Estadística Descriptiva en R: Parámetros y estadísticos

José Enrique Martín García
Universidad Politécnica de Gimialcón

Parámetros y Estadísticos

Parámetro: Es una cantidad numérica calculada sobre una población.

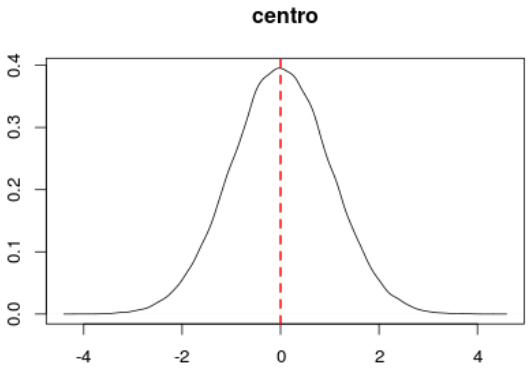
•La altura media de los individuos de un país.

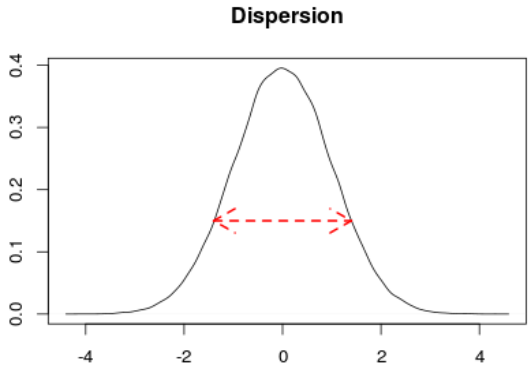
Estadístico: Es una cantidad numérica calculada sobre una muestra.

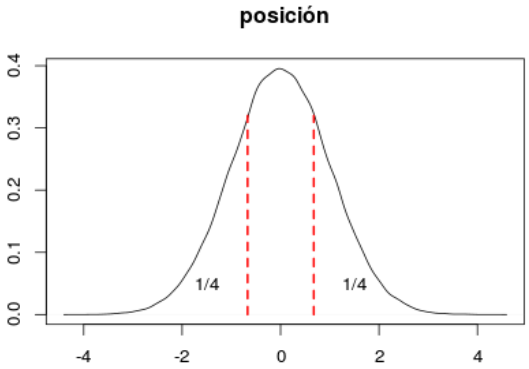
•La altura media de los que estamos en este aula.

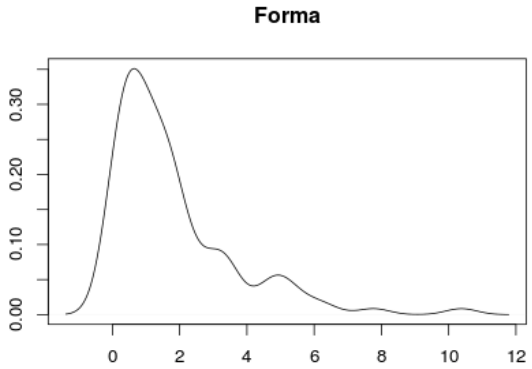
Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.
Los estadísticos se calculan, y estos estiman parámetros.

Parámetros y Estadísticos







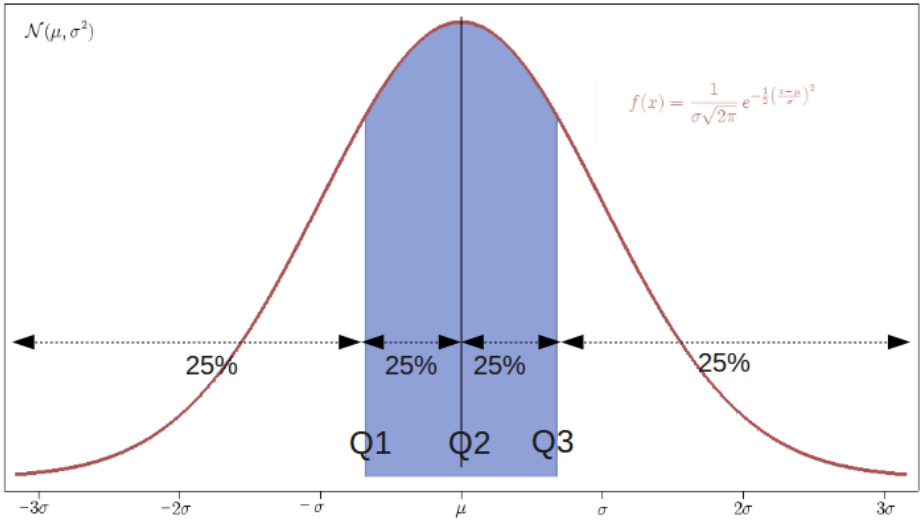


Medidas de posición

Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.

Las más populares:

- Cuantiles,
- Percentiles,
- Cuartiles,
- Deciles.



Medidas de posición: El Cuantil

El cuantil de orden p de una distribución (con $0 < p < 1$) es el valor de la variable xp que marca un corte de modo que una proporción p de valores de la población es menor o igual que xp.

Por ejemplo, el cuantil de orden 0,3 dejaría un 30% de valores por debajo y el cuantil de orden 0,50 se corresponde con la mediana de la distribución.

Los cuantiles suelen usarse por grupos que dividen la distribución en partes iguales; entendidas estas como intervalos que comprenden la misma proporción de valores.

Medidas de posición: El Cuantil

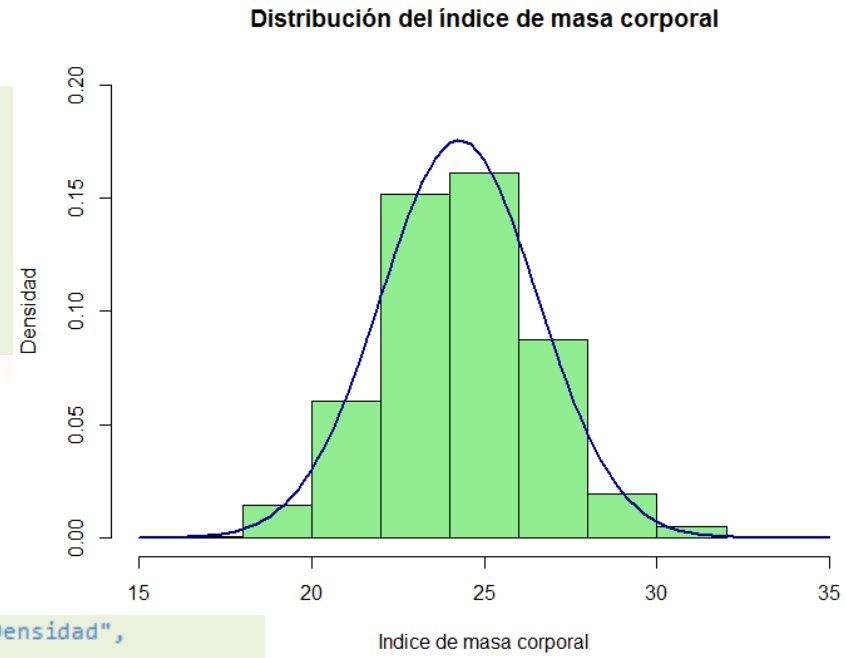
Los mas utilizados son:

- **Los cuartiles**, que dividen a la distribución en cuatro partes (corresponden a los cuantiles 0,25; 0,50 y 0,75);
- **Los quintiles**, que dividen a la distribución en cinco partes (corresponden a los cuantiles 0,20; 0,40; 0,60 y 0,80);
- **Los deciles,** que dividen a la distribución en diez partes;
- **Los percentiles**, que dividen a la distribución en cien partes.

Medidas de posición: El Cuantil

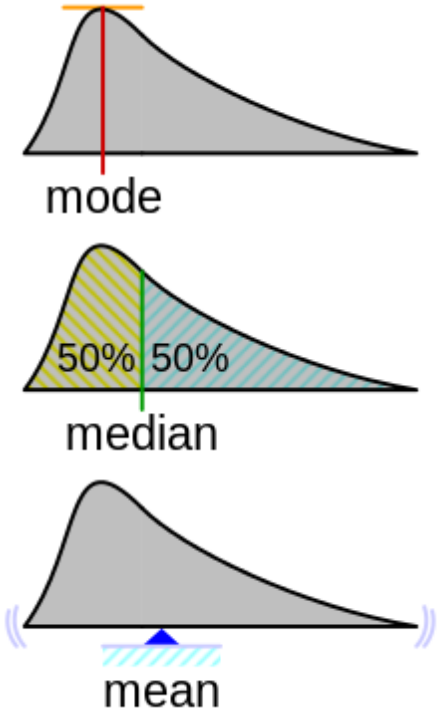
Ejemplo: El 5% de los españoles se consideran que tienen infrapeso con riesgo de anoréxia. ¿Qué Indice de Masa Corporal se considera “demasiado bajo” o infrapeso? Percentil 5 o cuantil 0,05

```
>BMI<-rnorm(n=1000, m=24.2, sd=2.2)
>quantile(BMI, 0.05)
 5%
20.41249
```




```
>hist(BMI, freq=FALSE, xlab="Indice de masa corporal", ylab="Densidad",
 + main="Distribución del índice de masa corporal", col="lightgreen",
 + xlim=c(15,35),  ylim=c(0, .20),breaks=10)
>curve(dnorm(x, mean=mean(BMI), sd=sd(BMI)), add=TRUE, col="darkblue", lwd=2)
```

Medidas de centralización o tendencia central



Indican valores con respecto a los que los datos “parecen” agruparse.

La media, moda y mediana son parámetros característicos de una distribución de probabilidad

La media se confunde a veces con la mediana o moda.; sin embargo, para las distribuciones con sesgo, la media no es necesariamente el mismo valor que la mediana o que la moda

Medidas de centralización o tendencia central

Media aritmética

La media aritmética es el promedio de un conjunto de valores, o su distribución que a menudo se denomina "promedio".

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

La media aritmética "mean" es la suma de los valores dividido por el tamaño muestral.

```
>x  <- c(1,  2,  3,  4,5  )
>mean(x)
3
>media <-  (1  +  2  +3  +  4  +  5)/5
>media
3
```

La media aritmética se trata de un parámetro conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos (en estos casos hay otras 'medias', menos intuitivas, pero que pueden ser útiles: media aritmética, geométrica, ponderada...)

Medidas de centralización o tendencia central

Mediana

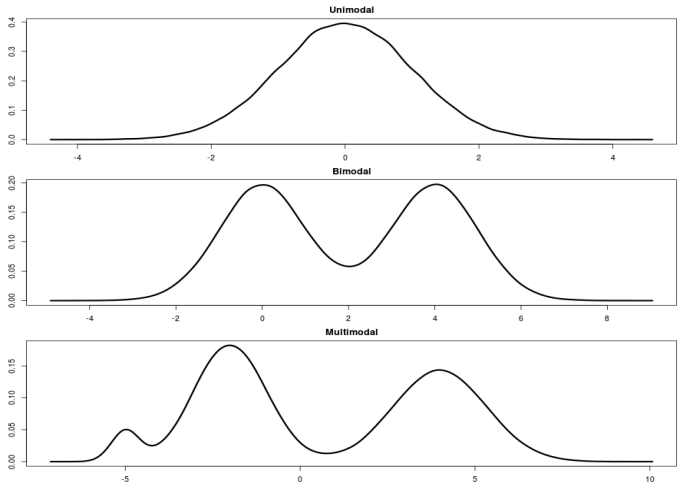
Representa el valor de la variable de posición central en un conjunto de datos ordenados.
Mediana(“median”): Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
Mediana de 1,2,4,5,6,6,8 es 5.
Mediana de 1,2,4,5,6,6,8,9 es 5.5
Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
Mediana de 1,2,4,5,6,6, 800 es 5. (¡La media es 117,7!)

```
> y <- c(1, 2, 4, 5, 6, 6, 8, 9)
> z <- c(1, 2, 4, 5, 6, 6, 800)
> median(y)
[1] 5.5
> median(z)
[1] 5
```

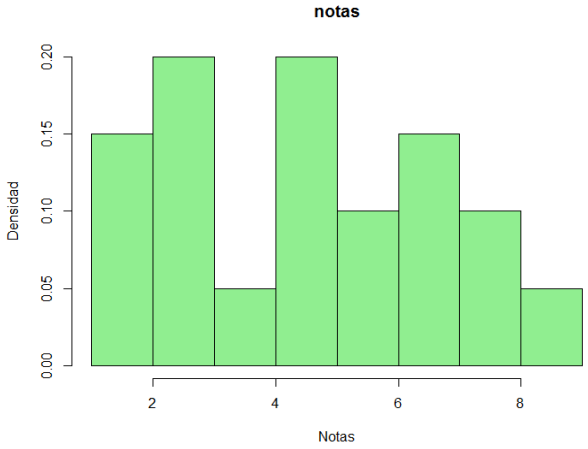
Medidas de centralización o tendencia central

Moda

La moda es el valor con una mayor frecuencia en una distribución de datos. Se hablará de una distribución bimodal de los datos cuando encontremos dos modas, es decir, dos datos que tengan la misma frecuencia absoluta máxima.



Para calcular la moda de una distribución tenemos que utilizar un paquete específico. En este caso utilizamos el paquete "modeest"




```
> library(modeest)
> notas<-c(3,3,5,2,7,8,5,7,6,4,5,6,5,3,9,7,8,3,1,1)
> mfv(notas)
[1] 3 5
> hist(notas, freq=FALSE, xlab="Notas", ylab="Densidad",
+ main="notas", col="lightgreen",breaks=11)
```

Medidas de dispersión

Rango

Rango de una variable estadística es la diferencia entre el mayor y el menor valor que toma la misma. Es la medida de dispersión más sencilla de calcular, Basta con que uno de estos dos datos varíe para que el parámetro también lo haga, aunque el resto de la distribución siga siendo, esencialmente, la misma.

Existen otros parámetros dentro de esta categoría, como los **recorridos o rangos intercuantílicos,** que tienen en cuenta más datos y, por tanto, permiten afinar en la dispersión.

Entre los más usados está el **rango intercuartílico**, que se define como la diferencia entre el cuartil tercero y el cuartil primero. En ese rango están, por la propia definición de los cuartiles, el 50% de las observaciones.

Este tipo de medidas también se usa para determinar valores atípicos.

Medidas de dispersión

Desviaciones Medias

Dada una variable estadística X y un parámetro de tendencia central, c, se llama desviación de un valor de la variable, xi, respecto de c, al número $|x_i - c|$. Este número mide lo lejos que está cada dato del valor central c, por lo que una media de esas medidas podría resumir el conjunto de desviaciones de todos los datos.

Así pues, se denomina desviación media de la variable X respecto de c a la media aritmética de las desviaciones de los valores de la variable respecto de c, esto es, si

$$X = x_1, x_2, \ldots, x_n, \text{ entonces } DM_c = \frac{\sum_{i=1}^{n} |x_i - c|}{n}$$

De este modo se definen la desviación media respecto de la media ($c = \overline{x}$) o la desviación media respecto de la mediana ($c = \overline{Me}$), cuya interpretación es sencilla en virtud del significado de la media aritmética

Sin embargo, el uso de valores absolutos impide determinados cálculos algebraicos que obligan a desechar estos parámetros,

Medidas de dispersión

Varianza y desviación típica

La suma de todas las desviaciones respecto al parámetro más utilizado, la media aritmética, es cero. Por tanto si se desea una medida de la dispersión sin los inconvenientes para el cálculo que tienen las desviaciones medias, una solución es elevar al cuadrado tales desviaciones antes de calcular el promedio.

Se define la **varianza** como:

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

o sea, la media de los cuadrados de las desviaciones respecto de la media.

La desviación típica, σ, se define como la raíz cuadrada de la varianza, esto es,

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

Medidas de dispersión

Coeficiente de variación de Pearson

Es la razón entre la desviación típica y la media. Se define como:

$$C_V = \frac{\sigma}{\bar{x}}$$

Donde σ es la desviación típica y x es la media aritmética.

Se interpreta como el número de veces que la media está contenida en la desviación típica. Suele darse su valor en tanto por ciento, multiplicando el resultado anterior por 100. De este modo se obtiene un porcentaje de la variabilidad.

Mide la desviación típica en forma de “qué tamaño tiene con respecto a la media”.

También se la denomina variabilidad relativa. Es una cantidad adimensional. Interesante para comparar la variabilidad de diferentes variables.

Medidas de dispersión

Medidas de dispersión en R

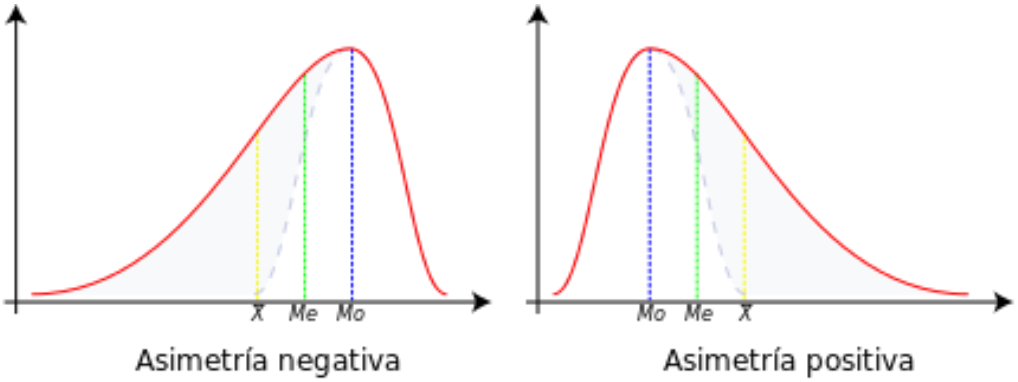
```
> pesos <- rnorm(1000,  3,  0.8)
> range(pesos)
[1] 0.4445757 5.5031138
> IQR  <- (quantile(pesos,  0.75,  names  =  F) - quantile(pesos,
+ 0.25,  names  =  F))
> IQR
[1] 1.088238
> var(pesos)
[1] 0.6307161
> sd(pesos)
[1] 0.7941764
> CV <- sd(pesos)/mean(pesos)
> CV
[1] 0.2634968
```

Medidas de forma

Medidas de asimetría

Se dice que una distribución de datos estadísticos es simétrica cuando la línea vertical que pasa por su media, divide a su representación gráfica en dos partes simétricas. Ello equivale a decir que los valores equidistantes de la media, a uno u otro lado, presentan la misma frecuencia.

En las distribuciones simétricas los parámetros media, mediana y moda coinciden, mientras que si una distribución presenta cierta asimetría, de un tipo o de otro, los parámetros se sitúan como muestra el siguiente gráfico:



Medidas de forma

Medidas de asimetría

La asimetría resulta útil en muchos campos. Muchos modelos simplistas asumen una distribución normal, esto es, simétrica en torno a la media. La distribución normal tiene una asimetría cero. Pero en realidad, los valores no son nunca perfectamente simétricos y la asimetría de la distribución proporciona una idea sobre si las desviaciones de la media son positivas o negativas. Una asimetría positiva implica que hay más valores distintos a la derecha de la media.

Las medidas de asimetría, sobre todo el coeficiente de asimetría de Fisher, junto con las medidas de apuntamiento o curtosis se utilizan para contrastar si se puede aceptar que una distribución estadística sigue la distribución normal. Esto es necesario para realizar numerosos contrastes estadísticos en la teoría de inferencia estadística

Medidas de forma

Medidas de asimetría

Coeficiente de asimetría de Pearson

Sólo se puede utilizar en distribuciones uniformes, unimodales y moderadamente asimétricas. Se basa en que en distribuciones simétricas la media de la distribución es igual a la moda.

$$A_p = \frac{\mu - moda}{\sigma},$$

Donde μ es el momento central de orden 1, que corresponde a la media aritmética de la variable X.

Si la distribución es simétrica, μ=moda y Ap = 0. Si la distribución es asimétrica positiva la media se sitúa por encima de la moda y, por tanto, Ap > 0.

Medidas de forma

Medidas de asimetría

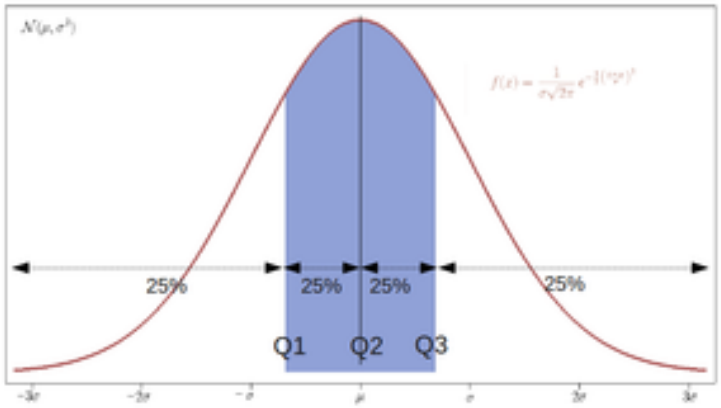
Coeficiente de asimetría de Bowley

Está basado en la posición de los cuártiles y la mediana, y utiliza la siguiente expresión:

$$A_B = \frac{Q_{3/4} + Q_{1/4} - 2Me}{Q_{3/4} - Q_{1/4}}$$

En una distribución simétrica el cuartil estará a la misma distancia de la mediana que el primer cuartil. Por tanto AB=0.

Si la distribución es positiva o a la derecha, AB > 0



Medidas de forma

Medidas de apuntamiento o curtosis

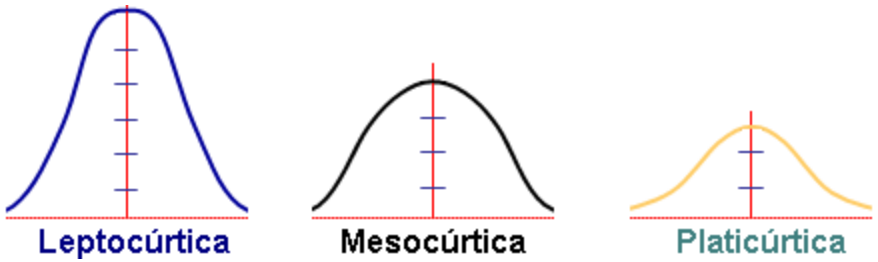
Con estos parámetros se pretende medir cómo se reparten las frecuencias relativas de los datos entre el centro y los extremos, tomando como comparación la campana de Gauss.
El parámetro usado con más frecuencia para esta medida es el **coeficiente de curtosis de Fisher**, definido como:

,

$$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3$$

La comparación con la distribución normal permite hablar de distribuciones

- Platicúrtica (aplanada): curtosis < 0
- Mesocúrtica (como la normal): curtosis = 0
- Leptocúrtica (apuntada): curtosis > 0



Otros parámetros

Proporción

La proporción de un dato estadístico es el número de veces que se presenta ese dato respecto al total de datos. Se conoce también como frecuencia relativa y es uno de los parámetros de cálculo más sencillo. Tiene la ventaja de que puede calcularse para variables cualitativas.
El dato con mayor proporción se conoce como **moda**.

Número índice

Un número índice es una medida estadística que permite estudiar las fluctuaciones o variaciones de una magnitud o de más de una en relación al tiempo o al espacio. Algunos ejemplos de uso cotidiano de este parámetro son el índice de precios o el IPC.

Tasa

La tasa es un coeficiente que expresa la relación entre la cantidad y la frecuencia de un fenómeno o un grupo de fenómenos. Se utiliza para indicar la presencia de una situación que no puede ser medida en forma directa. Esta razón se utiliza en ámbitos variados, como la demografía o la economía, donde se hace referencia a la tasa de interés.