# MACHINE LEARNING

CIENCIAS DE LA COMPUTACIÓN
M.Sc. Richard F. Fernández Vásquez


Recogida de Datos en "crudo" (RAW)
Limpieza de Datos
Transformación y Feature Engineering
Aprendizaje y Entrenamiento del Modelo
Evaluación de calidad de las predicciones
80%
20%

Otorga Ventaja Competitiva

Genera Conocimiento

Mejores Decisiones

# ¿Por qué es importante el Análisis de Datos?

# ESTADÍSTICA


M.Sc. Richard F. Fernández Vásquez

# Definición de Estadística

Ciencia que se ocupa de **recolectar, organizar, resumir, analizar e interpretar** los **DATOS**, con el objetivo de establecer conclusiones y tomar decisiones confiables.

# Definición de Estadística

Ciencia que se ocupa de recolectar, organizar, resumir, analizar e interpretar los **DATOS**, con el objetivo de establecer conclusiones y tomar decisiones confiables.

1. Recolectar => Tener claro el objetivo. Fuentes secundarias o primarias (tomar muestras, aplicaciones, etc). Bases de datos internas de las organizaciones o externas.
2. Organizar => Estructurar la información, limpieza de datos, etc.
3. Resumir => estadística descriptiva, inferencial. Modelos estadísticos: regresión lineal, regresión logística, clúster, etc.
4. Analizar => resultados importantes.
5. Interpretar => conclusiones y accionar.

# Logro Unidad 1

Al finalizar la unidad, el alumno es capaz de aplicar adecuadamente técnicas de pre procesamiento de datos para posibilitar la implementación de una solución de Machine Learning para un problema del mundo real.

# Contenido 2

- Datos, instancias y atributos
- Análisis univariado de datos
- Análisis bivariado de datos
- Visualización de datos
- Caso de Aplicación

# Contenido 2

- Datos, instancias y atributos
- Análisis univariado de datos
- Análisis bivariado de datos
- Visualización de datos
- Caso de Aplicación

# Matriz de datos

Atributos
(Variables)

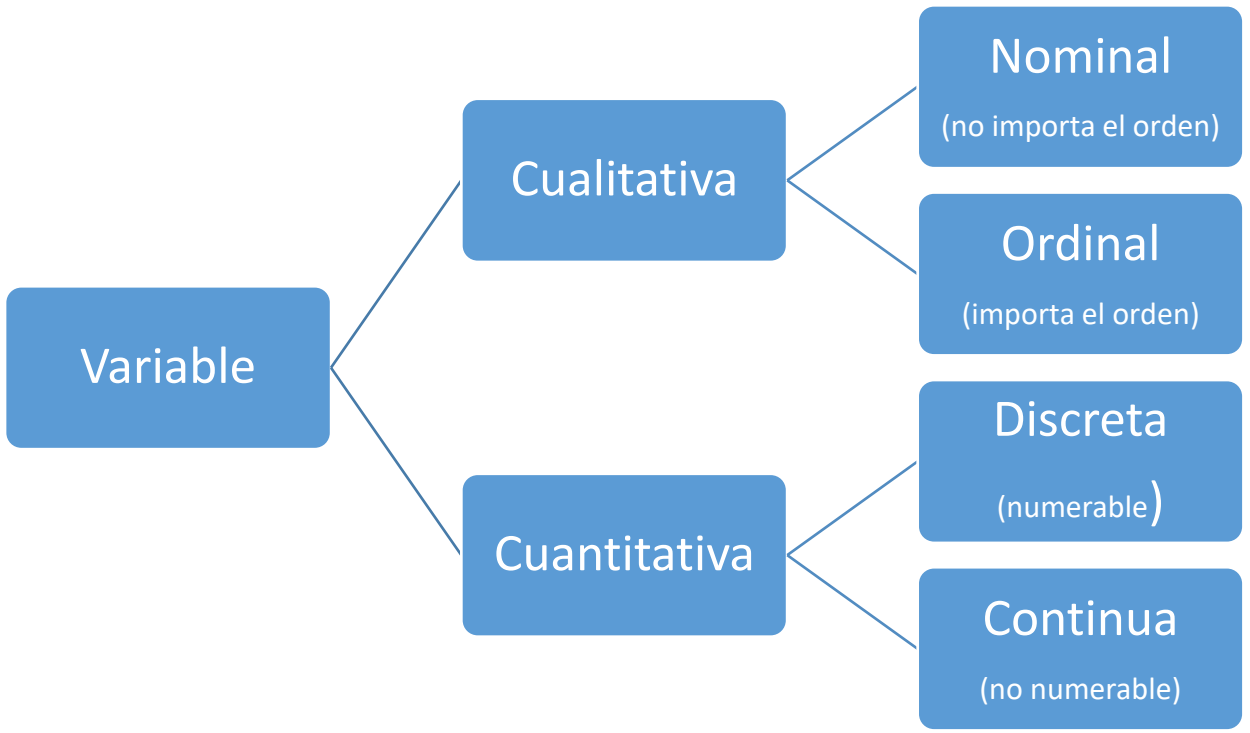
Instancias
(Registros)

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Unidades vendidas | Precio Unitario | |
| 2 | Producto A | 8 | $ 1,500.00 | |
| 3 | Producto B | 6 | $ 2,004.00 | |
| 4 | Producto C | 2 | $ 2,283.00 | |
| 5 | Producto D | 9 | $ 1,921.00 | |
| 6 | Producto E | 8 | $ 1,521.00 | |
| 7 | Producto F | 1 | $ 1,770.00 | |
| 8 | Producto G | 3 | $ 1,599.00 | |
| 9 | Producto H | 5 | $ 1,609.00 | |
| 10 | Producto I | 6 | $ 2,149.00 | |
| 11 | Producto J | 6 | $ 1,669.00 | |

Datos

# Variable

Es toda característica que se encuentra en estudio en una muestra o población.




**EJEMPLOS**

Estado civil, color del cabello, etc

Grado de instrucción, nivel socioeconómico, etc

Número de accidentes de tránsito, número de atenciones en una agencia bancaria, etc.

Peso (en kg) de una persona, el ingreso de un cliente, etc.

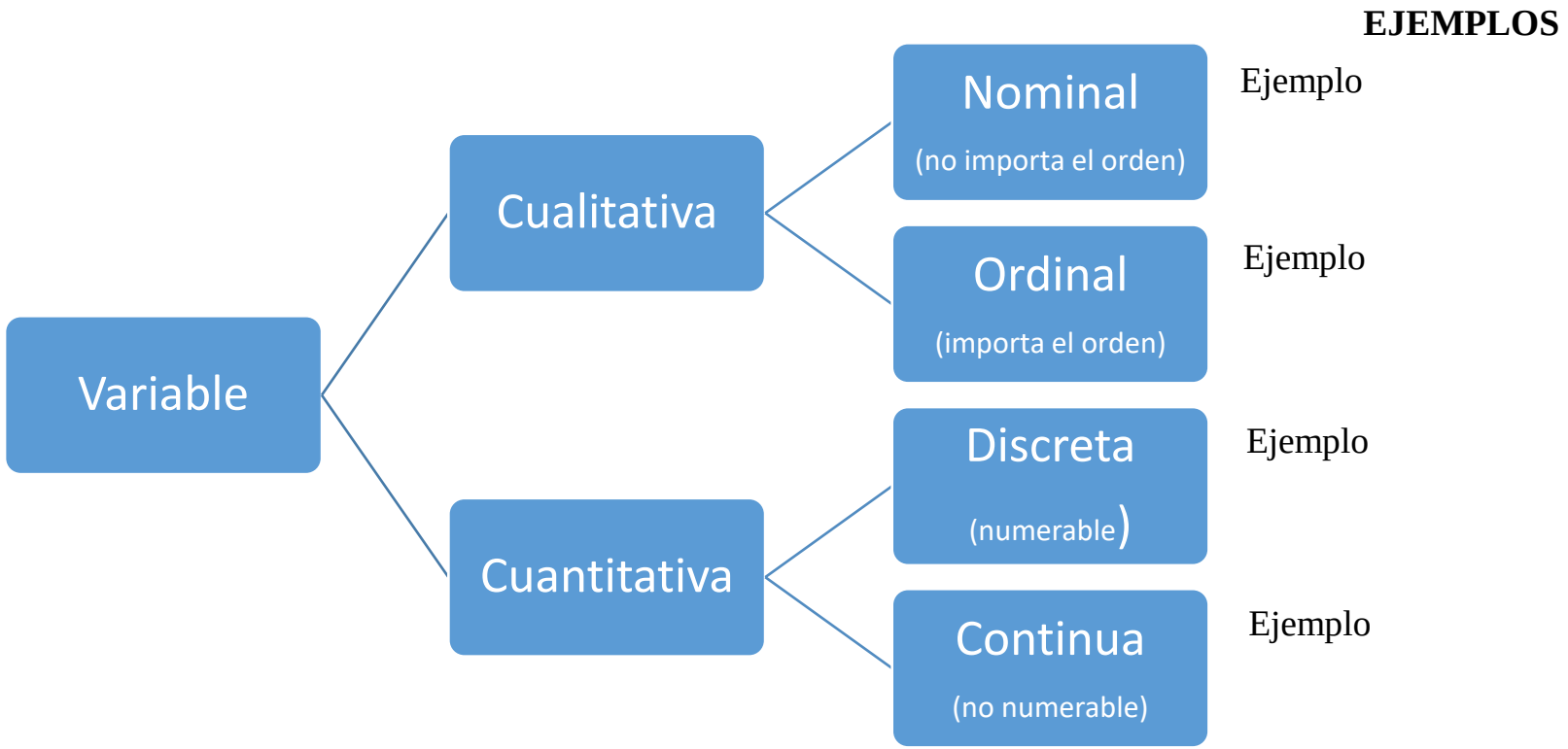
# Variable

Ejemplo:

| ID | Edad | Estado Civil | Consumo Promedio | Salario | Costo Unitario |
|---|---|---|---|---|---|
| 1 | Joven | Si | 150 | Alto | 10 |
| 2 | Joven | No | 250 | Medio | 10 |
| 3 | Joven | Si | 256 | Medio | 10 |
| 4 | Joven | Si | 500 | Bajo | 10 |
| 5 | Mayor | Si | 1200 | Bajo | 10 |
| 6 | Mayor | No | 508 | Medio | 10 |
| 7 | Joven | No | 205 | Medio | 10 |
| 8 | Joven | Si | 300 | Alto | 10 |
| 9 | Mayor | Si | 100 | Medio | 10 |
| 10 | Mayor | No | 150 | Bajo | 10 |

# Actividad

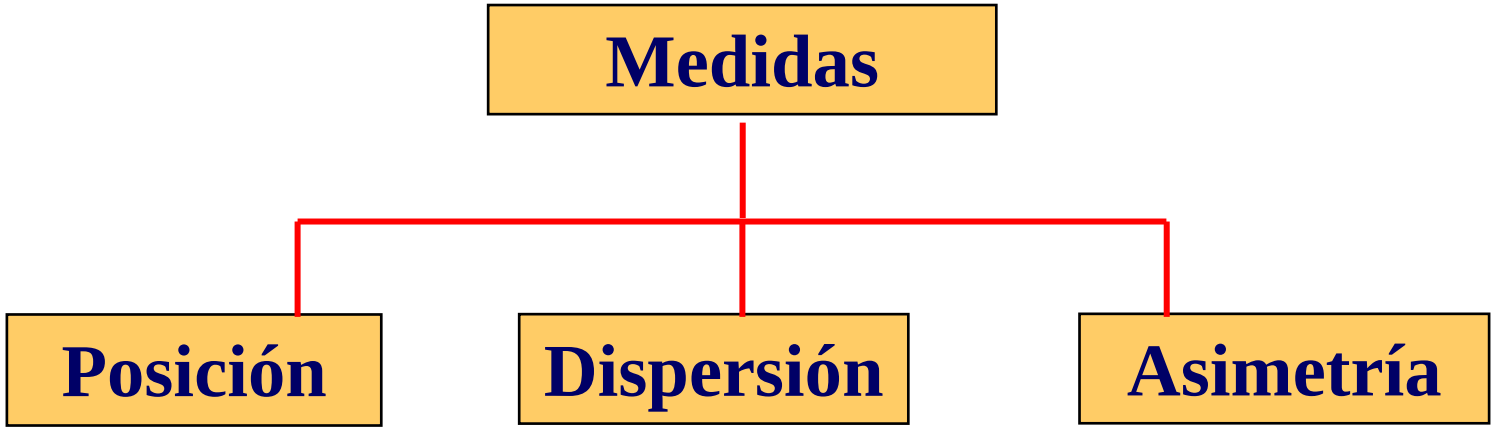
Brindar 4 ejemplos según el tipo de variable:



# Contenido 2

- Datos, instancias y atributos
- Análisis univariado de datos
- Análisis bivariado de datos
- Visualización de datos
- Caso de Aplicación

# Análisis Univariado

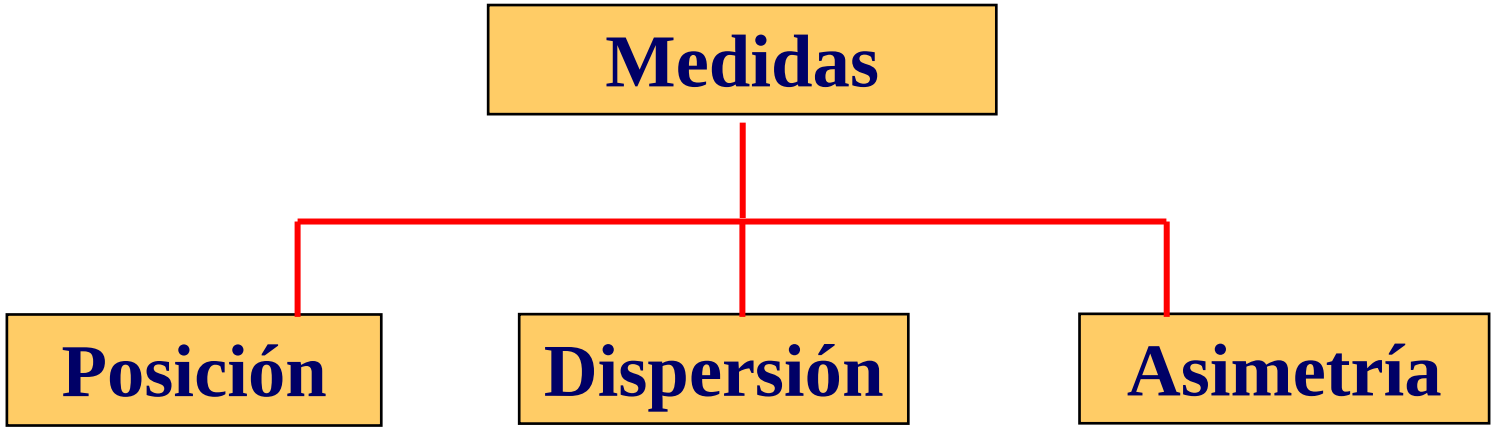



- La media
- La mediana
- La moda
- Percentiles

- El rango
- El rango intercuartílico
- La varianza
- La desviación estándar
- El coeficiente de variabilidad

- El coeficiente de asimetría
- El coeficiente de curtosis

# Análisis Univariado




- La media
- La mediana
- La moda
- Percentiles

- El rango
- El rango intercuartílico
- La varianza
- La desviación estándar
- El coeficiente de variabilidad

- El coeficiente de asimetría
- El coeficiente de curtosis

# Media

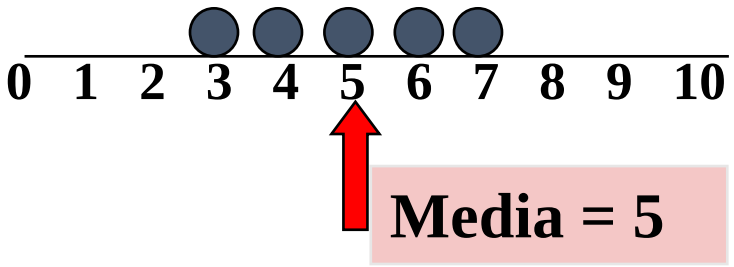
- Es la medida más común de tendencia central.

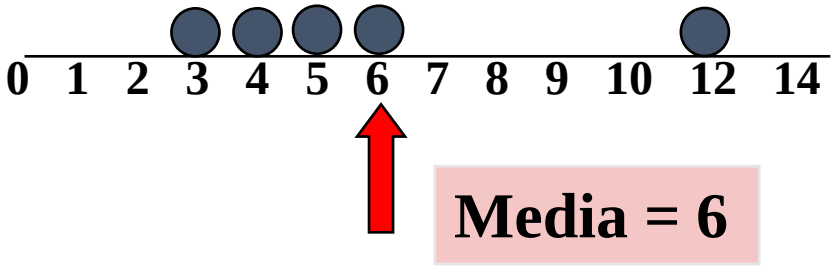
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Media Poblacional

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

Media Muestral



# Media

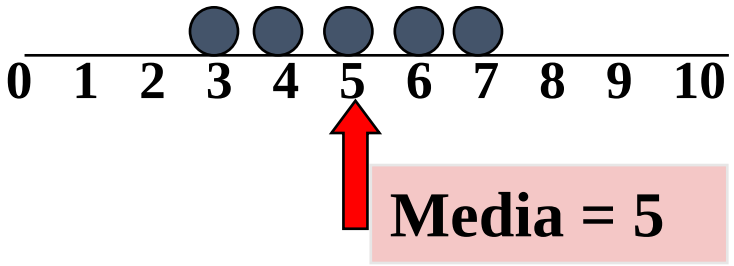
- Es la medida más común de tendencia central.

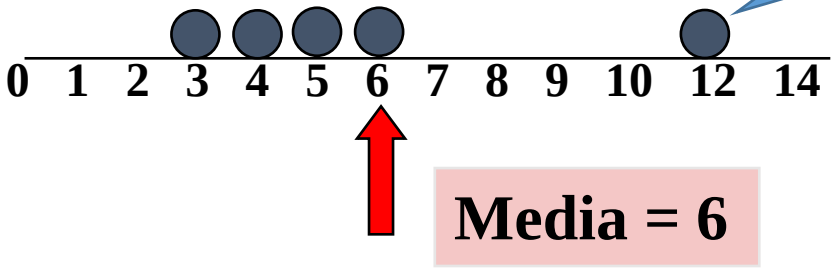
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Media Poblacional

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

Media Muestral




Outlier o Valor Atípico

Tendrán una explicación de negocio para determinar si es necesario removerlos o no.

# Mediana

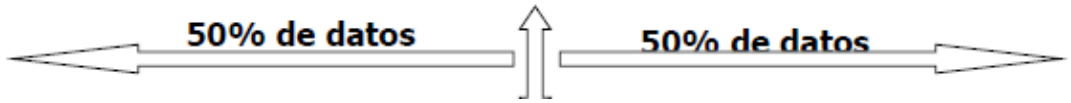
- Es el punto medio de los valores **después de ordenarlos** de menor a mayor, o de mayor a menor.
- No es afectada por valores extremos.

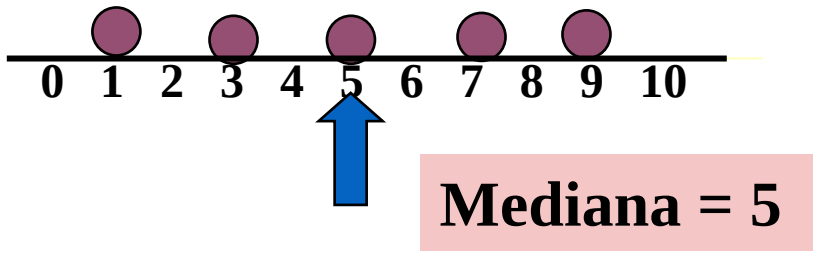
$$me = x_{\left(\frac{n+1}{2}\right)}$$

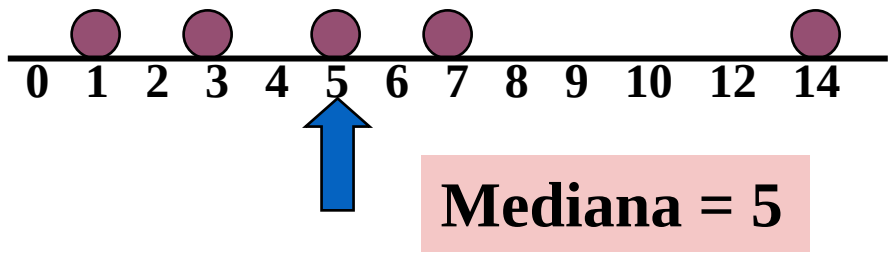
Si n es impar

$$me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

Si n es par

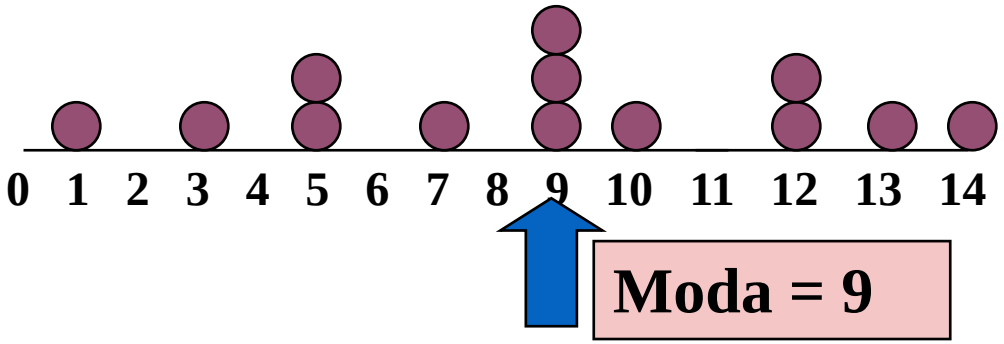


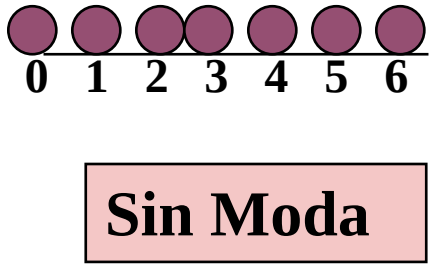



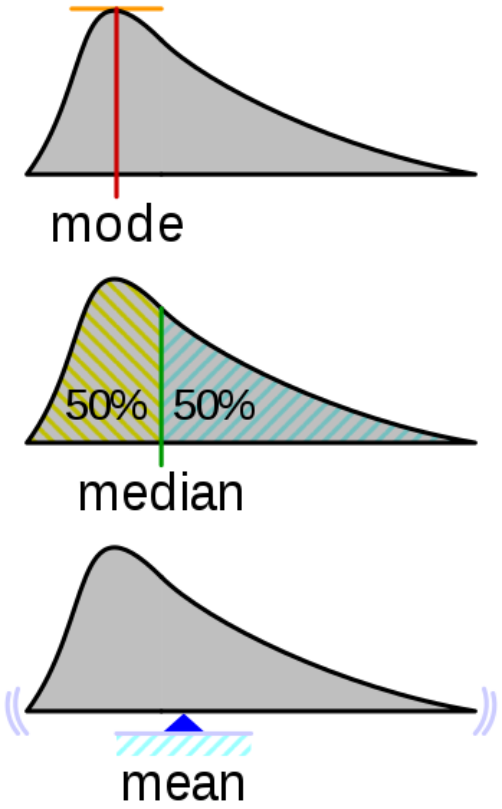
# Moda

- Valor que ocurre más a menudo.
- No es afectada por valores extremos.
- Puede no existir una moda.
- Pueden haber varias modas.



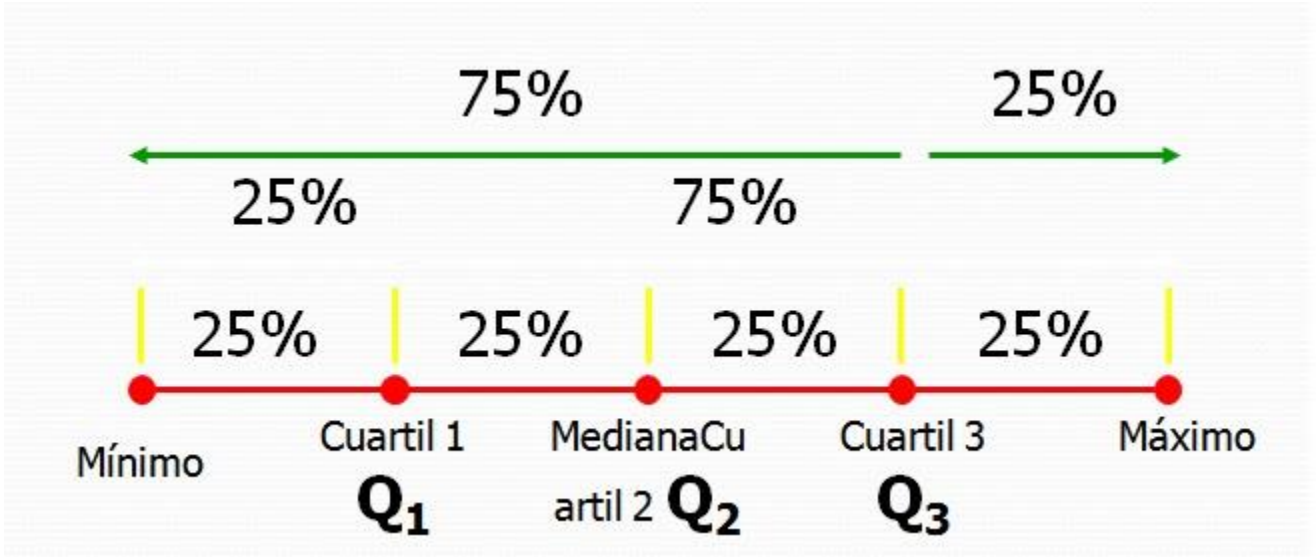



# Media, Mediana y Moda

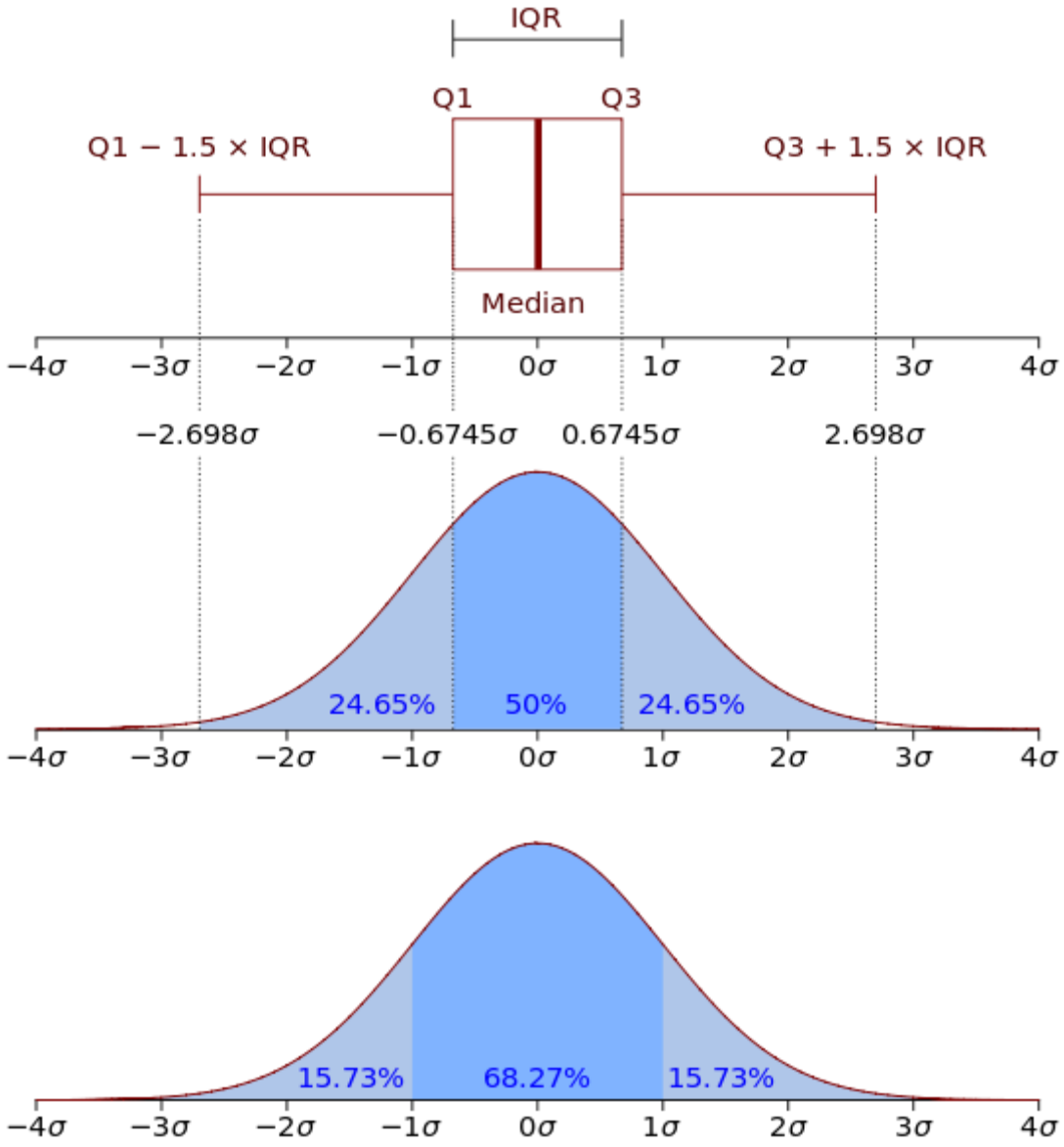


# Cuartil

- Dividen la distribución de frecuencias en cuatro partes iguales.

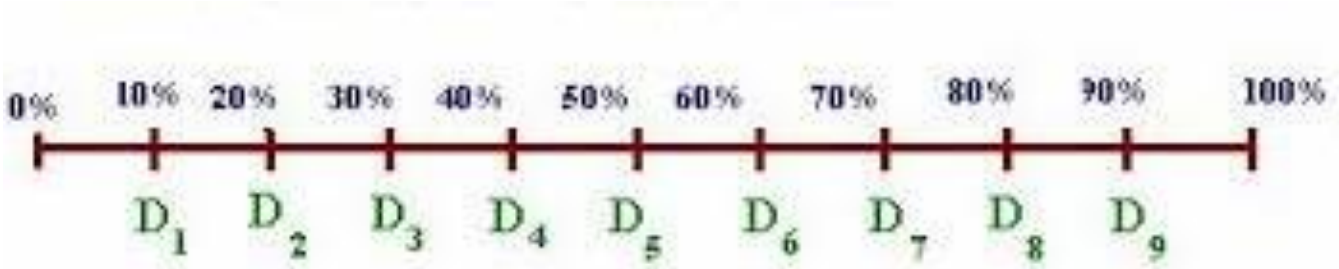


# Cuartil


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

# Decil

- Dividen la distribución de frecuencias en diez partes iguales.

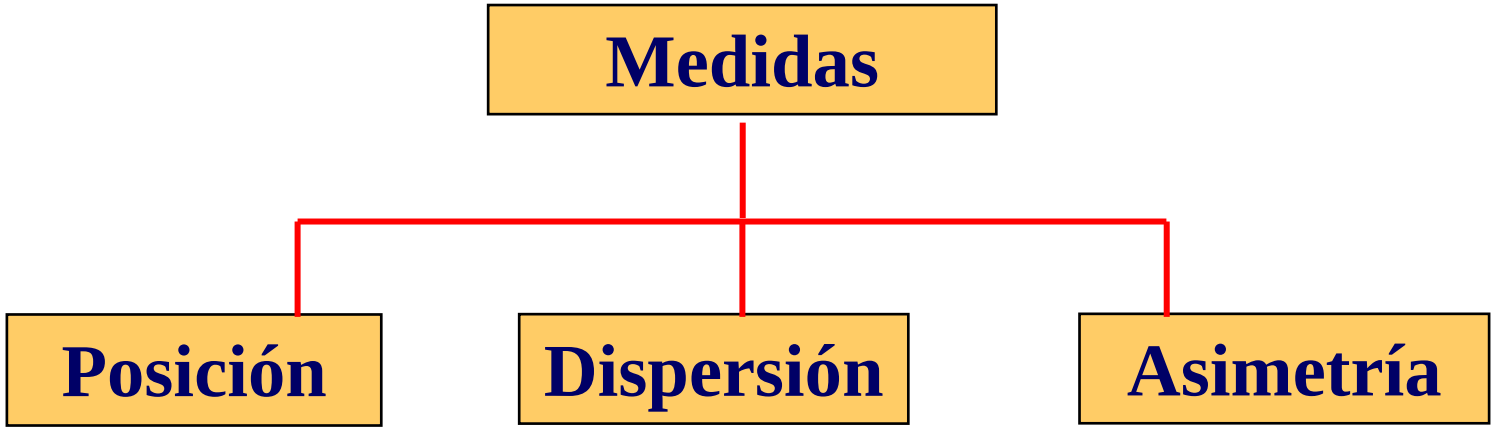



Representación de los deciles

# Percentiles

- Dividen la distribución de frecuencias en 100 partes iguales.

# Análisis Univariado



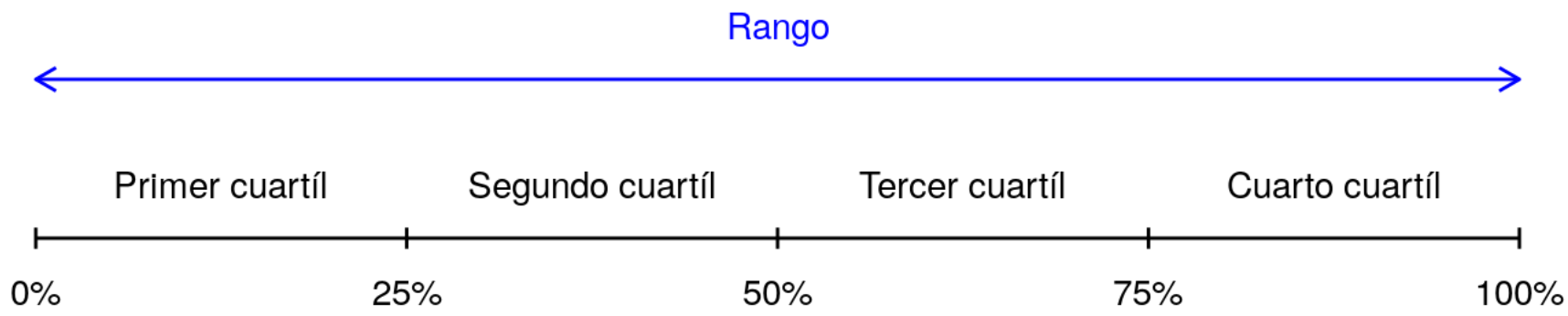

- La media
- La mediana
- La moda
- Percentiles

- El rango
- El rango intercuartílico
- La varianza
- La desviación estándar
- El coeficiente de variabilidad

- El coeficiente de asimetría
- El coeficiente de curtosis

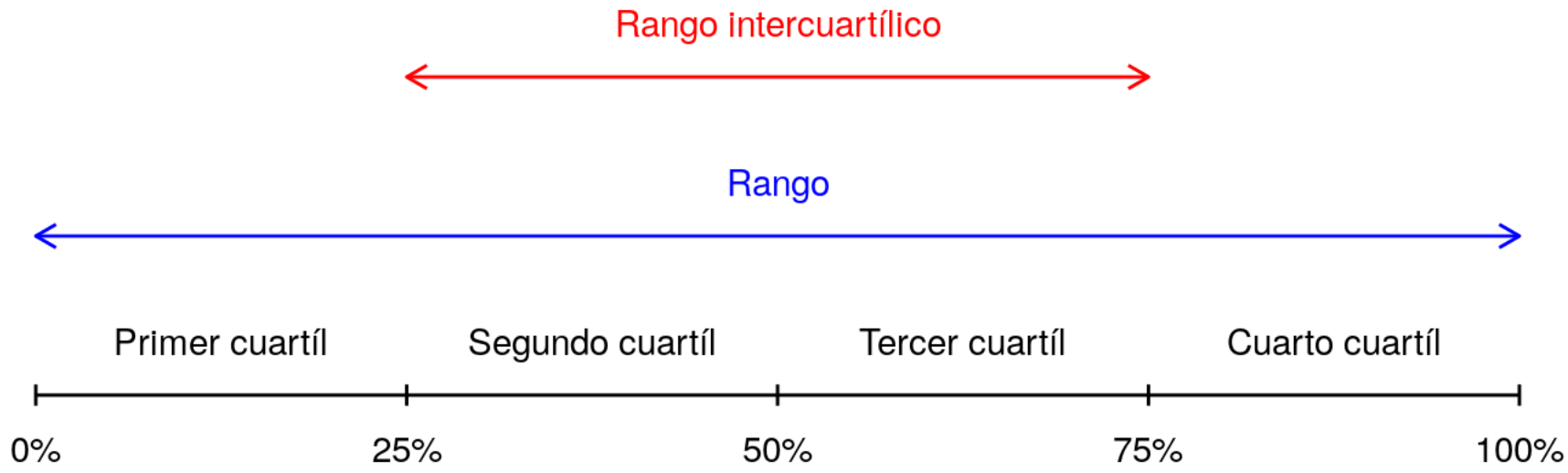
# Rango

- Se calcula como la diferencia entre el valor máximo y mínimo de una distribución de datos.



# Rango Intercuartílico

- Se calcula como la diferencia entre el cuartil 3 y cuartil 1 de una distribución de datos.



# Varianza y Desviación Estándar

| | Varianza | Desviación Estándar |
|---|---|---|
| Población | $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ | $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$ |
| Muestra | $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ | $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ |

# Medidas de dispersión

Distribución A

$\bar{x} = 10$

$s = 2$

Distribución B

$\bar{x} = 100$

$s = 5$

**¿Cuál de las dos tiene menor dispersión?**

# Coeficiente de variación

Distribución A

$$CV = \frac{2}{10} \times 100 = 20\%$$

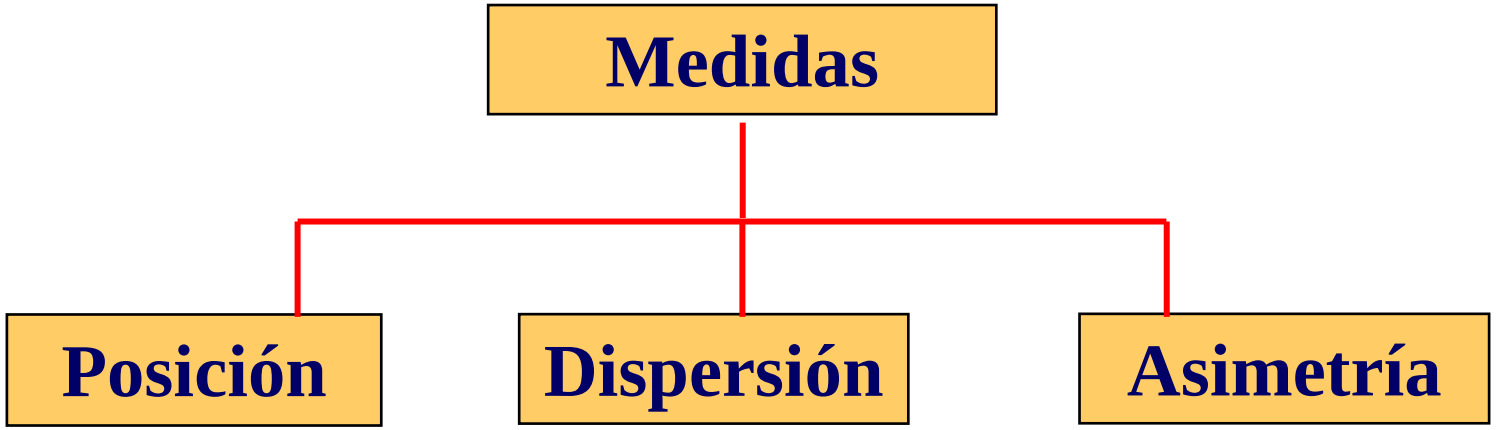
Distribución B

$$CV = \frac{5}{100} \times 100 = 5\%$$

**La distribución B tiene menor dispersión**

$$CV = \frac{\sigma_x}{|\bar{X}|}$$

# Análisis Univariado




- La media
- La mediana
- La moda
- Percentiles

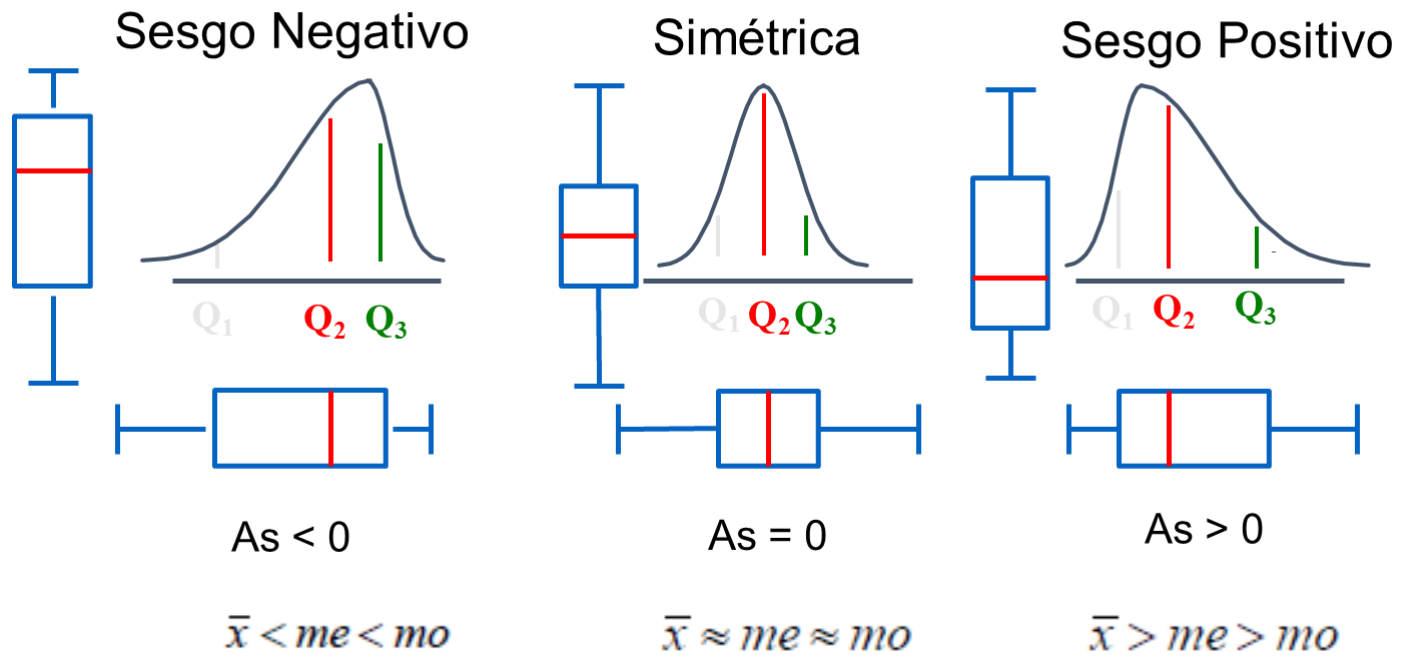
- El rango
- El rango intercuartílico
- La varianza
- La desviación estándar
- El coeficiente de variabilidad

- El coeficiente de asimetría
- El coeficiente de curtosis

# Coeficiente de Asimetría

Si el conjunto de observaciones es la población: $As = \frac{3(\mu - Me)}{\sigma}$

Si el conjunto de observaciones es una muestra: $as = \frac{3(\bar{x} - me)}{S}$



# Coeficiente de Curtosis

$$k = \frac{\frac{1}{2}(q_3 - q_1)}{(d_9 - d_1)}$$

Distribución Platicúrtica

$K < 0.25$

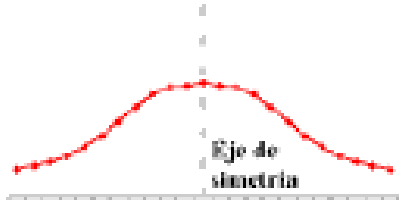
Distribución Mesocúrtica

$K \approx 0.25$

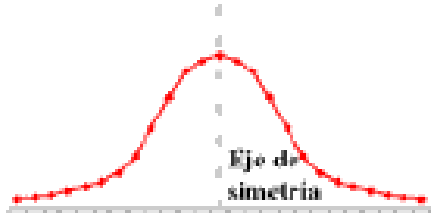
Distribución Leptocúrtica

$K > 0.25$

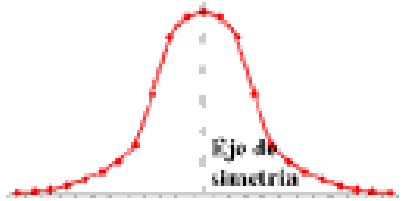
CURVA PLATICURTICA




CURVA MESOCURTICA




CURVA LEPTOCURTICA



# Contenido 2

- Datos, instancias y atributos
- Análisis univariado de datos
- Análisis bivariado de datos
- Visualización de datos
- Caso de Aplicación

# Prueba de normalidad

- Ho: La variable se aproxima a una distribución normal
- Ha: La variable no se aproxima a una distribución normal

Si el pvalor < 5% => se rechaza la Ho, al 95% de confianza se concluye que la variable no se aproxima a una distribución normal

Si el pvalor > 5% => no se rechaza la Ho, al 95% de confianza se concluye que la variable se aproxima a una distribución normal

α: Nivel de significación, por lo general toma el valor de 5%

1-α: Nivel de confianza, por lo general toma el valor de 95%

**Se usa para variables cuantitativas**.

# Correlación (r)

- Se usa para variables cuantitativas.
- Trata de establecer la relación o dependencia entre dos variables.
- Varía entre -1 y 1.
- Teniendo una relación directa al tratarse de 1 (cuando una variable aumenta, la otra también), mientras que existirá una relación inversa al tratarse de -1 (cuando una variable aumenta la otra disminuye).
- Mientras que, Si r = 0 (o cercano a este valor) no existe relación lineal, aunque puede existir algún otro tipo de relación no lineal.

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

# Correlación Pearson

- Se usa para datos que se aproximan a una distribución normal
- Es más sensible a los valores extremos

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

# Correlación Spearman

- Es un método no paramétrico.
- Se aplica cuando no se cumple el supuesto de normalidad.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

Siendo $d_i$ la distancia entre los rangos de cada observación $(x_i - y_i)$ y $n$ el número de observaciones.

# Correlación Kendall

- Es un método no paramétrico.
- Se aplica cuando no se cumple el supuesto de normalidad.
- Útil cuando se tienen pocos datos.

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

siendo $C$ el número de pares concordantes, aquellos en los que el rango de la segunda variable es mayor que el rango de la primera variable. $D$ el número de pares discordantes, cuando el rango de la segunda es igual o menor que el rango de la primera variable.

# Prueba de correlación

- Ho: no existe correlación entre las variables
- Ha: existe correlación entre las variables

Si el pvalor < 5% => se rechaza la Ho, al 95% de confianza se concluye que existe correlación entre las variables

Si el pvalor > 5% => no se rechaza la Ho, al 95% de confianza se concluye que no existe correlación entre las variables

α: Nivel de significación, por lo general toma el valor de 5%

1-α: Nivel de confianza, por lo general toma el valor de 95%

**Se usa para variables cuantitativas**.

# Prueba Chi-Cuadrado

- Ho: no existe dependencia entre las variables
- Ha: existe dependencia entre las variables

Si el pvalor < 5% => se rechaza la Ho, al 95% de confianza se concluye que existe dependencia entre las variables

Si el pvalor > 5% => no se rechaza la Ho, al 95% de confianza se concluye que no existe dependencia entre las variables

α: Nivel de significación, por lo general toma el valor de 5%

1-α: Nivel de confianza, por lo general toma el valor de 95%

**Se usa para variables cualitativas**.

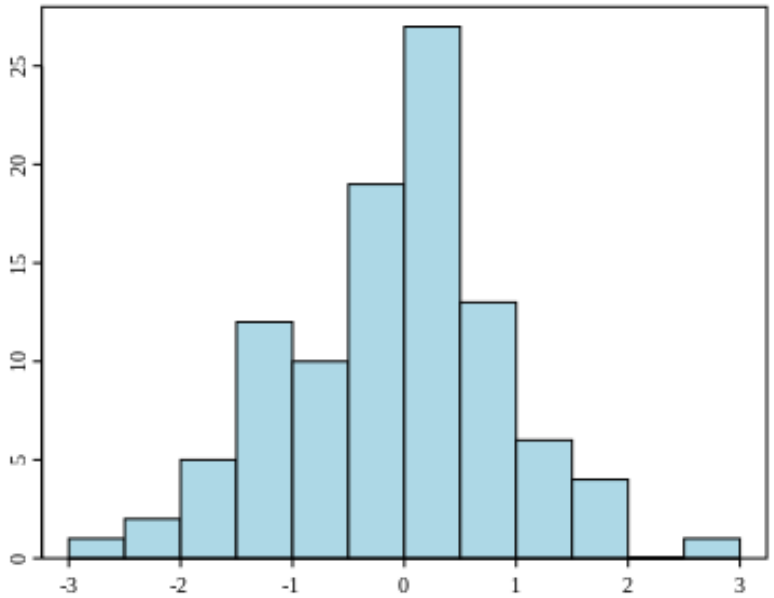
# Contenido 2

- Datos, instancias y atributos
- Análisis univariado de datos
- Análisis bivariado de datos
- Visualización de datos
- Caso de Aplicación

# Visualización de datos

**Histograma**

Usado para variables cuantitativas.
Los intervalos son de igual tamaño.




Número de Intervalos (Regla de Sturgess):

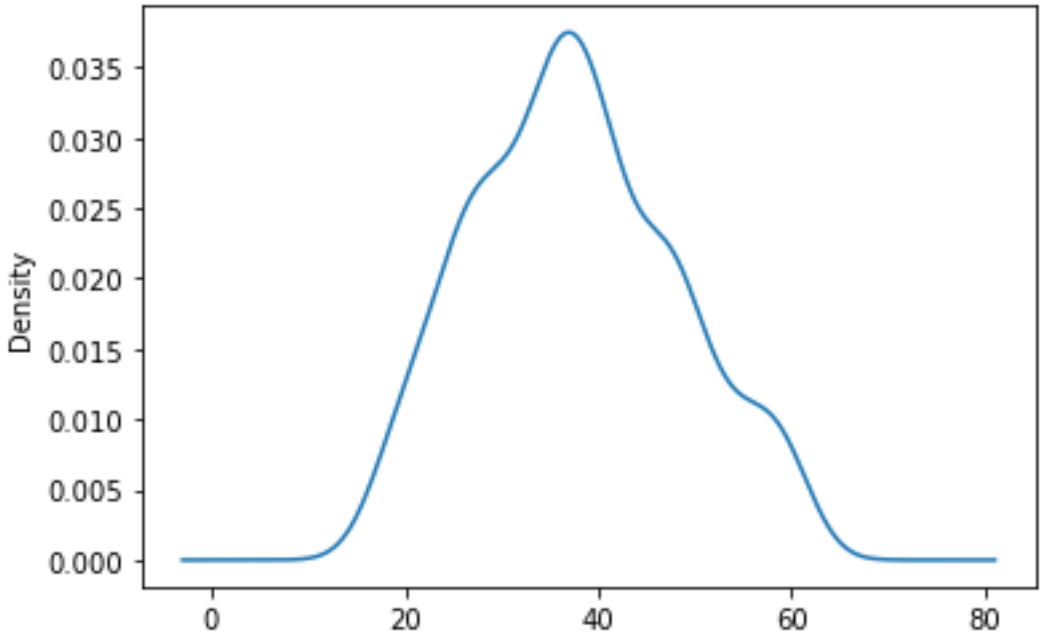
$$c = 1 + 3.322 \cdot \log N$$

Donde N es el tamaño de la muestra

# Visualización de datos

**Densidad**

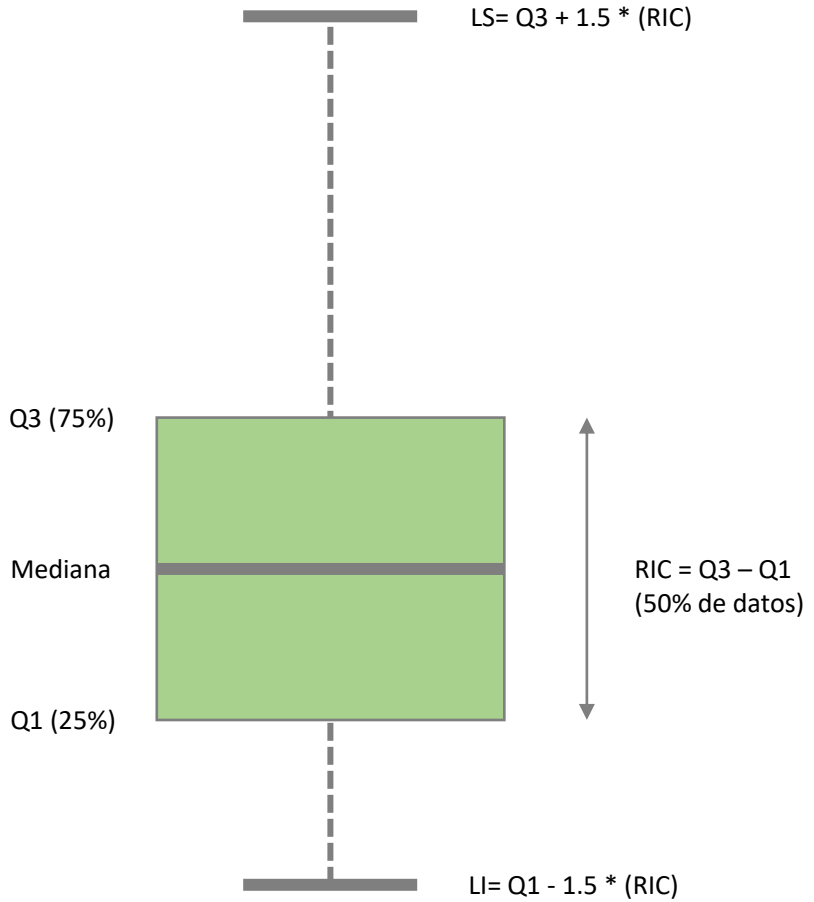
Usado para variables cuantitativas.
Se visualiza la distribución de datos.



# Visualización de datos

**Boxplot**

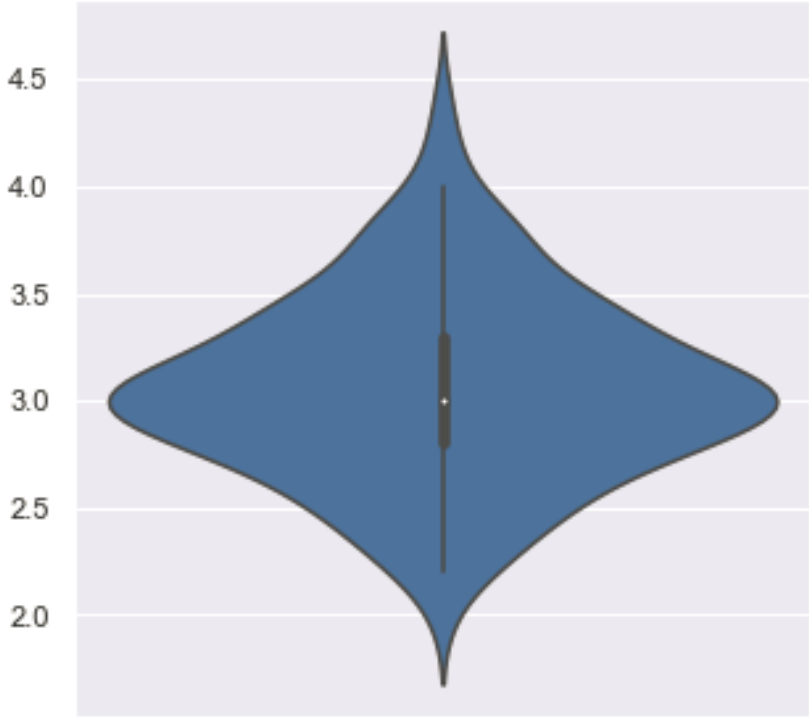
Usado para variables cuantitativas.



# Visualización de datos

**Violin Plot**

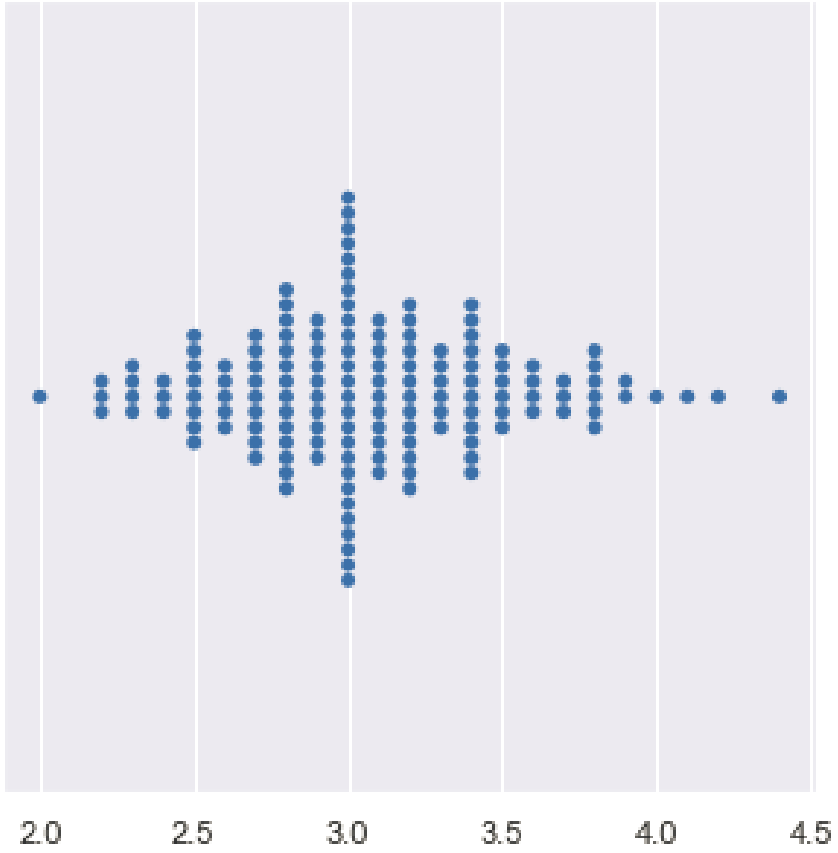
Usado para variables cuantitativas.
Usado para ver la dispersión de valores.
Similar al Boxplot.



# Visualización de datos

**Swarm Plot**

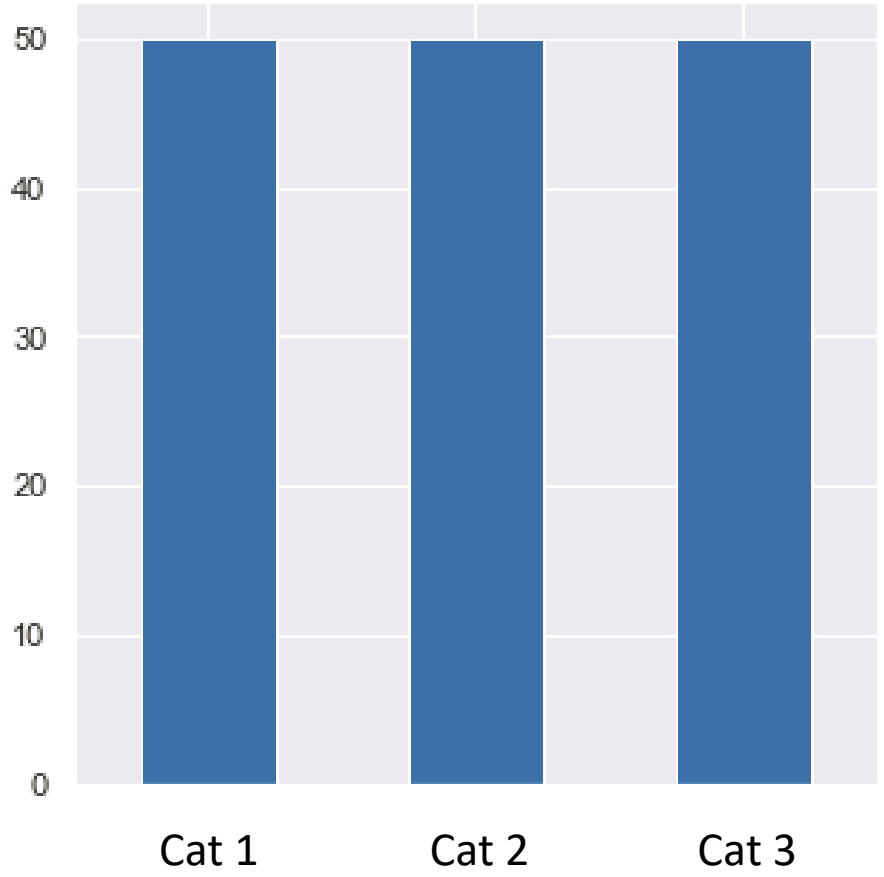
Usado para variables cuantitativas.
Usado para ver la dispersión de valores.



# Visualización de datos

**Gráfico de Barras**

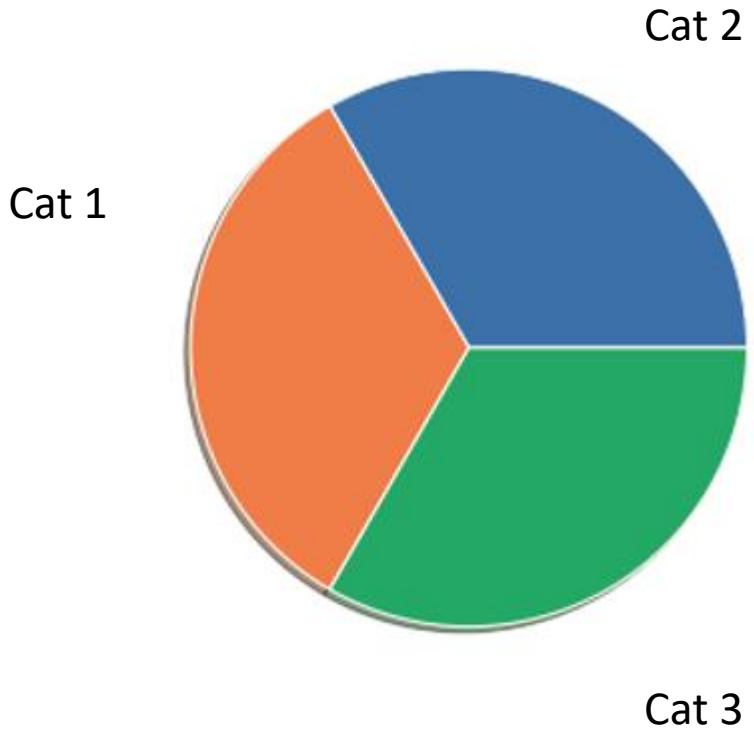
Se usa para variables cualitativas
Para la visualización de datos univariante en un eje bidimensional
El eje X indica la categoría
El eje Y indica el valor numérico de cada categoría, indicado por la longitud de la barra.



# Visualización de datos

**Gráfico de Pie**

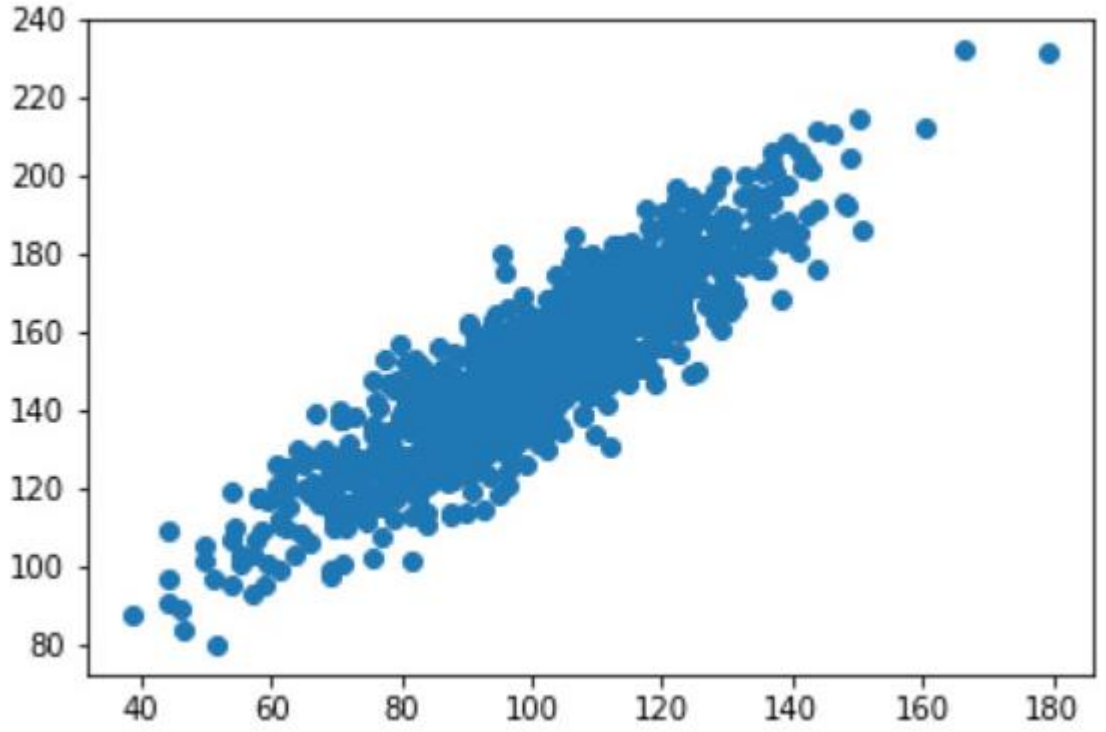
Se usa para variables cualitativas
Para la visualización de datos univariante de la proporción numérica ocupada por cada una de las categorías



# Visualización de datos

**Gráfico de Dispersión**

Usado para analizar 2 variables cuantitativas.
Útil para un análisis previo en un modelo de regresión lineal.



# CIERRE

Medida de posición útil en presencia de valores atípicos.

Medida de dispersión útil para comparar entre variables.

¿Qué es una matriz de datos?

# CONSULTAS

pcsirife@upc.edu.pe