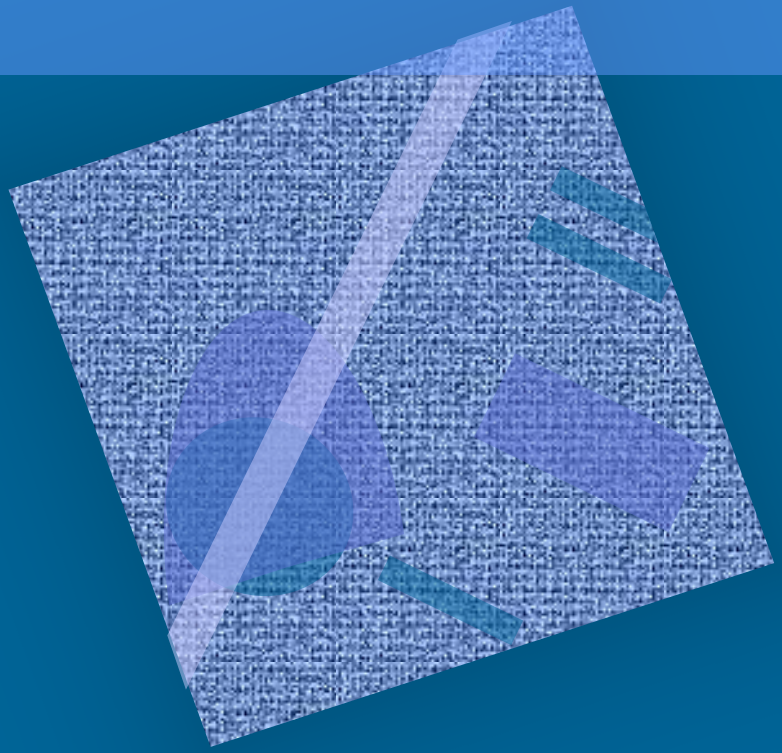
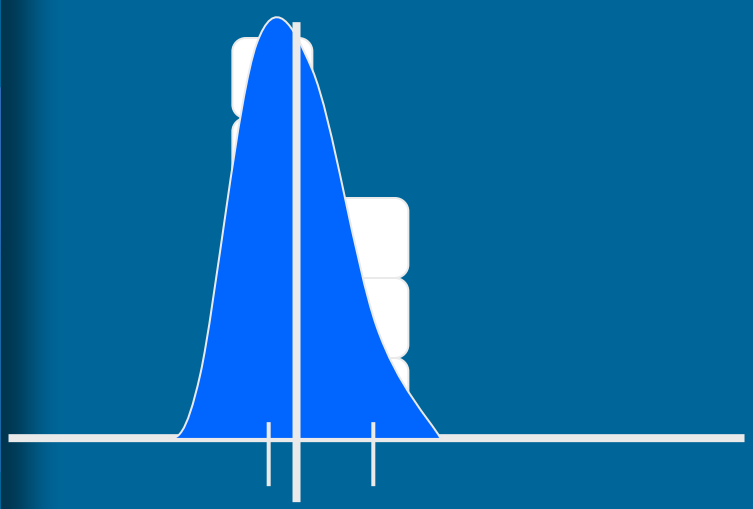
Medidas de dispersión



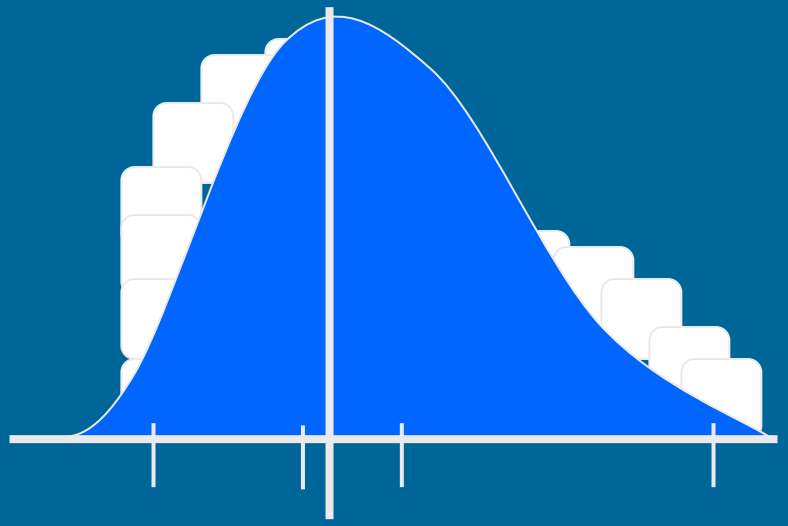
- En el análisis estadístico no basta el cálculo e interpretación de las medidas de tendencia central o de posición, ya que, por ejemplo, cuando pretendemos representar toda una información con la media aritmética, no estamos siendo absolutamente fieles a la realidad, pues suelen existir datos extremos inferiores y superiores a la media aritmética.

Medidas de dispersión

Ejemplo de dos conjuntos de datos con igual media



Datos con baja dispersión



Datos con alta dispersión

Amplitud o Rango

- **Rango**

Una manera de medir la dispersión es calcular el ***recorrido*** de la distribución empírica, es decir, la diferencia entre las observaciones máxima y mínima.

Su mayor ventaja es que se puede calcular facilmente, sin embargo, no brinda información sobre la dispersión existente entre ambos valores extremos.

- El rango depende sólo de las observaciones máxima y mínima, que podrían ser observaciones atípicas.
- Podríamos mejorar nuestra descripción de la dispersión fijándonos, por ejemplo, también en la dispersión del 50% de los valores centrales de nuestros datos.
- Un conjunto de estadísticos de utilidad son los **cuartiles** de una distribución.

- Esto se calcula por la diferencia entre el dato mayor (H) y el dato menor (L).

Ejemplo :

muestra:

4, 4, 5, 7, 8, 9

Solución:

- dato mayor H = 9
- dato menor L = 4
- A = 9 — 4 = 5

- La amplitud señala que los 6 datos se encuentran dentro de una distancia de 5 unidades en la recta numérica.

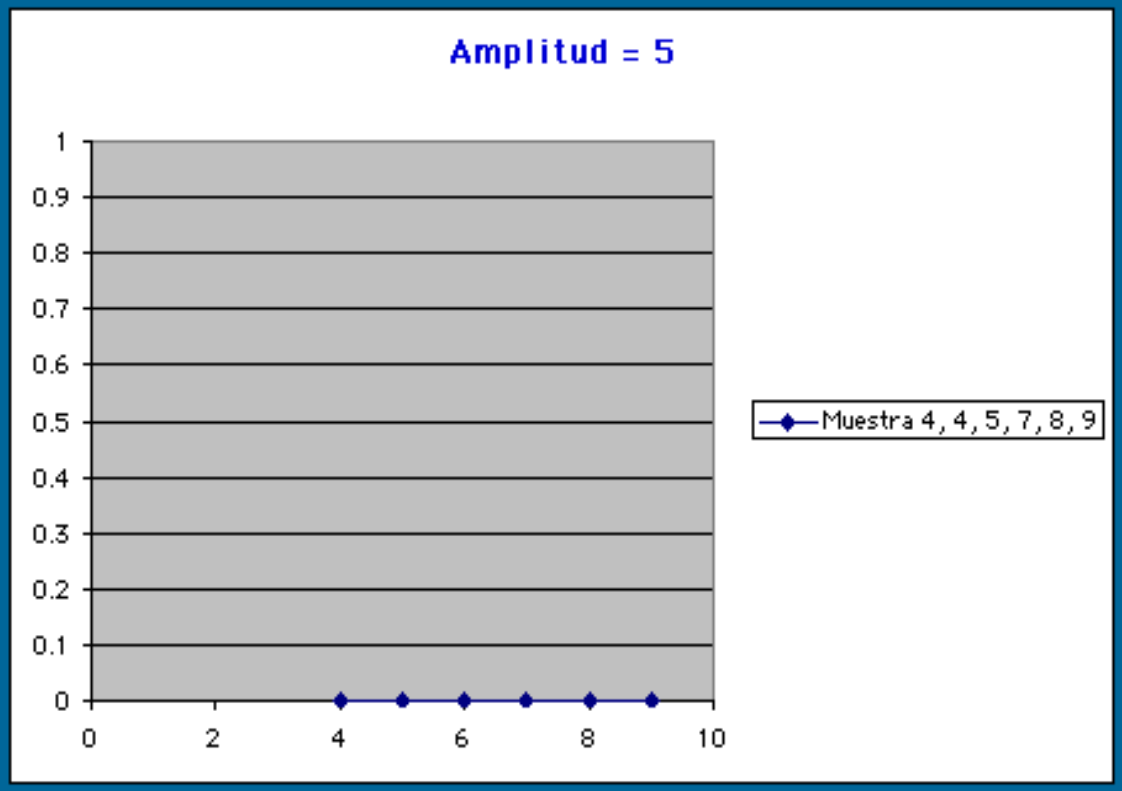


Diagrama de caja

- Los cinco números resumen de una distribución son representados gráficamente por un ***diagrama de caja***.

  - L - Observación máxima
  - Q_3 - Tercer cuartil
  - Q_2 - Mediana
  - Q_1 - Primer cuartil
  - S - Observación mínima

Diagrama de caja

- Los lados inferior y superior de la caja van del primer al tercer cuartil. Por tanto, la altura de la caja es la amplitud del 50% de los datos centrales.
- El segmento del interior de la caja indica la mediana. Los extremos de los segmentos perpendiculares a los lados superior e inferior indican, respectivamente, los valores máximo y mínimo de la distribución.

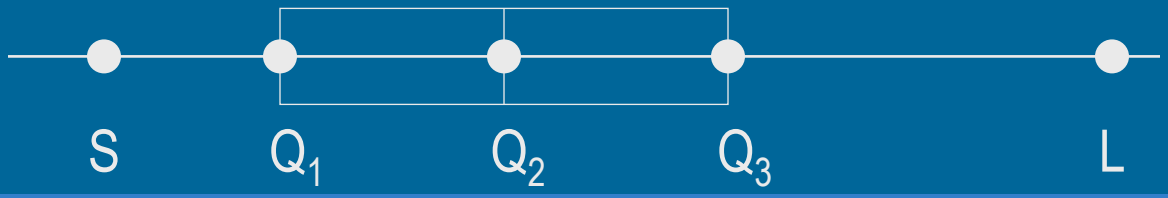
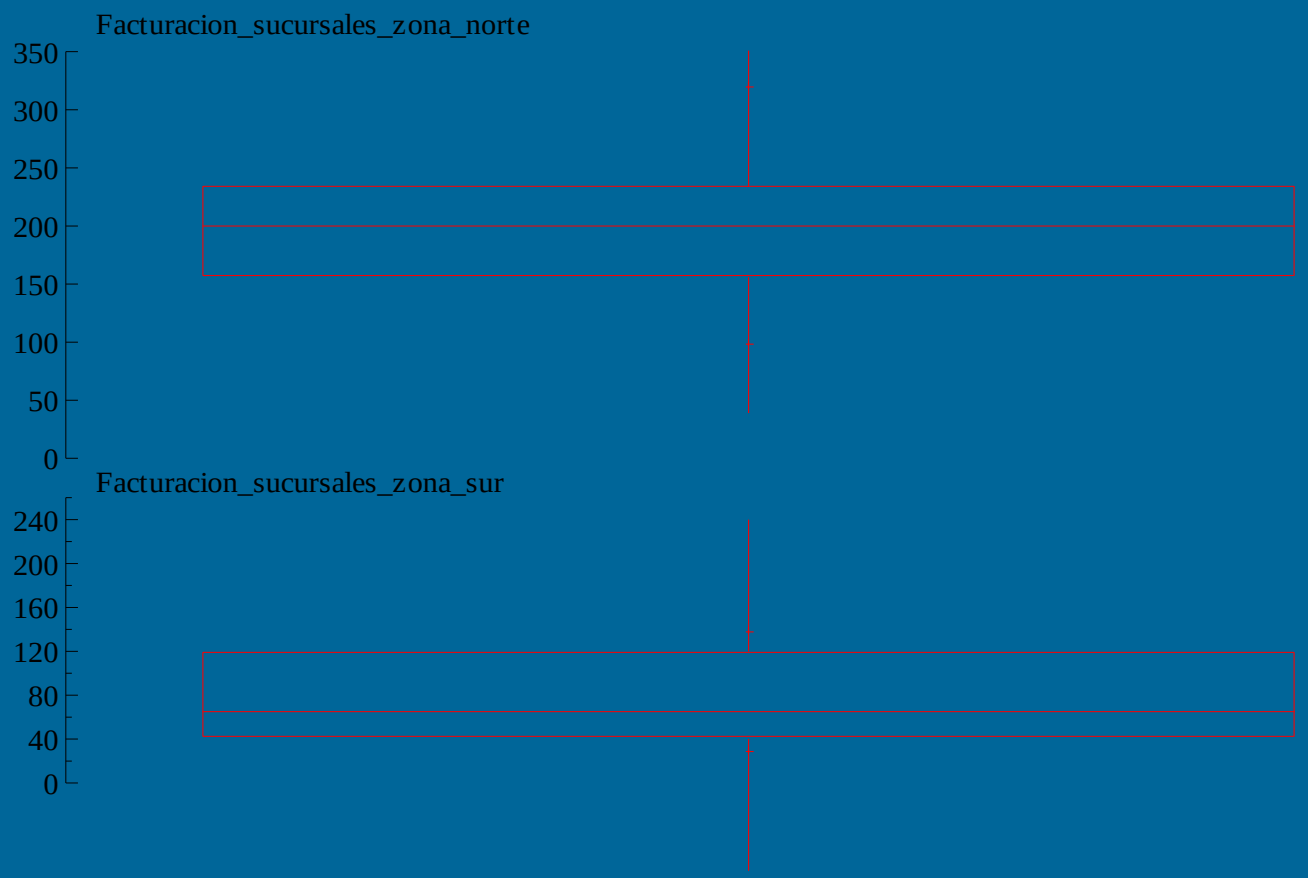


Diagrama de caja



Varianza

- La **varianza**, **S^2**, se define como la media de las diferencias cuadráticas de *n* puntuaciones con respecto a su media aritmética, es decir:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

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

- Para datos agrupados en tablas, usando las notaciones establecidas anteriormente, la varianza se puede escribir como

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \, \mathrm{n_i}}{n-1}$$

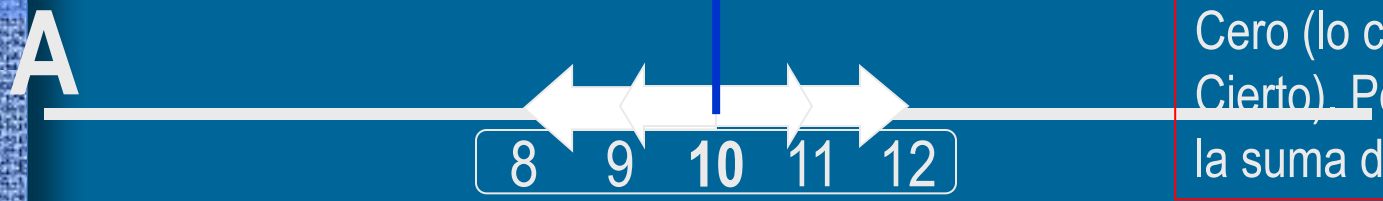
La varianza

Considere dos poblaciones:
Población A: 8, 9, 10, 11, 12
Población B: 4, 7, 10, 13, 16

9 –10 = -1
11 –10 = +1
8 –10 = -2
12 – 10 = +2
Suma = 0

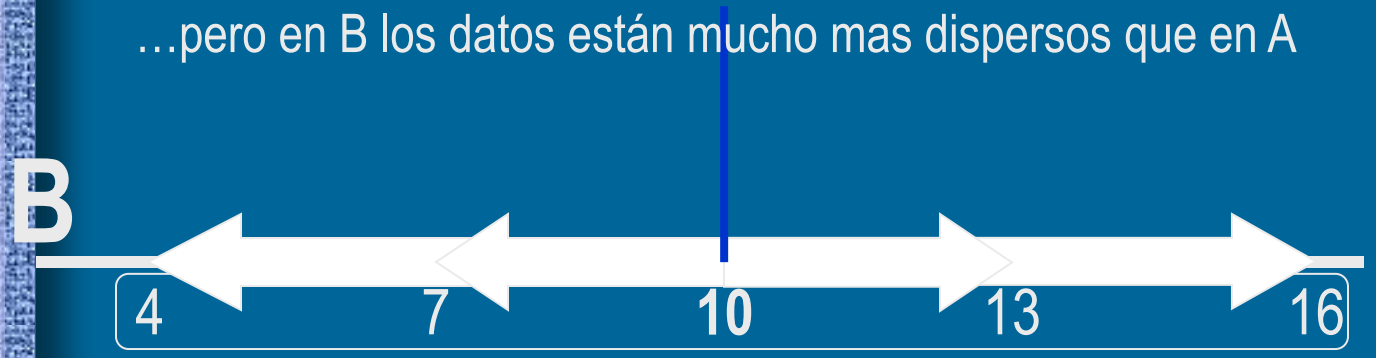
Comencemos calculando la suma de las desviaciones

La media de ambas poblaciones es 10...



En ambos casos, la suma de las desviaciones es Cero (lo cual es siempre Cierto). Por lo tanto, usamos la suma de los cuadrados.

…pero en B los datos están mucho mas dispersos que en A



4 -10 = - 6
16 -10 = +6
7- 10 = -3
13 -10 = +3
Suma = 0

La varianza

Calculemos la suma de las desviaciones al cuadrado para ambas poblaciones:

$$\sigma_A^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = 2$$

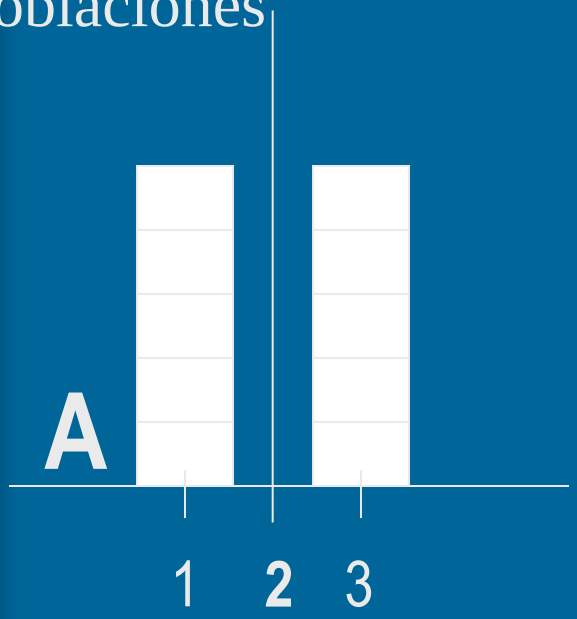
$$\sigma_B^2 = \frac{(4-10)^2 + (7-10)^2 + (10-10)^2 + (13-10)^2 + (16-10)^2}{5} = 18$$

¿Por qué la varianza esta definida como un promedio de desviaciones al cuadrado y no como su simple suma?

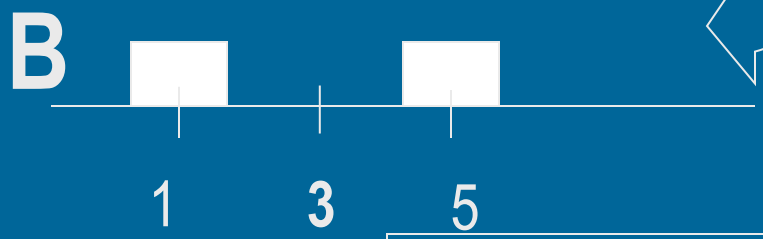
La suma de las desviaciones al cuadrado aumentan cuando la dispersión de aumenta!!

La varianza

Calculemos la suma de las desviaciones cuadradas para ambas poblaciones



B es mas dispersa alrededor de su media que A. Sin embargo, la suma no muestra eso.
Es por ello que se usa el promedio



$Suma_A = (1-2)^2 + \ldots + (1-2)^2 + (3-2)^2 + \ldots + (3-2)^2 = 10$

5 veces 5 veces

$Suma_B = (1-3)^2 + (5-3)^2 = 8$!

$\sigma_A^2 = Sum_A/N = 10/5 = 2$

$\sigma_B^2 = Sum_B/N = 8/2 = 4$

Desviación estándar

- La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en m^2). 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 estándar**, S, como

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

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

Grados de libertad

- ¿Por qué calculamos la varianza dividiendo por n - 1, en lugar de dividir por n?
- Como la suma de las desviaciones es 0, la última desviación es una combinación lineal de las n - 1 desviaciones restantes.
- Por lo tanto, no estamos calculando el promedio de n números independientes (los desvíos). Solo n -1 de las desviaciones al cuadrado pueden variar libremente y por ello, promediamos la suma de los desvíos al cuadrado dividiendo por n -1.
- Al numero n -1 se lo denomina **grados de libertad** de la varianza o de la desviación típica.

Ejemplo

- Calcular la varianza y desviación estándar de las siguientes cantidades medidas en metros:

3,3,4,4,5

Solución

- Para calcular dichas medidas de dispersión es necesario calcular previamente el valor con respecto al cual vamos a medir las diferencias. Ésta es la media:

$$\bar{x} = \frac{3+3+4+4+5}{5} = 3{,}8\ m$$

- La varianza es:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i^2 - \bar{x})^2 n_i}{n-1} = \frac{(3^2 + 3^2 + 4^2 + 4^2 + 5^2) - 3{,}8^2 \times 5}{4} = 0{,}70 \ \mathrm{m}^2$$

- siendo la desviación estándar su raíz cuadrada:

$$S = \sqrt{S^2} = \sqrt{0{,}70} = 0{,}84 \, \mathrm{m}$$

“La desviación estándar y la varianza son las medidas de separación con respecto a la media”

Propiedades del desvío standard

- S mide la dispersión respecto a la media. Debe emplearse solo cuando se escoge la media como medida central de la distribución.
- S = 0 solo ocurre cuando no hay dispersión: todas las observaciones toman el mismo valor. De lo contrario S > 0.
- Cuanto más dispersión hay entre las observaciones, mayor es S.
- S, al igual que la media, se encuentra fuertemente influenciado por las observaciones extremas.

Descripción de una distribución asimétrica

- Una distribución asimétrica con unas pocas observaciones en la cola larga de la distribución tendrá un desvío standard grande. En tal caso, **S** no proporciona información útil sobre la dispersión de la distribución.
- Como en una distribución muy asimétrica la dispersión de cada una de las colas es muy distinta, es imposible describir bien la dispersión con un solo número.
- Los cinco números resumen proporcionan mejor información sobre la dispersión de la distribución.
- *Es preferible utilizar los cinco números resumen en lugar de la media y el desvío standard para describir una distribución asimétrica*

Coeficiente de variación

- El coeficiente de variación es una medida de dispersión **relativa.**
- Muestra la dispersión de una distribución en relación a su media.
- Se utiliza para comparar distintas distribuciones.
- Su fórmula es:

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

- Por ejemplo, un desvio standard de 10, puede ser grande si la media es 100, pero no lo es si la media es 500.

Ejemplo

- Comparamos pesos de elefantes y ratas:

$$CVe = 3{,}09\%$$

$$CVr = 15{,}53\%$$

Elefantes

$$\overline{X} = 1088Kg$$

$$S = 58{,}3Kg$$

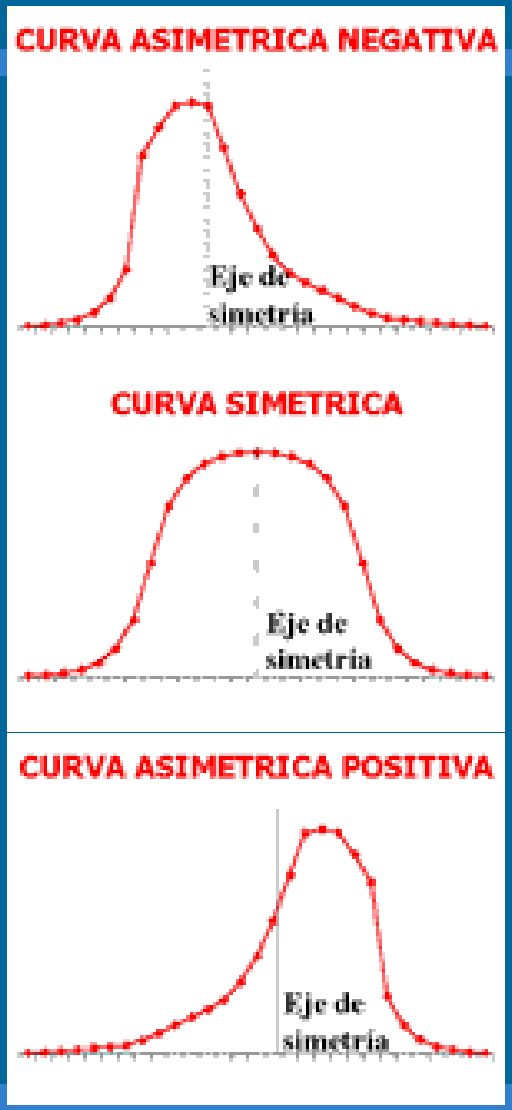
Ratas

$$\overline{X} = 0{,}047Kg$$

$$S = 0{,}0073Kg$$

Asimetría o sesgo

- **Asimetría:** El concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a izquierda y derecha de un valor central (media aritmética)



Coeficiente de asimetría de Pearson

$$S_{KP} = \frac{3 \times \left(\bar{x} - M_{ed} \right)}{S}$$

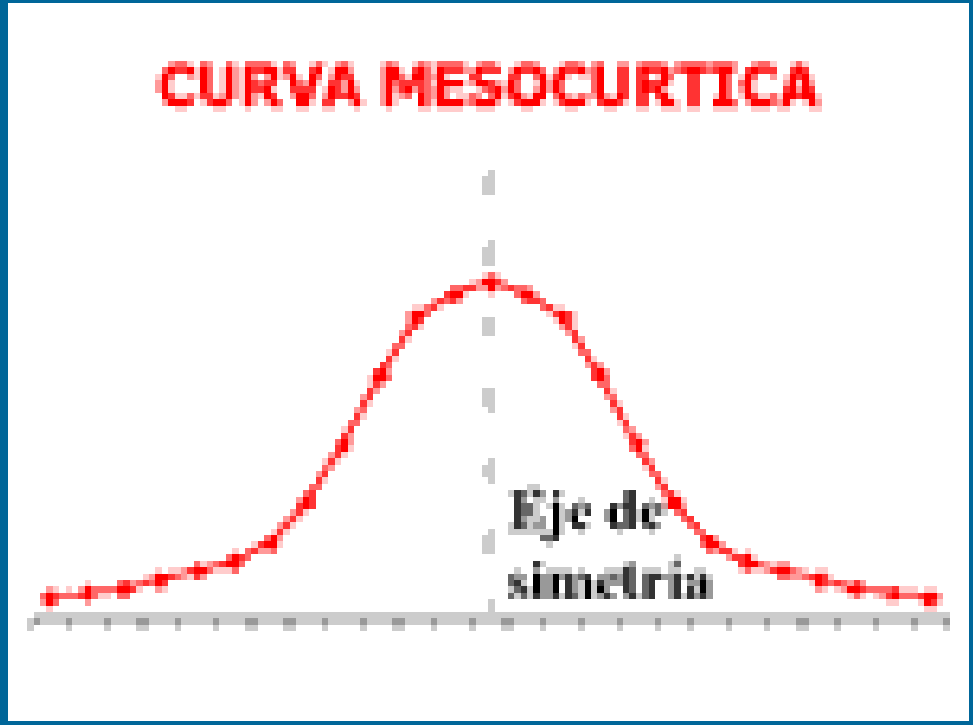
- **$S_{KP} = 0$** Distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media.
- **$S_{KP} > 0$** Distribución a simétrica positiva; existe mayor concentración de valores a la derecha de la media que a su izquierda.
- **$S_{KP} < 0$** Distribución a simétrica negativa; existe mayor concentración de valores a la izquierda de la media que a su derecha.

Apuntamiento (Curtosis)

- El Coeficiente de Curtosis analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución.

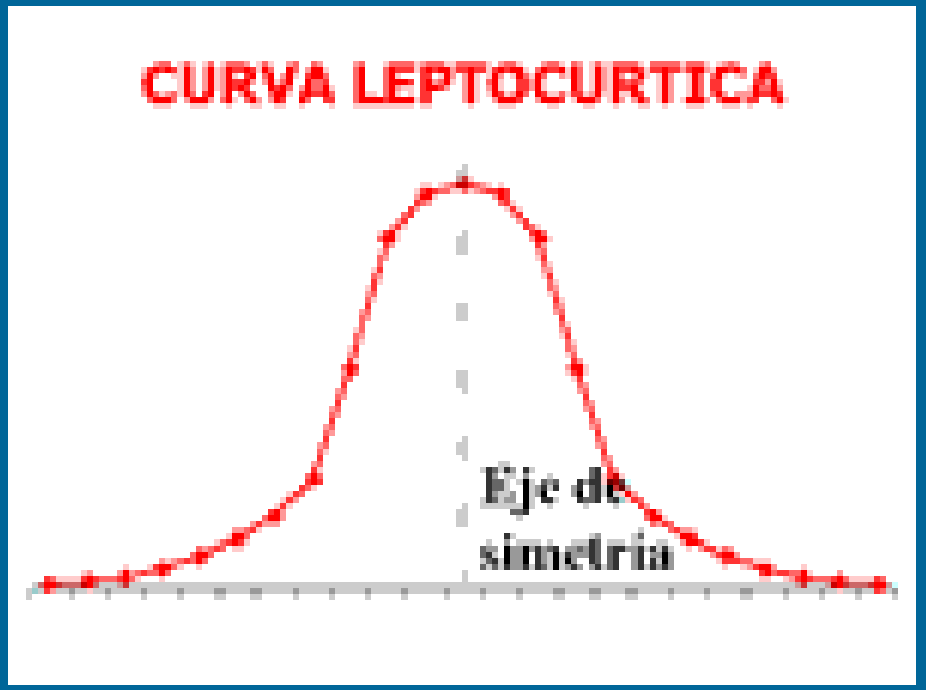
Distribución Mesocúrtica

- Presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).



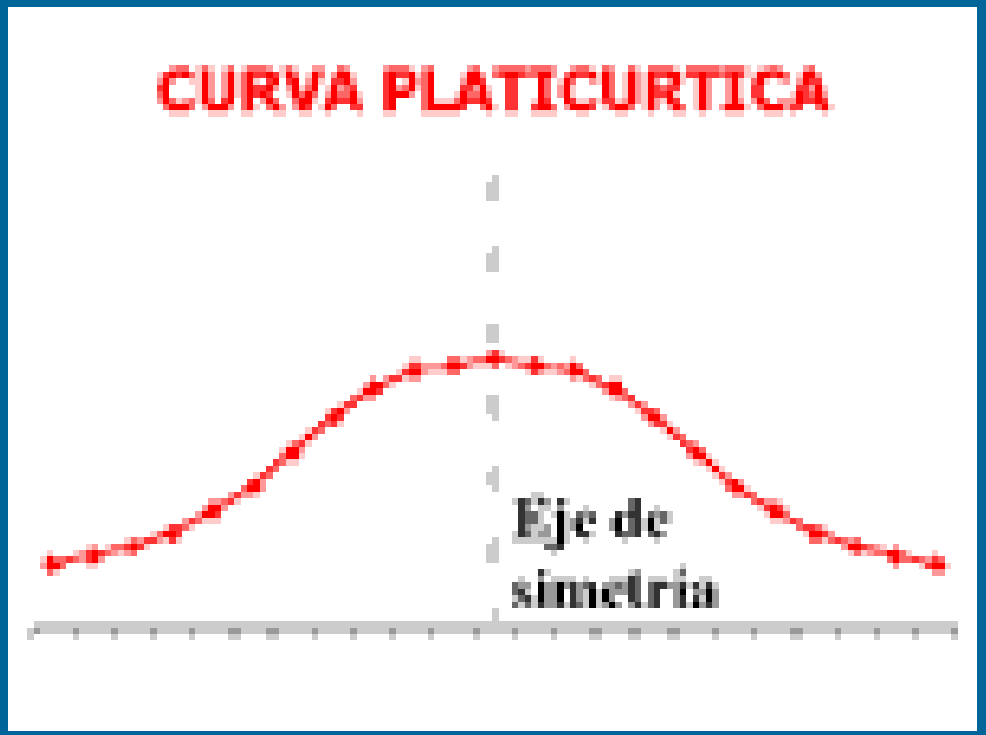
Distribución Leptocúrtica

- Presenta un elevado grado de concentración alrededor de los valores centrales de la variable.



Distribución Platicúrtica

- Presenta un reducido grado de concentración alrededor de los valores centrales de la variable.



Coeficiente de Curtosis

- $g_2 = 0$ (distribución mesocúrtica).
- $g_2 > 0$ (distribución leptocúrtica).
- $g_2 < 0$ (distribución platicúrtica).

$$g_2 = \frac{\left(\frac{1}{n}\right) \times \sum \left(x_i - \bar{x}\right)^4 \times n_i}{\left(\left(\frac{1}{n}\right) \times \sum \left(x_i - \bar{x}\right)^2 \times n_i\right)^2} - 3$$

Ejemplo

Vamos a calcular el Coeficiente de Curtosis de la serie de datos referidos a la estatura (altura a la cruz) de los terneros de un lote a remate visto anteriormente.

Variable	Frecuencias absolutas		Frecuencias relativas	
(Valor)	Simple	Acum	Simple	Acum
1,20	1	1	3,3%	3,3%
1,21	4	5	13,3%	16,6%
1,22	4	9	13,3%	30,0%
1,23	2	11	6,6%	36,6%
1,24	1	12	3,3%	40,0%
1,25	2	14	6,6%	46,6%
1,26	3	17	10,0%	56,6%
1,27	3	20	10,0%	66,6%
1,28	4	24	13,3%	80,0%
1,29	3	27	10,0%	90,0%
1,30	3	30	10,0%	100,0%

Recordemos que la media de esta muestra es 1,253

$$g_2 = \frac{\left(\frac{1}{n}\right) \times \sum \left(x_i - \bar{x}\right)^4 \times n_i}{\left(\left(\frac{1}{n}\right) \times \sum \left(x_i - \bar{x}\right)^2 \times n_i\right)^2} - 3$$

$$g_2 = \frac{\left(\frac{1}{30}\right) \times 0{,}00004967}{\left(\left(\frac{1}{30}\right) \times (0{,}03046667)\right)^2} - 3 = -1{,}39$$

- El Coeficiente de Curtosis de esta muestra es -1,39. Se trata de una distribución ***Platicúrtica***, es decir, con una reducida concentración alrededor de los valores centrales de la distribución.