# Estadística descriptiva

Jessica Nathaly Pulzara Mora
jessica.pulzara@udea.edu.co

Departamento de ingeniería de sistemas

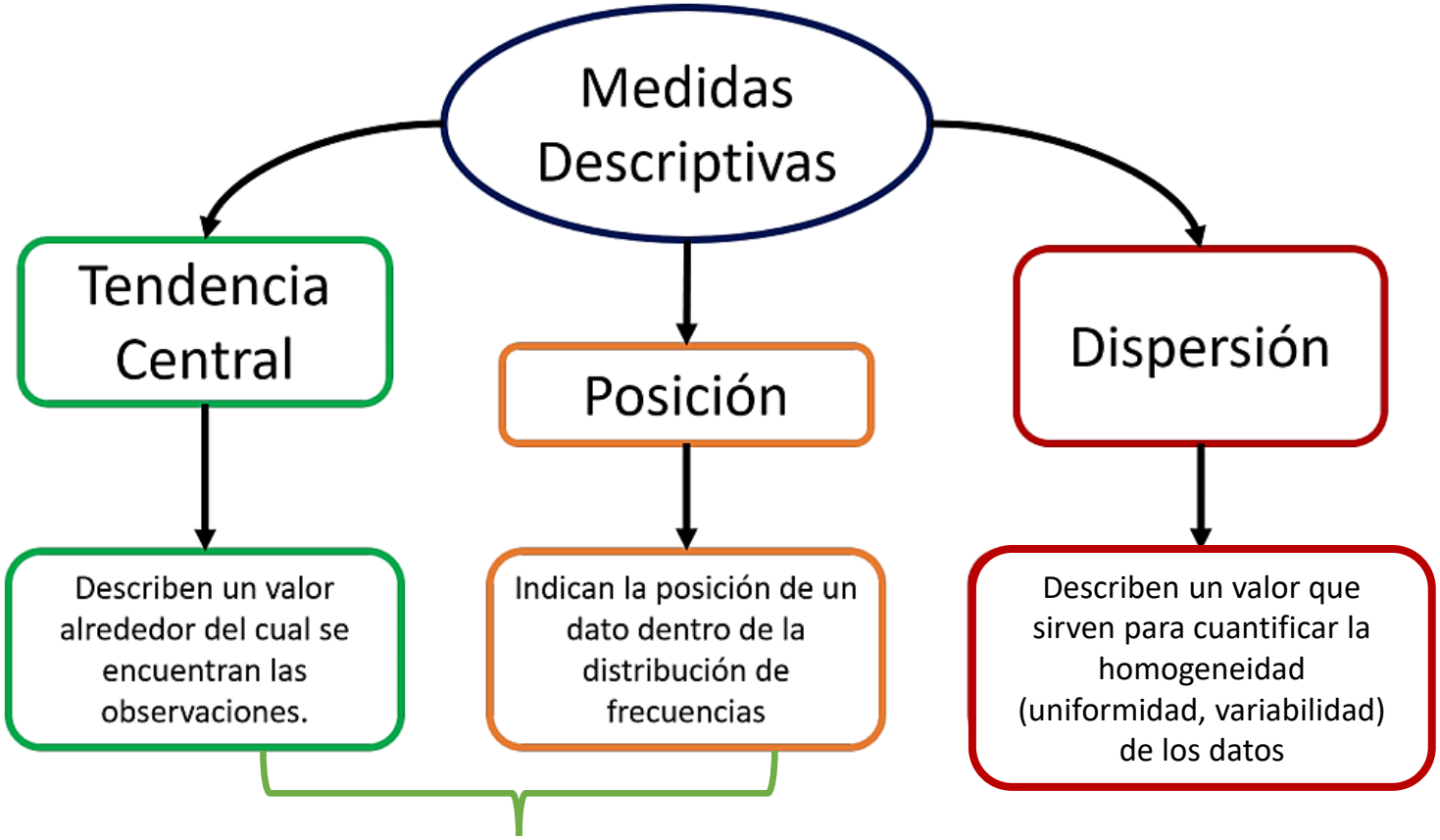
UNIVERSIDAD
DE ANTIOQUIA

En esta sección introduciremos distintas formas de resumir la distribución muestral o poblacional de una variable NUMÉRICA y finalmente presentaremos un tipo de gráfico que se construye a partir de medidas resúmenes.

Las medidas resúmenes son útiles para comparar conjuntos de datos cuantitativos

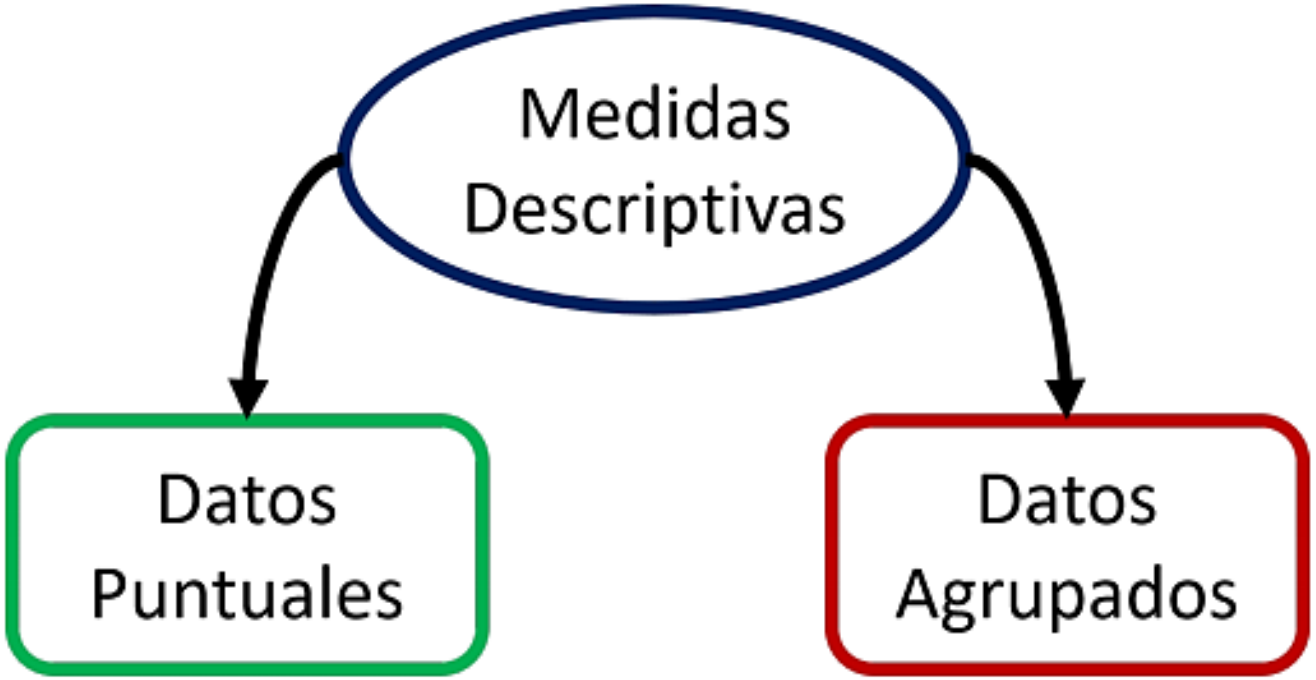
# Medidas descriptivas

Las medidas se clasifican en dos grupos principales:




## Medidas de localización

Además, pueden calcularse de dos maneras:



# Medidas de Tendencia Central
## Datos puntuales

# Media

- Es la medida de posición más frecuentemente usada.
- También le llaman **Promedio**.
- Para calcularla se suman todos los valores y se divide por el número total de observaciones.

Si tenemos una muestra de $n$, observaciones, denotadas por $x_1, x_2, ..., x_n$, la media se calcula así:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

# Ejemplo

Considere los ingresos mensuales en dólares de 8 empleados públicos:

500, 750, 600, 550, 700, 650, 550, 550

Calcule el ingreso mensual medio.

$$\bar{X} = \frac{500 + 750 + \cdots + 650 + 550 + 550}{8} = 606.25$$

# Ejemplo

La media es **sensible a los valores extremos**. Ahora, considere los siguientes ingresos:

500, 750, 600, 550, 700, 2000, 550, 550

Calcule el ingreso mensual medio.

$$\bar{X} = \frac{500 + 750 + \cdots + 2000 + 550 + 550}{8} = 775$$

# Mediana

La mediana es el dato que ocupa la posición central en la muestra ordenada de menor a mayor.

**¿Cómo calculamos la mediana de una muestra de $n$ observaciones?**

- Ordenamos los datos de menor a mayor.
- Determinar la mediana:

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

# Mediana

Explicación de la fórmula:

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

Si el número de datos es impar, la mediana que ocupa la posición $\frac{n+1}{2}$. Si el número de datos es par, la mediana es el promedio de los dos datos centrales.

# Ejemplo

Considere los ingresos mensuales en dólares de 8 empleados Públicos:

$$500,\ 750,\ 600,\ 550,\ 700,\ 2000,\ 550,\ 550$$

Calcule la mediana.

La muestra ordenada es: 500, 550, 550, 550, 600, 700, 750, 2000.
Como $n = 8$ es par, entonces

$$\tilde{X} = \frac{X_{8/2} + X_{\frac{8}{2}+1}}{2} = \frac{550 + 600}{2} = 575$$

# Moda

Es el dato que ocurre con mayor frecuencia en el conjunto. Es una medida de poca utilidad salvo para datos categóricos en los que suele interesar identificar la categoría con mayor cantidad de datos.

## Ejemplo

Encuentre la moda de las edades de los siete empleados.

**53 32 61 57 39 44 57**

La moda es 57 porque ocurre la mayoría de las veces.

# Medidas de Posición
## Datos puntuales

# Percentiles

- Indican el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo.
- Los datos deben estar ordenados de forma ascendente.
- Por ejemplo, el percentil 33 es el valor bajo el cual se encuentran el 33 por ciento de las observaciones.

# Percentiles

**Algorítmo de cálculo:**

Proceda así para calcular el percentil $P_i$ $(1 \leq i \leq 100)$

1. Ordene la muestra de menor a mayor.

2. Calcule la posición:

$$p = \frac{ni}{100}$$

3. Halle el percentil correspondiente así:

$$P_i = \begin{cases} \frac{x_p + x_{(p+1)}}{2} & \text{si } p \text{ es entero,} \\ x_{(||p||+1)} & \text{si } p \text{ es decimal.} \end{cases}$$

# Ejemplo

Considere los siguientes datos ordenados de menor a mayor:

$$900,\ 950,\ 500,\ 550,\ 700,\ 750,\ 750,\ 800,\ 550,\ 600$$

Calcule el percentil 76.

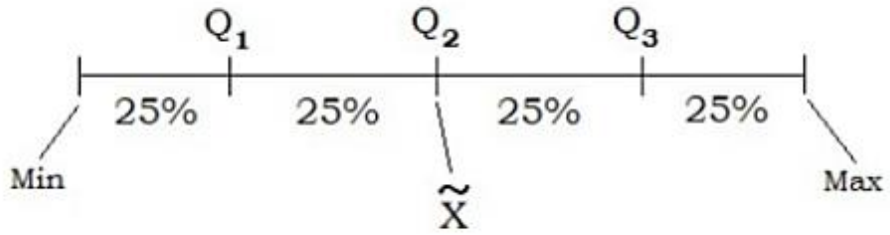
La muestra ordenada es: 500, 550, 550, 600, 700, 750, 750, 800, 900, 950.
Como $n = 10$, $i = 76$, entonces

$$p = \frac{10 \cdot 76}{100} = 7.6 \quad \text{(es decimal)}$$

$$P_{76} = x_{(||7.6||+1)} = x_8 = 800$$

**Equivalencias importantes:**

Representemos el cuartil $i$ con el símbolo $Q_i$, con $i = 1, 2, 3, 4$.




Entonces:

- $P_{25} = Q_1, \qquad P_{75} = Q_3$
- $P_{50} = Q_2 = \tilde{X}$

# Medidas de Dispersión (o variabilidad)
# Datos puntuales

Nos dicen cuán disperso es el conjunto de datos. Consideremos los siguientes conjuntos de datos:

Muestra A: 55 55 55 55 55 55 55
Muestra B: 47 51 53 55 57 59 63
Muestra C: 39 47 53 55 57 63 71

En todos ellos $\bar{X} = \tilde{X} = 55$, pero los datos son diferentes en cada caso. Las medidas de dispersión nos dirán:

- Cuán cercanos se encuentran los datos entre ellos.
- Cuán cercanos se encuentran a una medida de posición/tendencia central.

# Rango intercuantil

Esta medida es la diferencia entre el percentil 75 y el 25. Mide que tan disperso está el 50 % de los datos más centrales.

$$RIQ = Q_3 - Q_1$$

# Ejemplo

Consideremos los datos anteriormente revisados para calcular el $RIQ$:

500, 550, 550, 600, 700, 750, 750, 800, 900, 950.

$$Q_3 = P_{75} = x_{(||7.5||+1)} = x_8 = 800$$
$$Q_1 = P_{25} = x_{(||2.5||+1)} = x_3 = 550$$

$$RIQ = 800 - 550 = 250$$

# Varianza y desviación estándar

Consideremos una muestra de observaciones $X_1, X_2, ...X_n$ con media $\bar{X}$. Su *varianza* se calcula así:

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

$$= \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n-1}$$

- Puede interpretarse como “promedio cuadrátio” de las distancias de los datos a la media.

- No tiene las mismas unidades de los datos. Por ejemplo, si los datos son estaturas en metros ($m$), la varianza está en $m^2$.

- La desviación estándar si tiene las mismas unidades de los datos, porque es la raiz cuadrada de la varianza:

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

# Ejemplo

Calculemos la varianza y la desviación estándar de las muestras antes consideradas:

- Muestra A: 55 55 55 55 55 55 55
- Muestra B: 47 51 53 55 57 59 63
- Muestra C: 39 47 53 55 57 63 71

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

# Ejemplo

Calculemos la varianza y la desviación estándar de las muestras antes consideradas:

- Muestra A: 55 55 55 55 55 55 55 $\quad S_A^2 = 0, \quad S_A = 0$
- Muestra B: 47 51 53 55 57 59 63 $\quad S_B^2 = 28, \quad S_B = 5.29$
- Muestra C: 39 47 53 55 57 63 71 $\quad S_C^2 = 108, \quad S_C = 10.39$

# Desviación estándar

Interpretación:

- $S$ es útil para comparar la variabilidad de dos conjuntos de datos en los que la variable ha sido medida en las mismas unidades. Por ejemplo, si $S_1 = 5.4$ y $S_2 = 10.4$, entonces $S_2$ está más disperso.

- La desviación estándar nos da idea de la distancia promedio de los datos a la media (aunque estrictamente hablando no es el promedio).

# Coeficiente de variación

Es una medida de dispersión relativa.
El coeficiente de variación que se define como

$$CV = \frac{S}{\bar{X}} \times 100\,\%$$

Es una fracción de la media muestral. Se usa para comparar la variabilidad de dos o más conjuntos de datos.

# Ejemplo


1000 ml pack

$s = 15\ ml$

$\bar{x} = 1005\ ml$

$$CV = \frac{s}{\bar{x}}100\% = \frac{15\ ml}{1005\ ml}100\% = 1.49\%$$


50 ml pack

$s = 3\ ml$

$\bar{x} = 45\ ml$

$$CV = \frac{s}{\bar{x}}100\% = \frac{3\ ml}{45\ ml}100\% = 6.67\%$$