de las acciones y mirando el resto de los valores podremos decir que si muchos días el precio ha estado alejado de la media (por debajo o por encima), diremos que la inversión es arriesgada o volátil. Y por el contrario, si la mayoría de los días el precio ha estado cerca de la media diremos que la inversión es segura.

La medida estadística que nos indica esta variación de los datos respecto de la media es la **varianza**.



Importante

La **varianza** mide la dispersión de una muestra en función de la diferencia (la distancia) de cada uno de los elementos de la muestra con el valor medio de la misma. Se representa por **s²**.

Se calcula mediante la fórmula:

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

La **desviación típica** es la raíz cuadrada positiva de la varianza. Se representa por **$\sigma = \sqrt{s^2}$** .

Aunque tanto la varianza como la desviación típica son medidas de la variación con respecto a la media y tienen un significado semejante, la desviación típica se utiliza

con más frecuencia porque se expresa en las mismas unidades que los valores de las variables. Al contrario que la varianza, cuyos valores se expresan en unidades al cuadrado.

No debes asustarte al ver la fórmula que es necesario utilizar para calcular la varianza, y por tanto la desviación típica.

Recuerda que N no es más que la suma de las frecuencias absolutas, es decir $\sum_{i=1}^n n_i$, y \bar{x} es la media.

Además, realizando ciertos cálculos, es posible obtener una fórmula más operativa que la anterior:

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

Veamos con un ejemplo cómo se aplica la expresión anterior.



Caso práctico

Volvemos a la encuesta del INE sobre actividades de formación en el año 2007 de las personas adultas en España. Veamos en la siguiente presentación los pasos que hay que dar para calcular la varianza y la desviación típica de los datos obtenidos en dicha encuesta.

http://www.slideshare.net/slideshow/embed_code/5072864

[Medidas de dispersion](#) from [Jesús Fernández](#)



Curiosidad

La varianza es el nombre técnico que dan los jugadores de poker a una racha de mala suerte. La varianza justifica que aunque un jugador esté jugando bien vaya perdiendo: todas sus acciones tienen EV (valor esperado) positivo, pero su banca registra pérdidas.

Los jugadores dicen (decimos) que no importa perder porque "ha sido culpa de la varianza". El buen jugador de póker parece que sólo tiene un enemigo: **la varianza**.

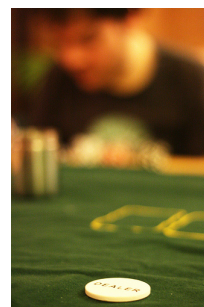


Imagen de obo-bobolina

Para defenderse de la varianza lo más común es mejorar la gestión de los recursos propios (banca), para que "un zarpazo de la varianza" no termine en "bancarrota". Es decir que una racha negativa no acabe con toda nuestra banca y podamos seguir jugando hasta que llegue una racha positiva.

en [Flickr](#). Licencia [cc by-nc/2.0](#)

Siguiendo con el ejemplo de la inversión en bolsa, nos encontramos con dos empresas que muestran la misma desviación típica en el precio de sus acciones pero el precio medio de las acciones es diferente. ¿En cuál de las dos empresas deberíamos invertir?

Existe una medida estadística que evalúa la dispersión de los datos (volatilidad del precio) respecto de la media (precio medio de las acciones) que se llama **coeficiente de variación**.



Importante

El **coeficiente de variación** es la medida de dispersión más popular y se define como el cociente entre la desviación típica y la media. Se representa por **CV**.

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

Cuanto mayor sea el coeficiente más dispersión hay en la variable y menos representativa es la media.



Comprueba lo aprendido

Supongamos las dos empresas anteriores con los siguientes datos:

	Empresa 1	Empresa 2
Desviación típica	10€	10€
Media (precio medio por acción)	50€	20€

¿En cuál de las dos empresas debemos invertir?

En la empresa ya que su CV= que es que el CV de la empresa .

Como queremos una inversión lo más segura posible, cuanta menos variabilidad más seguridad.



Reflexiona

En el apartado anterior veíamos como el INE desglosaba la encuesta sobre actividades de formación, en enseñanzas formales y no formales. Los datos eran los siguientes:

Tramos de edad	Personas que cursaron enseñanza formal	Personas que cursaron enseñanza no formal
[25, 35)	896270	2475505
[35, 45)	334895	2224578
[45, 55)	200511	1496676
[55, 65)	85918	742246
[65, 75)	26912	277585

Halla la desviación típica y el coeficiente de variación para cada una de las modalidades de enseñanza. Compara los resultados que hayas obtenido.

Para la enseñanza formal la desviación típica es 10 años, como la media era 37, tenemos que el coeficiente de variación vale 0,27.

En el caso de la enseñanza no formal, la desviación típica es 11, la media era 42, por lo que el coeficiente de variación que se obtiene es 0,269.

Es curioso, pero al obtener coeficientes de variación muy cercanos, podemos decir que los datos de ambas encuestas están distribuidos de forma muy parecida.

4. Percentiles y cuartiles

¡Qué ilusionante es tener un hijo! Pensar cómo será, a quién se parecerá,... hasta que nace, claro.

Entonces llega el pediatra y te dice las palabras fatídicas: "este niño tiene un percentil muy bajo". Y claro, uno dice: "no sé qué será eso, pero suena mal. Mi niño tiene que tener un percentil más alto". Llega el momento de "torturar" al pobre niño dándole más de comer y haciéndole masajes de estiramiento para que crezca más rápido y dejar al pediatra con la boca abierta en la siguiente visita.

Todo esto es una exageración, pero ¿sabemos lo que son los percentiles?, el siguiente video nos da una idea.

[Enlace a recurso reproducible >> http://www.youtube.com/embed/JpFhxI4LPV8](http://www.youtube.com/embed/JpFhxI4LPV8)

Vídeo de IEDA Andalucía alojado en [Youtube](#)



Importante

Los **percentiles** son medidas de posición que generalizan el concepto de mediana, y dividen el conjunto de observaciones en 100 partes de igual frecuencia.

Además de los percentiles, en estadística también se utilizan los **cuartiles** y los **deciles** (que dividen el conjunto de las observaciones en 4 y 10 partes de igual frecuencia, respectivamente).

Los percentiles se representan por q_p donde p es un número entre 0 y 0,99 con dos decimales (Por ejemplo, $q_{0,95}$ es el percentil 95).

Los cuartiles se representan por Q_1 , Q_2 y Q_3 .



Comprueba lo aprendido

En la edición digital de el diario [El Mundo](#) apareció publicado el siguiente titular (para acceder a la noticia completa, haz clic sobre el titular).



'Newsweek' sitúa a España en el puesto 21 entre los 100 mejores países por calidad de vida

¿Qué te parece la posición en que queda España? Teniendo en cuenta el lugar en que queda, completa las siguientes frases.

Según la encuesta de Nesweek, España se encuentra en el cuartil , en el decil , y en el percentil q .

No es difícil, sólo es cuestión de contar bien.

Hasta el momento sólo hemos calculado datos y más datos. Supongo que alguno echará en falta algún gráfico como en el tema anterior. Porque seamos sinceros, *una imagen vale más que mil palabras* (y en estadística más). Por eso, a continuación, y para finalizar el tema veremos los llamados **diagramas de cajas y bigotes**, que se utiliza para representar de forma visual la concentración de datos.



Caso de estudio

Hemos cogido 20 personas aleatoriamente y les hemos preguntado su edad. Obteniendo los siguientes datos:

36 25 37 24 39 20 36 45 31 31 39 24 29 23 41 40 33 24 34 40

Calculemos un diagrama de caja y bigotes.

Para calcular los parámetros estadísticos necesarios, lo primero es **ordenar la distribución**

20 23 24 24 24 25 29 31 31 33 34 36 36 37 39 39 40 40 41 45

A continuación hay que **calcular los 3 cuartiles**

Q_1 , el cuartil Primero es el valor mayor que el 25% de los valores de la distribución. Como $N = 20$ resulta que $N/4 = 5$; el primer cuartil es la media aritmética de dicho valor y el siguiente:

$$Q_1 = (24 + 25) / 2 = 24,5$$

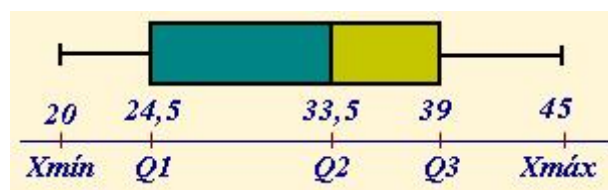
Q₂, el Segundo Cuartil es, evidentemente, la mediana de la distribución, es el valor de la variable que ocupa el lugar central en un conjunto de datos ordenados. Como $N/2 = 10$; la mediana es la media aritmética de dicho valor y el siguiente:

$$m_e = Q_2 = (33 + 34) / 2 = 33,5$$

Q₃ , el Tercer Cuartil, es el valor que sobrepasa al 75% de los valores de la distribución. En nuestro caso, como $3N / 4 = 15$, resulta

$$Q_3 = (39 + 39) / 2 = 39$$

Por último dibujaremos el diagrama



El *bigote* de la izquierda representa al colectivo de edades (**X_{mín}, Q₁**)

La primera parte de la caja a (**Q₁, Q₂**),

La segunda parte de la caja a (**Q₂, Q₃**)

El *bigote* de la derecha viene dado por (**Q₃, X_{máx}**).

Resumen



Importante

Los **parámetros estadísticos** son datos globales que nos informan sobre características de la población estudiada.



Importante

Actividad

Las **medidas de centralización** son datos que representan de forma global a toda la población, y entorno a los cuales están agrupados todos los valores.

Las más importantes son la **media**, la **mediana** y la **moda**.

- Decimos que la **media** se corresponde con la idea de repartir todo lo que hay en partes iguales para todos.

La media es el promedio aritmético de las observaciones, es decir, el cociente entre la suma de todos los datos y el número total de ellos. Se representa por \bar{x} . Se calcula:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_i \cdot n_i}{N}$$

Sólo tiene sentido en variables cuantitativas, pues las x_i no son números en las variables cualitativas.

- Si ordenamos los datos de menor a mayor, la **mediana**, es el valor que está en medio, es decir, tiene tantos valores a la izquierda como a la derecha. Se representa por **M_e** .

Esta medida tiene sentido calcularla para cualquier tipo de variable cuantitativa. Destacar que cuando los datos se agrupan en intervalos se calcula el **intervalo mediano**.

El **intervalo mediano** es aquel que verifica que la frecuencia absoluta acumulada en su extremo inferior (N_{i-1}) es menor que $N/2$ y la frecuencia absoluta acumulada en

su extremo superior (N_i) es mayor que $N/2$. De manera equivalente el intervalo mediano verifica que $N_{i-1}/N < 1/2$ y $N_i/N > 1/2$.

- Decimos que la **moda** es el valor que más se repite, es decir el de mayor frecuencia absoluta. Se representa por **M_o** .

Esta medida tiene sentido calcularla para cualquier tipo de variable.

Claramente cuando la variable es cuantitativa continua no tiene sentido hablar de valor más frecuente, si no, de **intervalo modal**.

En caso de que varias modalidades tengan la frecuencia máxima, se habla de **distribución multimoda** (con 2 o más modas).



Importante

Actividad

Las **medidas de dispersión** nos informan de hasta qué punto las medidas de centralización son representativas como síntesis de la información.

Las **medidas de dispersión** cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central.

Los más importantes son el **recorrido**, la **varianza**, la **desviación típica** y el **coeficiente de variación**.

- El **recorrido** de una variable es la diferencia entre el mayor y menor valor que toma esa variable. Se representa por **R** .
- La **varianza** mide la dispersión de una muestra en función de la diferencia (la distancia) de cada uno de los elementos de la muestra con el valor medio de la misma. Se representa por **s^2** .

Se calcula mediante la fórmula:

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

- La **desviación típica** es la raíz cuadrada positiva de la varianza. Se representa por **$\sigma = \sqrt{s^2}$** .

Aunque tanto la varianza como la desviación típica son medidas de la variación con respecto a la media y tienen un significado semejante, la desviación típica se

utiliza con más frecuencia porque se expresa en las mismas unidades que los valores de las variables. Al contrario que la varianza, cuyos valores se expresan en unidades al cuadrado.

- El **coeficiente de variación** es la medida de dispersión más popular y se define como el cociente entre la desviación típica y la media. Se representa por **CV**.

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

Cuanto mayor sea el coeficiente más dispersión hay en la variable y menos representativa es la media.



Importante

Los **percentiles** son medidas de posición que generalizan el concepto de mediana, y dividen el conjunto de observaciones en 100 partes de igual frecuencia.

Además de los percentiles, en estadística también se utilizan los **cuartiles** y los **deciles** (que dividen el conjunto de las observaciones en 4 y 10 partes de igual frecuencia, respectivamente).

Los percentiles se representan por q_p donde p es un número entre 0 y 0,99 con dos decimales (Por ejemplo, $q_{0,95}$ es el percentil 95).

Los cuartiles se representan por **Q_1 , Q_2 y Q_3** .

Aviso legal

Las páginas externas no se muestran en la versión imprimible

<http://www.juntadeandalucia.es/educacion/permanente/materiales/index.php?aviso#space>