

Fundamentos de Análisis de Datos

Estadística Descriptiva (Caso Univariado)

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA

La estadística descriptiva analiza, resume y ordena las características básicas de un conjunto de datos.

Análisis de **distibución** relacionado con la frecuencia de los datos

Análisis de **variabilidad** o **dispersión**

Análisis de **tendencia central** a partir de valores promedio

Ejemplo de base

Paquete	Peso en kilogramos
A	45
B	39
C	53
D	45
E	43
F	48
G	50
H	45

Medidas de Tendencia Central

Media

Es el valor promedio de los datos.

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, se define la **media muestral** como:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

En nuestro ejemplo, donde $n = 8$,

$$\bar{X} = \frac{1}{8}(45 + 39 + 53 + 45 + 43 + 48 + 50 + 45) = \frac{368}{8} = \boxed{46}.$$

Media

Observación

La media es sensible a valores atípicos.

Supongamos que ingresa un nuevo paquete que pesa 1 kilogramo.
¿Cuál es la media de la nueva muestra Y ?

$$\bar{Y} = \frac{1}{9}(45 + 39 + 53 + 45 + 43 + 48 + 50 + 45 + 1) = \frac{369}{9} = \boxed{41}.$$

Y si el nuevo paquete pesa 100 kilogramos, ¿cuál es la media de la nueva muestra Z ?

$$\bar{Z} = \frac{1}{9}(45 + 39 + 53 + 45 + 43 + 48 + 50 + 45 + 100) = \frac{468}{9} = \boxed{52}.$$

Mediana

Es el valor que se halla en la mitad de los datos ordenados de manera ascendente.

Definición

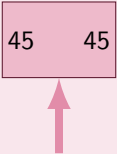
Dado un conjunto de datos ordenados $X = \{X_1 \leq X_2 \leq \dots \leq X_n\}$, se define la **mediana** como:

$$\tilde{X} = \begin{cases} X_{(n+1)/2} & \text{si } n \text{ es impar,} \\ \frac{X_{n/2} + X_{n/2+1}}{2} & \text{si } n \text{ es par.} \end{cases}$$

Mediana

Ordenando los datos de nuestro ejemplo:

39 43 45 45 45 48 50 53


$$\tilde{X} = \frac{45 + 45}{2} = 45$$

Mediana

Pregunta

¿Cómo se comporta la mediana frente a valores atípicos?

Siguiendo con el mismo ejemplo que usamos para la media,

1 39 43 45 45 45 48 50 53

\tilde{Y}

39 43 45 45 45 48 50 53 100

\tilde{Z}

Vemos que la mediana **no** es sensible a valores atípicos.

Moda

Es el valor más frecuente de los datos, que puede ser usado tanto para variables categóricas como numéricas.

En nuestro ejemplo todos los pesos aparecen una única vez, excepto por el valor 45 que aparece tres veces.

Entonces, la moda de los pesos de los paquetes es 45.

Medidas de Variabilidad

Rango

Mide el grado de dispersión de los datos como la distancia entre los valores máximo y mínimo de los datos.

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, sean X_{\min} y X_{\max} los valores mínimo y máximo de X , respectivamente. El **rango** se define como:

$$\text{rg}(X) = X_{\max} - X_{\min}.$$

En nuestro ejemplo:

$$\text{rg}(X) = 53 - 39 = \boxed{14}.$$

Varianza

Refleja el grado de dispersión de los datos.

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, se define la **varianza muestral** como:

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Observación

El denominador de la varianza muestral es $n-1$ para obtener un estimador insesgado de la varianza poblacional.

Varianza

En nuestro ejemplo:

$$\begin{aligned}s_X^2 &= \frac{1}{7} [(45 - 46)^2 + (39 - 46)^2 + (53 - 46)^2 + (45 - 46)^2 + \\ &\quad (43 - 46)^2 + (48 - 46)^2 + (50 - 46)^2 + (45 - 46)^2] \\ &= \frac{1}{7} [1^2 + 7^2 + 7^2 + 1^2 + 3^2 + 2^2 + 4^2 + 1^2] = \frac{130}{7} = \boxed{18.57143}.\end{aligned}$$

Desvío estándar

Proporciona información sobre la distancia entre un valor del conjunto de datos y la media de este conjunto.

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, se define el **desvío estándar muestral** como:

$$s_X = \sqrt{s_X^2}.$$

En nuestro ejemplo:

$$s_X = \sqrt{18.57143} = \boxed{4.309458}.$$

Desvío estándar

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, se define el **desvío estándar poblacional** como:

$$\sigma_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

Cuartiles

Se utilizan para describir la dispersión de los datos.

Definición

Los **cuartiles** son los valores que separan el conjunto de datos en cuatro partes iguales. Explícitamente:

- ➡ Q_0 es el valor mínimo de los datos,
- ➡ Q_1 es el valor que separa el primer cuarto de los datos del resto,
- ➡ Q_2 es el valor que separa los datos en la mitad,
- ➡ Q_3 es el valor que separa el tercer cuarto del último cuarto de los datos,
- ➡ Q_4 es el valor máximo de los datos.

Cuartiles

Observación

- ▢ Q_2 coincide con la mediana.
- ▢ Entre Q_0 y Q_1 se encuentra el 25% de los valores más bajos de los datos. Entre Q_1 y Q_2 el siguiente 25%. Y así, sucesivamente.

Pregunta

¿Cómo calculamos los cuartiles de un conjunto de datos
 $X = \{X_1, X_2, \dots, X_n\}$?

Cuartiles

- 1 Ordenamos los datos en forma creciente.
- 2 Calculamos el primer cuartil:

$$Q_1 = \begin{cases} (X_{n/4} + X_{n/4+1})/2 & \text{si } n/4 \text{ es un entero,} \\ X_{r_1} & \text{en caso contrario, donde } r_1 \text{ es el primer} \\ & \text{entero mayor que } n/4. \end{cases}$$

Cuartiles

- 3 Calculamos el segundo cuartil:

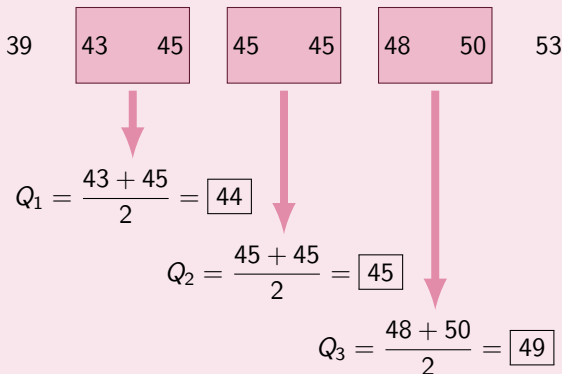
$$Q_2 = \begin{cases} (X_{2n/4} + X_{2n/4+1})/2 & \text{si } 2n/4 \text{ es un entero,} \\ X_{r_2} & \text{en caso contrario, donde } r_2 \text{ es el primer} \\ & \text{entero mayor que } 2n/4. \end{cases}$$

- 4 Calculamos el tercer cuartil:

$$Q_3 = \begin{cases} (X_{3n/4} + X_{3n/4+1})/2 & \text{si } 3n/4 \text{ es un entero,} \\ X_{r_3} & \text{en caso contrario, donde } r_3 \text{ es el primer} \\ & \text{entero mayor que } 3n/4. \end{cases}$$

Cuartiles

Volvamos a nuestro ejemplo del peso de los paquetes, donde $n = 8$. Entonces $n/4 = 2$, $2n/4 = 4$ y $3n/4 = 6$ son todos números enteros.



Rango intercuartil

Indica la dispersión del 50% de los datos intermedios.

Definición

El **rango intercuartil** está dado por:

$$IQR = Q_3 - Q_1.$$

En nuestro ejemplo, $IQR = 49 - 44 = \boxed{5}$.

Outlier

Un *outlier* es un valor atípico.

Definición

Un **outlier** puede definirse como una observación inferior al primer cuartil o superior al tercer cuartil cuya distancia excede $1.5/IQR$.

En nuestro ejemplo, $1.5/IQR = 7.5$.

Como

- $Q_1 - 7.5 = 44 - 7.5 = 36.5$,
- $Q_3 + 7.5 = 49 + 7.5 = 56.5$,

cualquier valor menor que 36.5 o mayor que 56.5 puede considerarse un outlier.

Percentiles

Se utilizan para conocer el valor que una variable debe tomar para estar dentro del $x\%$ de los valores superiores.

Definición

Los **percentiles** son los valores que separan el conjunto de datos en cien partes iguales.

Por ejemplo, el percentil 95 denotado por $P_{95\%}$, separa el 95% de los valores del 5% de los más altos.

Observación

- ⇒ El percentil 25, $P_{25\%}$, coincide con el primer cuartil Q_1 .
- ⇒ El percentil 50, $P_{50\%}$, coincide con el segundo cuartil Q_2 y con la mediana.
- ⇒ El percentil 75, $P_{75\%}$, coincide con el tercer cuartil Q_3 .

Ejemplo*

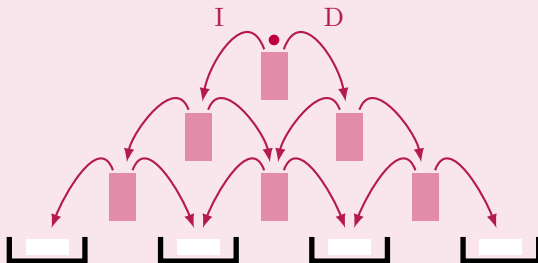


*Base de datos iris de R

Distribución

Tablero de Galton

Fue creado por Galton (1822-1911) para demostrar el Teorema Central del Límite.



Distribución Normal



Carl Friedrich Gauss (1777-1855)

La distribución Normal, también conocida como distribución Gaussiana, es quizás la distribución más importante en Estadística.

Distribución Normal

Importancia

- ➡ Muchos fenómenos naturales y estudios de mercado pueden modelarse con una distribución Normal.
- ➡ Es la base del Teorema Central del Límite, que establece la distribución Normal aproximada del promedio de variables aleatorias independientes e idénticamente distribuidas, sin importar el tipo de distribución de las variables muestreadas.
- ➡ Una gran cantidad de procedimientos estadísticos asumen la distribución Normal de los datos. Por ejemplo, análisis de correlación, modelos de regresión, aplicación de test estadísticos y ANOVA.

Distribución Normal

Función de densidad de probabilidad

Si X es una variable aleatoria con distribución Normal, su función de densidad de probabilidad para $x \in \mathbb{R}$ es de la forma:

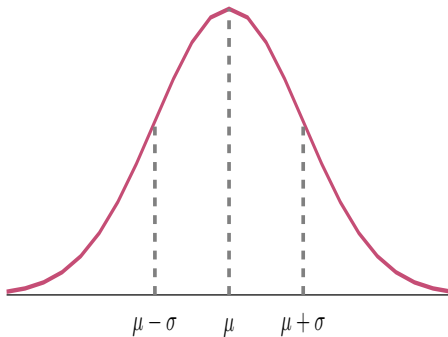
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right],$$

donde μ es la media y σ^2 la varianza.

En este caso notamos $X \sim \mathcal{N}(\mu, \sigma^2)$.

Distribución Normal

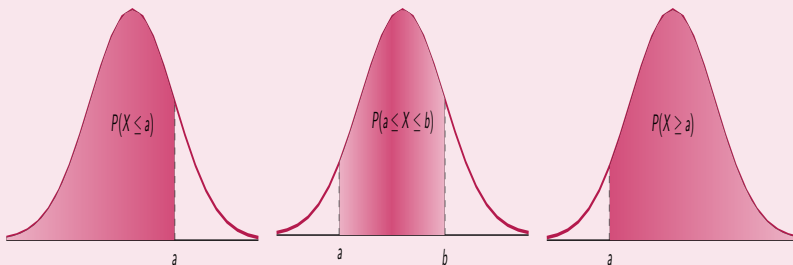
Gráfica



Distribución Normal

Pregunta

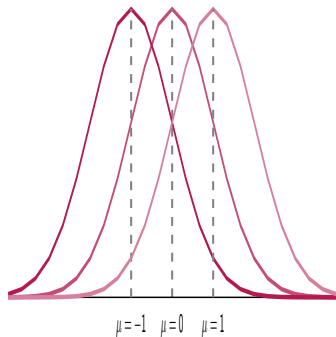
¿Cuál es la interpretación gráfica de la función de densidad de probabilidad?



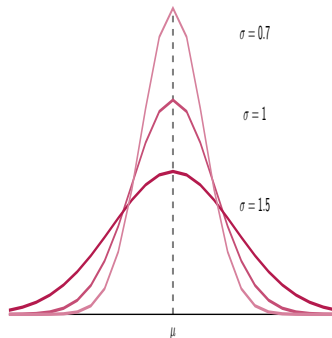
Distribución Normal

Gráfica

Misma varianza pero distinta media.



Misma media pero distinta varianza.



Distribución Normal

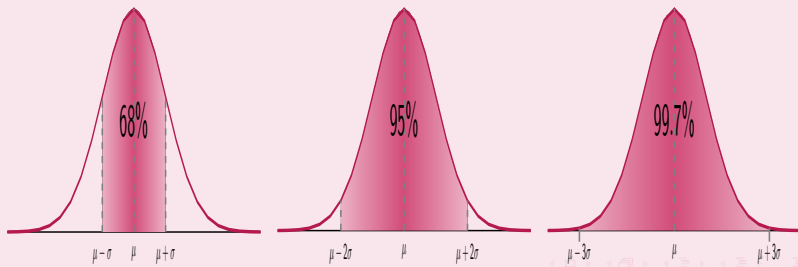
Propiedades

- ➡ El gráfico de la distribución tiene la forma de una campana.
- ➡ La distribución es simétrica respecto de la media; es decir, los lados izquierdo y derecho de la gráfica se espejan.
- ➡ Los datos cercanos a la media son más frecuentes que aquellos que están lejos de la misma.
- ➡ La media, la mediana y la moda coinciden.
- ➡ El pico -valor máximo- de la curva se alcanza en la media.
- ➡ La curva es asintótica respecto del eje de abscisas.
- ➡ El gráfico tiene dos puntos de inflexión -donde cambia la concavidad- en $\mu \pm \sigma$.
- ➡ Cuanto más angosta es la curva, menor es la probabilidad de que las observaciones se alejen de la media.

Distribución Normal

Observación

- ➡ El 68% de los valores se hallan a una desviación estándar de la media.
- ➡ El 95% de los valores se hallan a dos desviaciones estándar de la media.
- ➡ El 99.7% de los valores se hallan a tres desviaciones estándar de la media.



Distribución Normal Estándar

Es la distribución Normal con media cero y varianza unitaria; es decir, $\mu = 0$ y $\sigma^2 = 1$.

En este caso notamos $Z \sim \mathcal{N}(0, 1)$ y la función de densidad de probabilidad se reduce a:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Distribución Normal estándar

Observación

Sea $X \sim \mathcal{N}(\mu, \sigma^2)$, la variable $Z = \frac{X - \mu}{\sigma}$ tiene distribución Normal estándar.

Pregunta

¿Por qué estandarizar?

- Los valores estandarizados son útiles para comprender la ubicación relativa a la distribución completa de una observación.
- La estandarización permite comparar observaciones y calcular probabilidades de poblaciones diferentes.

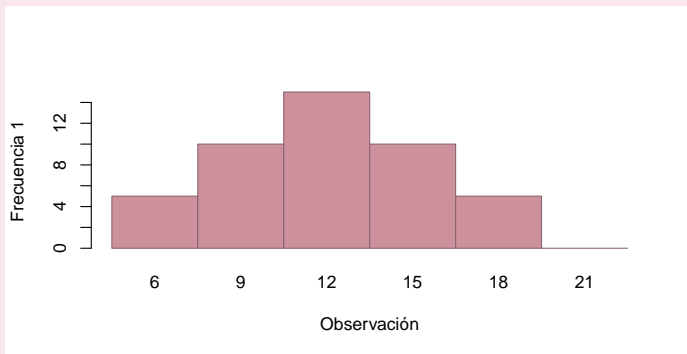
Frecuencia

Se usa tanto para datos cuantitativos como cualitativos, y cuenta la cantidad de veces que un valor ocurre en el conjunto de datos, presentando los datos de una manera más estructurada y organizada.

Vamos a analizar las siguientes tablas de frecuencia para un conjunto de observaciones $X = \{6, 9, 12, 15, 18, 21\}$.

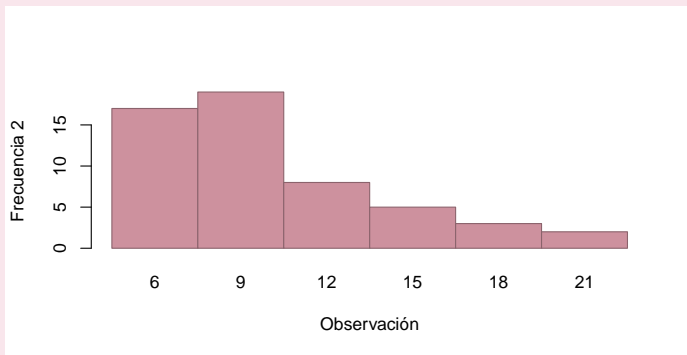
Observación	"6"	"9"	"12"	"15"	"18"	"21"
Frecuencia 1	5	10	15	10	5	0
Frecuencia 2	17	19	8	5	3	2
Frecuencia 3	2	13	5	10	15	19

Distribución



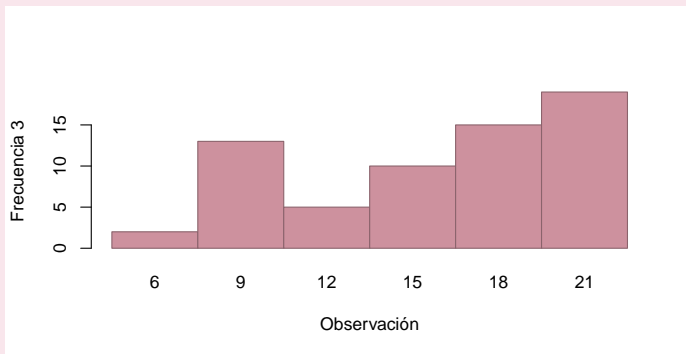
- ⇒ Media: 12
- ⇒ Mediana: 12
- ⇒ Moda: 12

Distribución



- ⇒ Media: 10
- ⇒ Mediana: 9
- ⇒ Moda: 9

Distribución

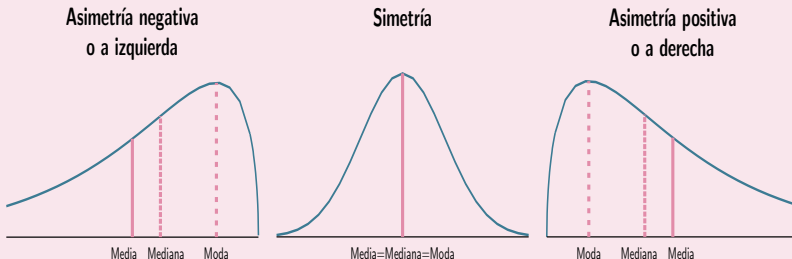


⇒ Media: 15.75

⇒ Mediana: 18

⇒ Moda: 21

Distribuciones sesgadas



- ➡ **Distribución con asimetría negativa:** la “cola” a la izquierda de la media es más larga que la de la derecha; es decir, hay valores más separados de la media a la izquierda.
- ➡ **Distribución simetría:** existe el mismo número de valores a la derecha que a la izquierda de la media.
- ➡ **Distribución con asimetría positiva:** la “cola” a la derecha de la media es más larga que la de la izquierda; es decir, hay valores más separados de la media a la derecha.

Distribuciones sesgadas

Observación

Para datos con distribución sesgada, la mediana es un mejor indicador que la media debido a que no se ve influenciada por valores atípicos.

Asimetría

Un conjunto de datos se dice simétrico si las observaciones se destruyen de igual manera a la izquierda y a la derecha del valor central.

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, el **coeficiente de asimetría de Fisher** se calcula como:

$$A_F = \frac{\sum_{i=1}^n (X_i - \bar{X})^3 / n}{\sigma_X^3}.$$

Asimetría

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, el **coeficiente de asimetría de Pearson** se calcula como:

$$A_P = \frac{\bar{X} - \text{Moda de } X}{\sigma_X}.$$

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, el **coeficiente de asimetría de Bowley** se calcula como:

$$A_B = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}.$$

Asimetría

Observación

- ➡ Un coeficiente de asimetría significativamente no nulo indica que los datos no son normales. Equivalentemente, si los datos son aproximadamente normales, entonces el coeficiente de asimetría es cercano a cero.
- ➡ Un valor de asimetría igual a cero no significa necesariamente que los datos son simétricos.
- ➡ Se utiliza el desvío poblacional para los cálculos de los coeficientes de Fisher y Pearson.
- ➡ El coeficiente de Pearson sólo se puede aplicar para datos unimodales.
- ➡ Su equivalente en inglés es *skewness*.

Asimetría

Comparemos los coeficientes de asimetría de las tres frecuencias de nuestro ejemplo.

	A_F	A_P	A_B
Frecuencia 1	0.00	0.00	0.00
Frecuencia 2	1.06	0.25	0.00
Frecuencia 3	-0.46	-1.10	-0.33

Curtosis

Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, la **curtosis** se calcula como:

$$K = \frac{\sum_{i=1}^n (X_i - \bar{X})^4 / n}{\sigma_X^4}.$$

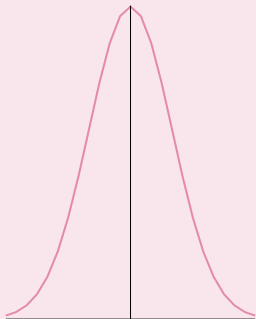
Definición

Dado un conjunto de datos $X = \{X_1, X_2, \dots, X_n\}$, el **exceso de curtosis** se calcula como:

$$K' = K - 3.$$

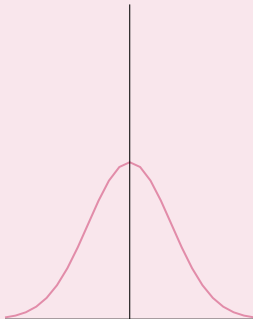
Tipos de curtosis

Leptocúrtica



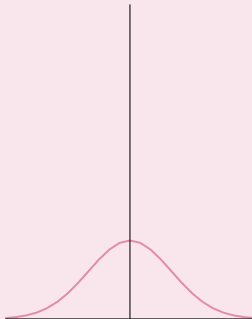
$$K' > 0$$

Mesocúrtica



$$K' = 0$$

Platicúrtica



$$K' < 0$$

Curtosis

Observación

- ➡ La razón de restar 3 en el exceso de curtosis se debe a que la curtosis de la distribución Normal es precisamente 3.
- ➡ Un conjunto de datos con exceso de curtosis positivo tiene una distribución de colas pesadas -llega a cero más lentamente-, por lo que tiene una tendencia a la presencia de outliers.
Por el contrario, si el valor de exceso de curtosis es negativo entonces las colas de la distribución no son pesadas indicando ausencia de outliers.
- ➡ Si los datos son normales, la curtosis es cercana a tres o, equivalentemente, el exceso de curtosis es cercano a cero.

Curtosis

Comparemos los valores de curtosis de las tres frecuencias de nuestro ejemplo.

	K	K'	Tipo
Frecuencia 1	2.25	-0.75	Platicúrtica
Frecuencia 2	3.42	0.42	Leptocúrtica
Frecuencia 3	1.83	-1.17	Platicúrtica

Ejemplo

