Notación: Data Frames

$X_{n \times p} = $ DATA FRAME

	edad	Voto	Ideo
1	19	P	CC
2	25	C	CD
⋮	⋮	⋮	⋮
n	31	P	CI

x_{ij} Indice

$j \to$ columna

$i \to$ fila

ej 1: $x_{11} = \{19\}$

$x_{23} = \{CD\}$

$x_{np} = \{CI\}$

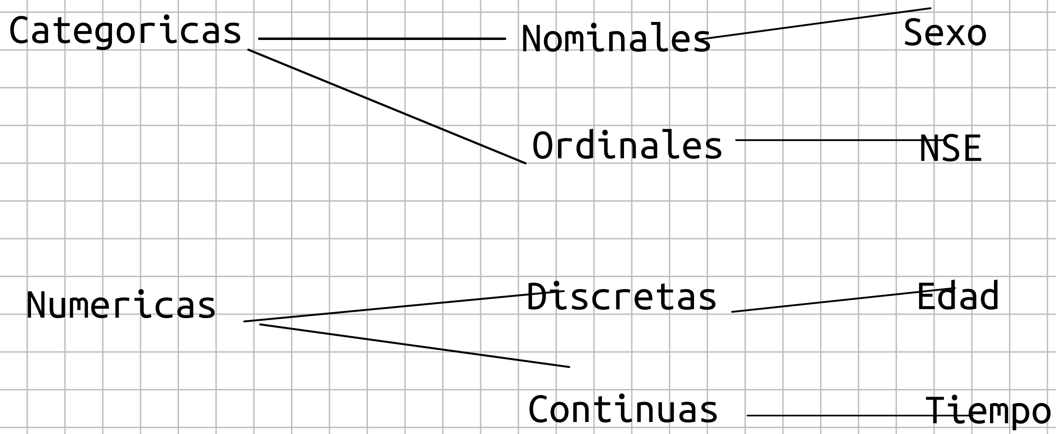
⋮ ⇒ Sigue para abajo

m-columnas

p-variables

n-casos

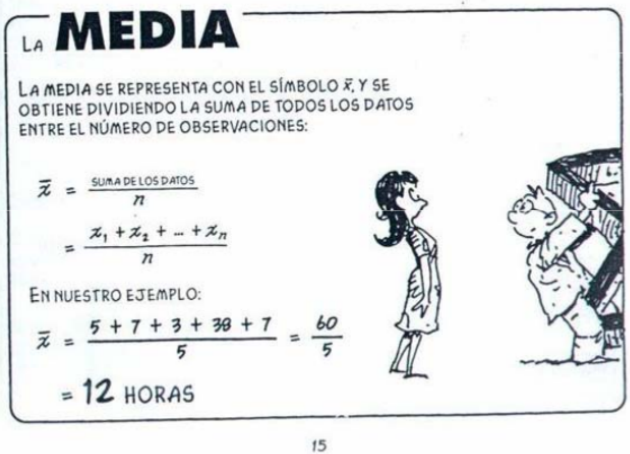
Tipos de Variables



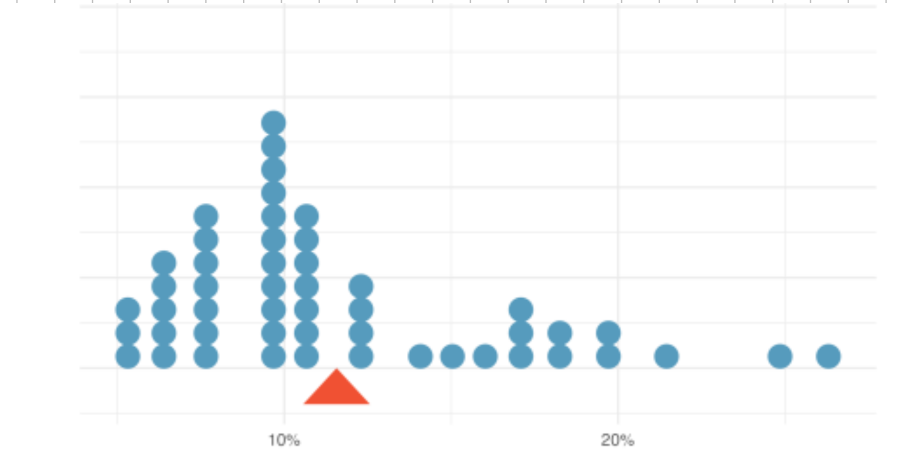
Resumen Numérico

Media (centro de masa) (medida de posición)

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{n-1} + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$



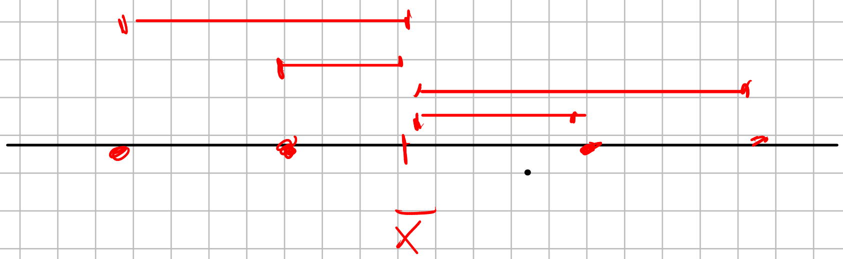
La media, como resumen numérico, explica el centro de masa.
Por tanto, es sencible a la contaminación de los datos



Varianza (distancia cuadrática a la media)
(medida de disperción)

$$s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n}$$

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



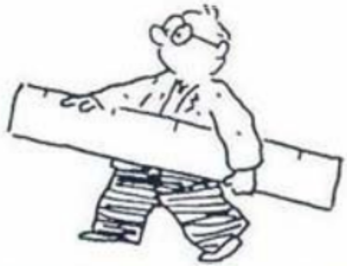
Promedio de
distancias
cuadráticas
de los casos
a la media

LA MEDIDA ESTÁNDAR DE LA DISPERSIÓN ES LA

DESVIACIÓN TÍPICA (TAMBIÉN DESVIACIÓN ESTÁNDAR)

A DIFERENCIA DEL IQR, QUE SE CALCULA A PARTIR DE LAS MEDIANAS, LA DESVIACIÓN TÍPICA MIDE LA DISPERSIÓN DE LOS DATOS DESDE LA MEDIA. UNA FORMA INTUITIVA DE VERLA ES COMO LA DISTANCIA MEDIA ENTRE LOS DATOS Y LA MEDIA $\bar{x}$...

Ver en siguient. diapos

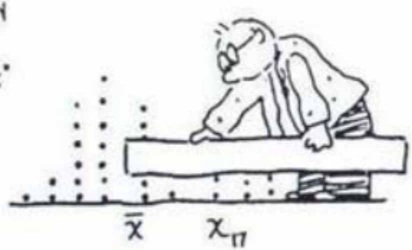


SIN EMBARGO, EN ESTA OCASIÓN UTILIZAMOS LAS DISTANCIAS ELEVADAS AL CUADRADO. O SEA, SI LA DISTANCIA AL CUADRADO ENTRE EL PUNTO x_i Y $\bar{x}$ ES $(x_i - \bar{x})^2$, ENTONCES

$$\text{LA DISTANCIA CUADRÁTICA MEDIA} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

POR MOTIVOS TÉCNICOS, SE UTILIZA $n-1$ EN EL DENOMINADOR EN LUGAR DE n, Y DEFINIMOS ENTONCES LA VARIANZA MUESTRAL s^2:*

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



* TAMBIÉN ES CORRECTA VARIANZA. [N.T.]

Ejemplo para una muesta de 50 casos

$$\begin{aligned}
x_1 - \bar{x} &= 10.9 - 11.57 = -0.67 \\
x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\
x_3 - \bar{x} &= 26.3 - 11.57 = 14.73 \\
&\vdots \\
x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49
\end{aligned}$$

$$\begin{aligned}
s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \cdots + (-5.49)^2}{50 - 1} \\
&= \frac{0.45 + 2.72 + \cdots + 30.14}{49} \\
&= 25.52
\end{aligned}$$

Desvío Estandar

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

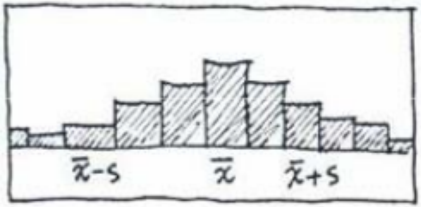
Al aplicar la raiz cuadrada, nos queda el desvío en las misma unidad que nuestros datos (piense en $^2 vs $)

En el ejemplo anterior, se obtendría

s = sqrt(25.52) = 5.051732

Puntaje Z (centrar y escalar los datos)

Las propiedades de $\bar{X}$ y S



LA MEDIA Y LA DESVIACIÓN TÍPICA SON MUY ÚTILES PARA RESUMIR LAS PROPIEDADES DE HISTOGRAMAS BASTANTE SIMÉTRICOS, SIN OBSERVACIONES ATÍPICAS, O SEA, HISTOGRAMAS CON FORMA DE MONTAÑA.

A MENUDO RESULTA ÚTIL SABER CUÁNTAS DESVIACIONES TÍPICAS DISTA UN PUNTO DE LA MEDIA. ENTONCES DEFINIMOS z, O VALORES ESTANDARIZADOS, COMO LA DISTANCIA DESDE $\bar{x}$ POR DESVIACIÓN TÍPICA.

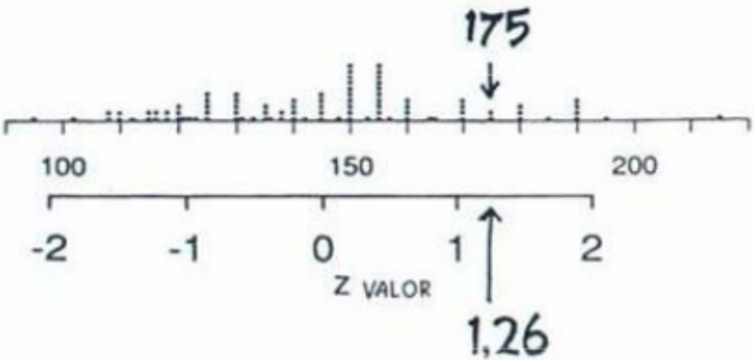
$$z_i = \frac{x_i - \bar{x}}{s} \quad \text{PARA CADA } i.$$

Puntaje Z (centrar y escalar los datos)

A MENUDO RESULTA ÚTIL SABER CUÁNTAS DESVIACIONES TÍPICAS DISTA UN PUNTO DE LA MEDIA. ENTONCES DEFINIMOS z, O VALORES ESTANDARIZADOS, COMO LA DISTANCIA DESDE $\bar{x}$ POR DESVIACIÓN TÍPICA.

$$z_i = \frac{x_i - \bar{x}}{s}$$ PARA CADA i.

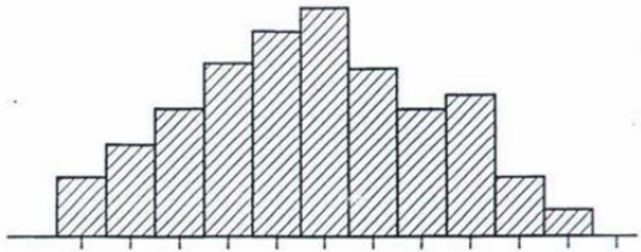
UNA z DE +2 QUIERE DECIR QUE LA OBSERVACIÓN SE ENCUENTRA DOS DESVIACIONES TÍPICAS POR ENCIMA DE LA MEDIA. EN LOS DATOS DE LOS PESOS DE LOS ESTUDIANTES ($\bar{x} = 145{,}2$ Y $S = 23{,}7$), PODEMOS REPRESENTAR LOS DATOS SIMULTÁNEAMENTE EN EL EJE x DEL PRINCIPIO Y UN EJE z.



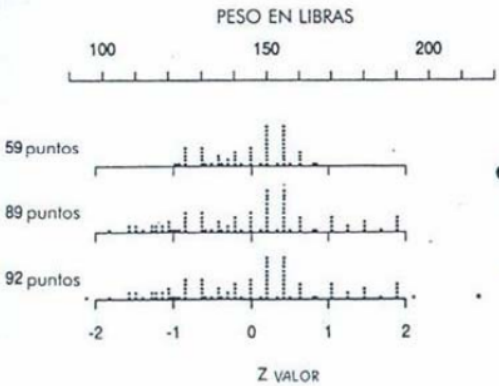
UN ESTUDIANTE QUE PESE 175 LIBRAS TIENE UNA z DE $\frac{175-145.2}{23.7} = 1{,}26$

una REGLA EMPÍRICA:

EN LOS CONJUNTOS DE DATOS CASI SIMÉTRICOS, CON FORMA DE MONTAÑA, ALREDEDOR DE UN **68%** DE LOS DATOS SE ENCUENTRA A MENOS DE UNA DESVIACIÓN TÍPICA DE LA MEDIA, Y EL **95%** ESTÁ A MENOS DE DOS DESVIACIONES TÍPICAS DE LA MEDIA.



SI MIRAMOS LOS PESOS, ESTA REGLA EMPÍRICA FUNCIONA BASTANTE BIEN: UN **64%** (= 59/92) DE LOS PESOS ESTÁN A MENOS DE UNA DESVIACIÓN TÍPICA DE LA MEDIA, Y UN **97%** (= 89/92) ESTÁN A MENOS DE DOS DESVIACIONES TÍPICAS DE LA MEDIA.



Y AHORA TOCA DESCANSAR DE TANTO NÚMERO.

Estadísticos de Orden

Notación y ejemplos [editar]

Por ejemplo, supongamos que se observan o son registrados 4 números, lo que resulta en una muestra de tamaño 4. Si los valores de la muestra son

6, 9, 3, 8,

que por lo general se denominan

$$x_1 = 6, \;\; x_2 = 9, \;\; x_3 = 3, \;\; x_4 = 8,$$

donde el subíndice *i* in x_i simplemente indica el orden en el que se registraron las observaciones y se supone por lo general no son significativos. Un caso en el que el orden es significativo es cuando las observaciones son parte de una serie de tiempo.

Los estadísticas de orden se indican

$$x_{(1)} = 3, \;\; x_{(2)} = 6, \;\; x_{(3)} = 8, \;\; x_{(4)} = 9,$$

donde el subíndice (i) entre paréntesis indica el orden º del estadística de la muestra *i*.

Ejemplo:

El primer estadístico de orden (o estadístico de orden más pequeño) es siempre el mínimo de la muestra, es decir,

$$X_{(1)} = \min\{\, X_1, \ldots, X_n \,\}$$

donde, tras una convención común, utilizamos letras mayúsculas para referirnos a variables aleatorias, y las letras minúsculas (como arriba) para los valores reales observados.

Del mismo modo, para una muestra de tamaño n, el n-ésimo estadístico de orden n (o más grande estadístico de orden) es el máximo, es decir:

$$X_{(n)} = \max\{\, X_1, \ldots, X_n \,\}.$$

El rango de la muestra es la diferencia entre el máximo y el mínimo. Note que es una función de los estadísticos de orden:

$$\mathrm{Range}\{\, X_1, \ldots, X_n \,\} = X_{(n)} - X_{(1)}.$$

Cuantiles

Una forma de obtener estadísticas resistentes es utilizar los cuantiles empíricos (percentiles / fractiles).

El cuantil (este término fue utilizado por primera vez por Kendall, 1940) de una distribución es el número x_p tal que una proporción p de los valores sea menor o igual que x_p. Por ejemplo, el cuantil 0.25 ($x_{0.25}$) es el valor tal que el 25% de todos los valores caen por debajo de ese valor.

Los cuantiles empíricos se pueden construir más fácilmente ordenando los datos en orden ascendente para obtener una secuencia de estadísticas de orden :

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

El p_ésimo cuantil ($Q_p(x)$) se obtiene tomando el rango $r = (n+1) * p$ del enésimo estadístico de orden

$$Q_p(X) = \begin{cases} x_{((n+1)*p)} & \text{si } (n+1)*p \text{ es n° entero} \\ 0.5*\left(x_{(\lfloor (n+1)*p \rfloor)} + x_{(\lfloor (n+1)*p+1 \rfloor)}\right) & \text{si no} \end{cases}$$

Notacion:

$\lfloor r \rfloor$

quiere decir redondeo para abajo. Ej: 3.3 quedaría 3

Ejemplo con la Imágen de un Candidato (de 0 a 100):

X={14, 60, 20, 19, 30, 12, 18, 13, 43, 19, 30}

Ordeno a X

X_ordenado = {12, 13, 14, 18, 19, 19, 20, 30, 30, 43, 60 }

Como #X = 11, tengo que

Q_.50(X) = x_(r=6) = 19   ya que r = (11+1)*0.5 = 6

Q_.85(X) = x_(r=10:11) = 0.5* (x_(10) + x(11)) = (43 + 60) / 2

ya que (11+1)*0.85 = 10.2
por lo que busco a r=10 y a r=11

Q_.25 = x_(r=3) = 14

Q_.75 = x_(r=9) = 30

OJO: Hay numerosas definiciones de Quantiles (problema de estimación). Ver:
https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile

Mediana (Q_.5(X))

PARA ENCONTRAR LA MEDIANA DE UN CONJUNTO DE DATOS, ORDENAMOS LOS DATOS DE MENOR A MAYOR. LA MEDIANA ES EL VALOR QUE QUEDA EN EL CENTRO.

3 5 7 7 38

MEDIANA

SI EL NÚMERO DE OBSERVACIONES ES PAR, EN CUYO CASO NO HAY NINGÚN PUNTO CENTRAL, HACEMOS LA MEDIA DE LOS DOS VALORES QUE QUEDAN EN EL CENTRO. ASÍ QUE SI LOS DATOS SON

3 5 7 7

ESPACIO CENTRAL

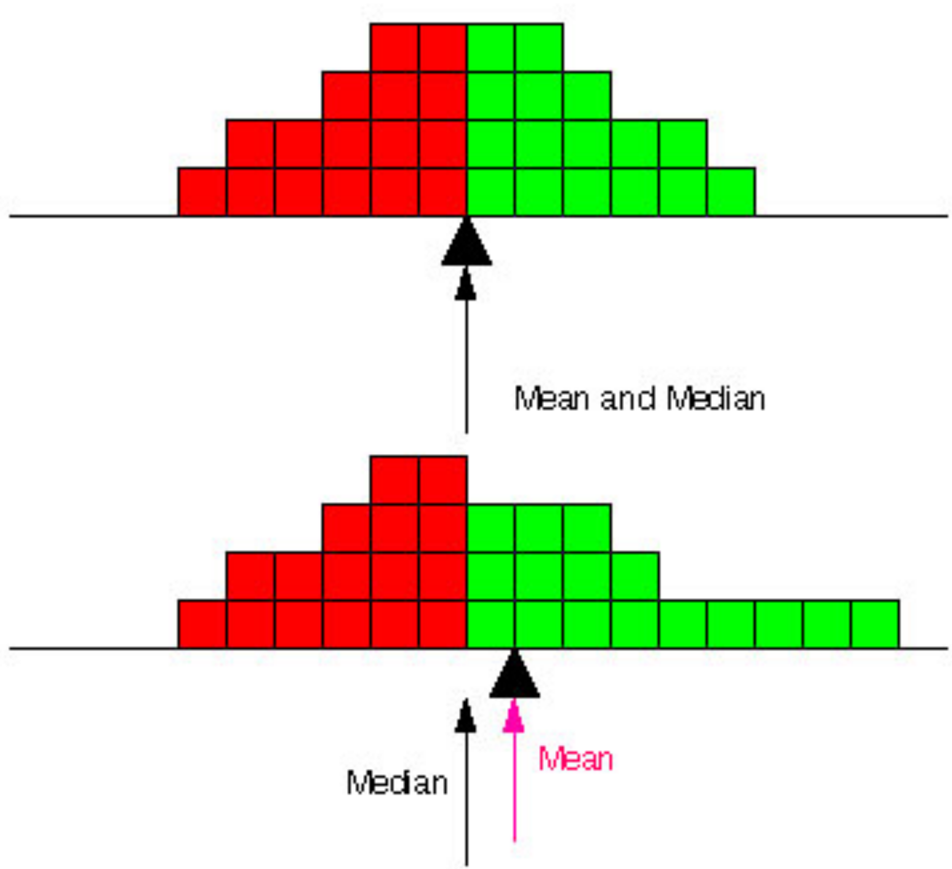
HACEMOS LA MEDIA DE 5 Y 7: $\frac{5+7}{2} = 6$

ESTO NOS DA UNA REGLA GENERAL: ORDENAR LOS DATOS DE MENOR A MAYOR.

SI EL NÚMERO DE DATOS ES IMPAR, LA MEDIANA ES EL VALOR CENTRAL.

SI EL NÚMERO DE DATOS ES PAR, LA MEDIANA ES LA MEDIA DE LOS DOS DATOS MÁS CERCANOS AL CENTRO.

Media vs Mediana: Caso 1 - Asimetría



En clase 2

se explicará

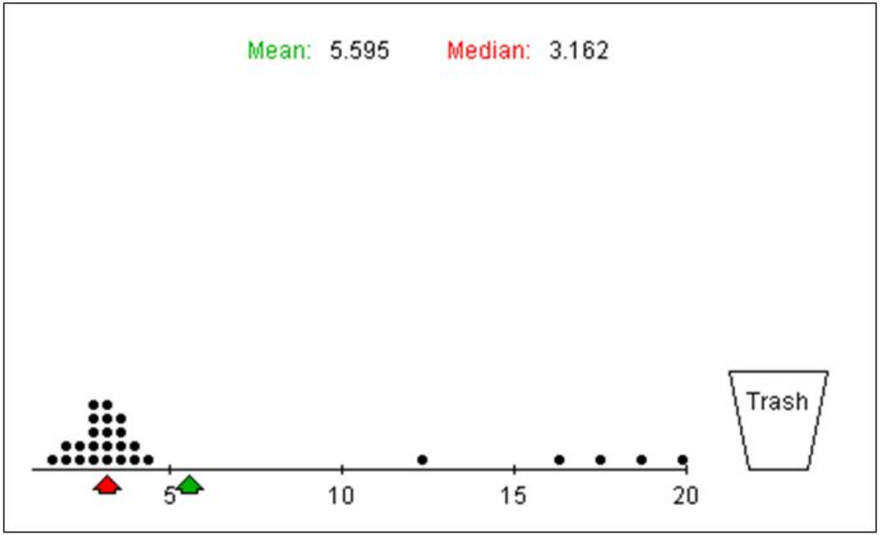
los histográmas

con mayor

detalle

Media vs Mediana: Caso 2 - Contaminación (outliers / valores atípicos)

Mean vs. Median

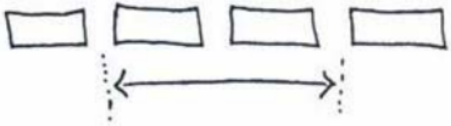


http://www.stat.tamu.edu/~west/ph/meanmedian.html

TAMBIÉN HAY VARIAS FORMAS DE MEDIR LA DISPERSIÓN. UNA ES EL

RECORRIDO INTERCUARTÍLICO

SE TRATA DE DIVIDIR LOS DATOS EN CUATRO GRUPOS IGUALES Y OBSERVAR LA DISTANCIA QUE SEPARA LOS GRUPOS EXTREMOS.



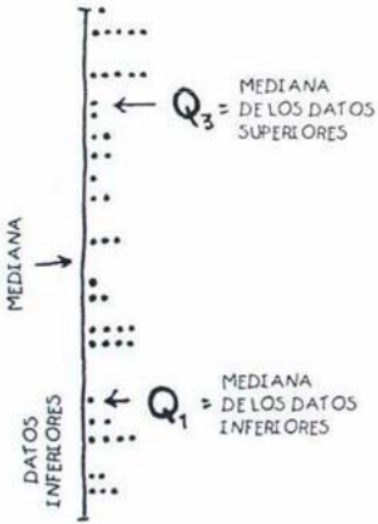
ÉSTA ES LA RECETA:

1) ORDENA LOS DATOS NUMÉRICAMENTE.

2) DIVIDE LOS DATOS POR LA MEDIANA EN DOS GRUPOS IGUALES (SI LA MEDIANA COINCIDE CON UN DATO, INCLÚYELO EN LOS DOS GRUPOS).

3) CALCULA LA MEDIANA DEL GRUPO INFERIOR. ÉSE ES EL PRIMER CUARTIL, O Q_1.

4) LA MEDIANA DEL GRUPO SUPERIOR ES EL TERCER CUARTIL, O Q_3.



IQR =

Q.75(X) - Q_.25(X)

Estimador Robusto
de desvío muestral
s(x)

Ej con imagen:

IQR = 30 - 14 = 16