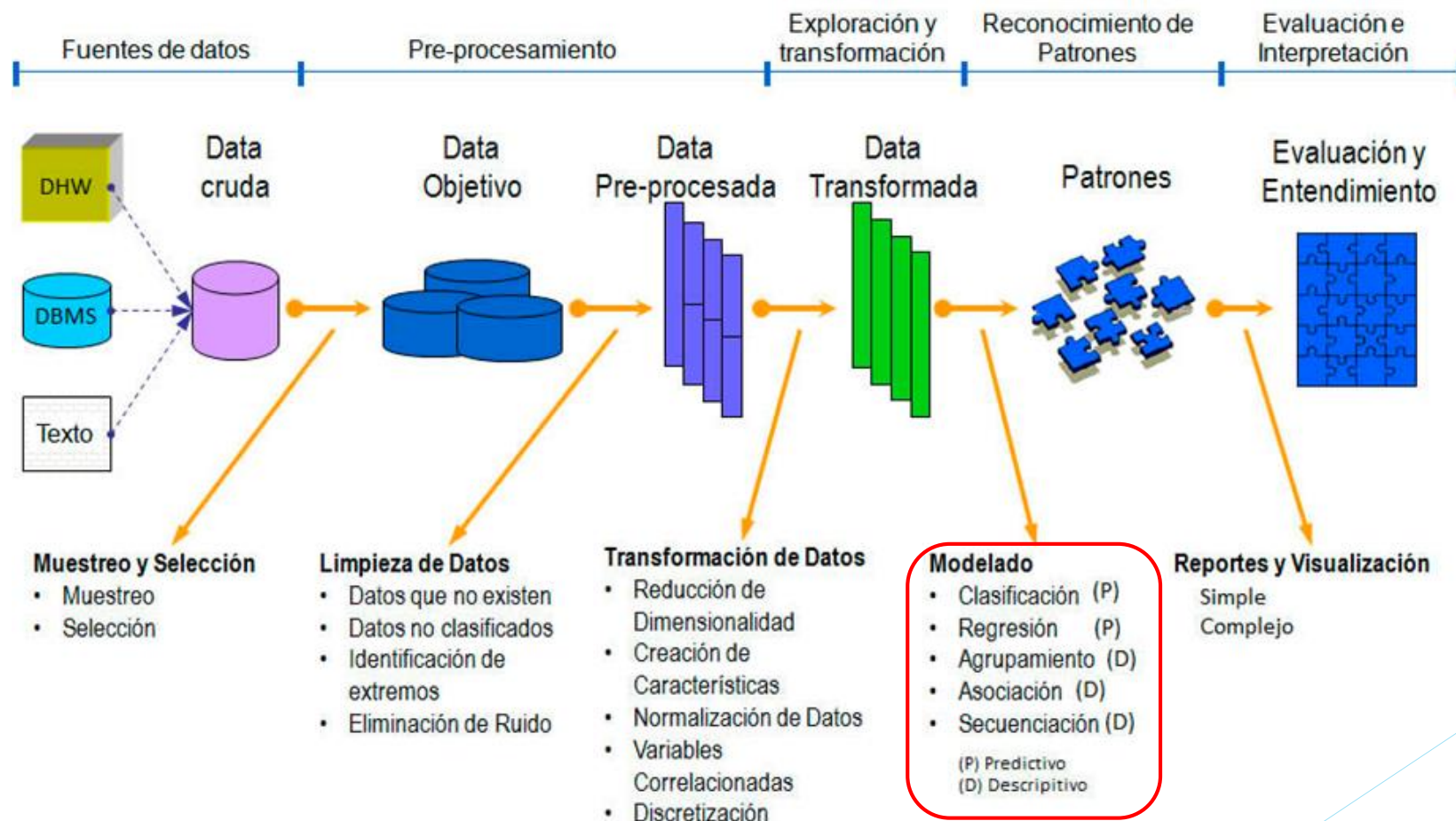


Introduccción al BigData

Ing. Ricardo Velasteguí (M.B.A.)

Análisis de Datos

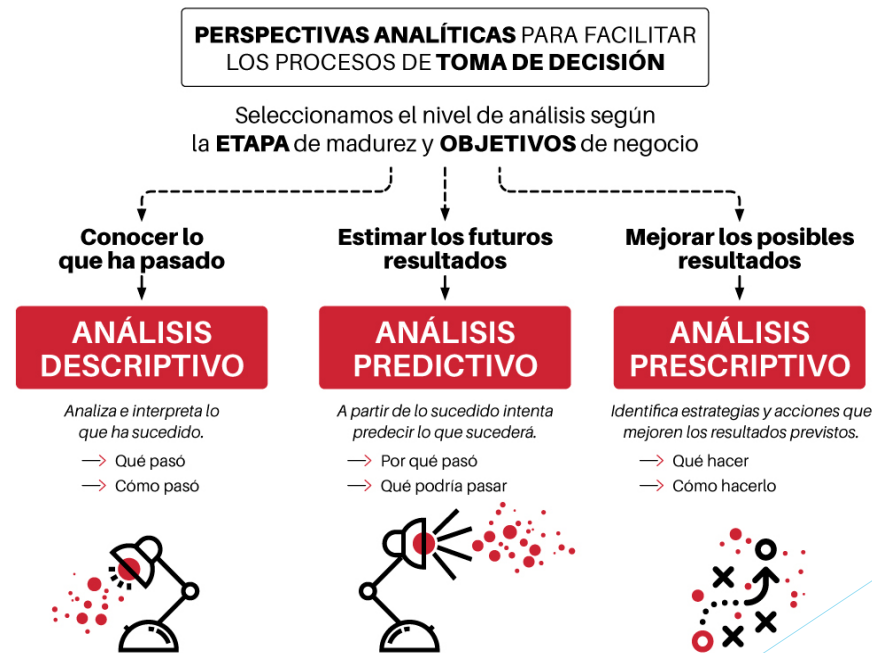
- El análisis de datos es un proceso que nos permite obtener conocimiento de la información inmersa de los datos con el propósito de extraer conclusiones que permitan tomar decisiones informadas.



Técnica de Modelado - BigData

- ▶ En el proceso de examinar grandes cantidades de datos y extraer información para descubrir patrones ocultos, correlaciones desconocidas y otra información útil, lo más importante del análisis de datos es procesar la información de manera eficaz y en un tiempo razonable, de tal manera, que se puedan obtener resultados óptimos.
- ▶ Existen distintos métodos para la Analítica de datos en función de los objetivos a alcanzar y los tipos de datos a utilizar. Estos métodos suelen ser agrupados en tres grandes categorías:
 - ▶ **Análisis Descriptivo.-** Se utiliza a menudo al examinar cualquier dato pasado o presente. Esto se debe a que los datos en bruto son difíciles de consumir e interpretar.
 - ▶ **Análisis Predictivo.-** Es el descubrimiento de información oculta en base a atributos a partir de la creación de modelos de ocurrencia de la mejor probabilidad de un resultado.
 - ▶ **Análisis Prescriptivo.-** Modifica ciertas variables para generar los mejores escenarios posibles de acuerdo al resultado.

Buscar en los datos

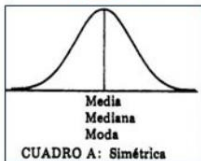


Análisis Descriptivo

- Consiste en describir las tendencias claves en los datos existentes y observar las situaciones que conduzcan a nuevos hechos. Este método se basa en una o varias preguntas de investigación y no tiene una hipótesis. Además, incluye la recopilación de datos relacionados, posteriormente, los organiza, tabula y describe el resultado.

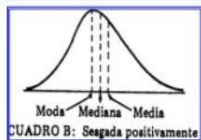
De tendencia central

Promedio (\bar{X}) – Mediana (Me) – Moda (Mo)



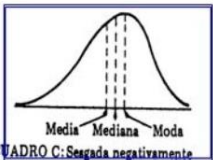
Media
Mediana
Moda

CUADRO A: Simétrica



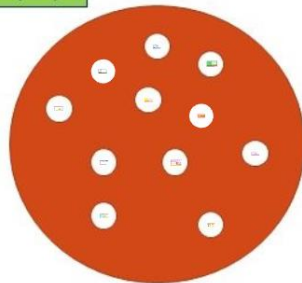
Moda Mediana Media

CUADRO B: Sesgada positivamente



Media Mediana Moda

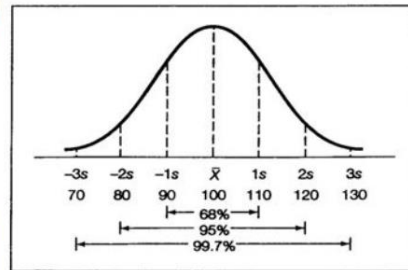
CUADRO C: Sesgada negativamente



Datos de la Muestra

De DISPERSIÓN

Rango (R)
Desviación Media (DM)
Varianza (V_a)
Desv. Estándar (δ)
Coef. de Variación (CV)



De tendencia No central

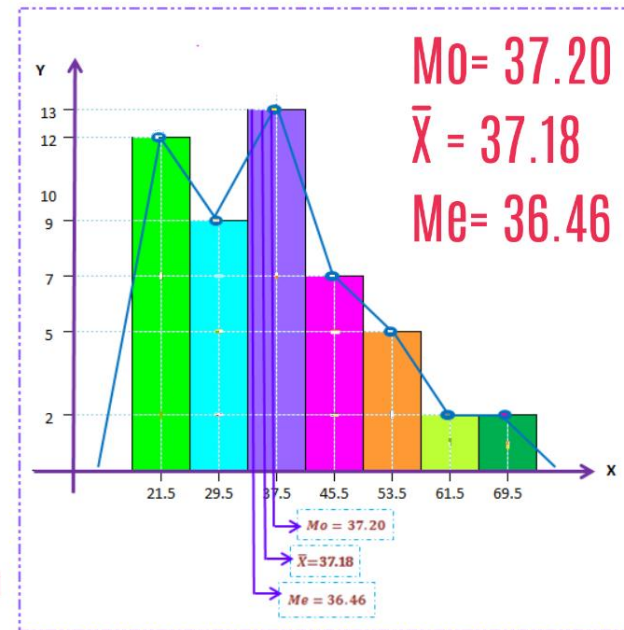
Cuartil (C = 4 partes)
Quintil (Q = 5 partes)
Decil (D = 10 partes)
Percentil (P = 100 partes)

Medidas de posición

✓ Medidas de tendencia central.

✓ Medidas de tendencia no central

✓ Medidas de Dispersión.



Análisis Descriptivo

► Medidas de tendencia central

- Informan sobre los valores medios del conjunto de datos.
- Sirven para describir características básicas de un análisis con datos cuantitativos, muestra promedios, compara resultados, interpreta resultados en relación a un valor central.

Las Medidas de Tendencia Central corresponden a los valores que generalmente se ubican en la parte central de un conjunto de datos.

1 MEDIA O PROMEDIO

Es la suma de todos los datos divididos por el número total de ellos

2 MEDIANA

Es el dato que ocupa la posición central en el conjunto de datos ordenados

3 MODA

Es el valor que más se repite en un conjunto de datos

NOTA: La mayoría del tiempo es necesario denotar un conjunto de datos por un solo valor, que sirva de referencia para interpretar información y pueda representar de la mejor manera a todos los valores del conjunto.

Análisis Descriptivo

► Media o Promedio

- Es la medida que indica el valor promedio de un conjunto de datos.
- Esta medida es sensible a los valores extremos, valores muy grandes aumentan el valor de la media, y valores muy pequeños disminuyen la media.
- Se calcula mediante la siguiente formula:

Ejemplo:

$$\text{Media} = \bar{x} = \frac{\text{Suma de valores}}{\text{Cantidad de valores}}$$

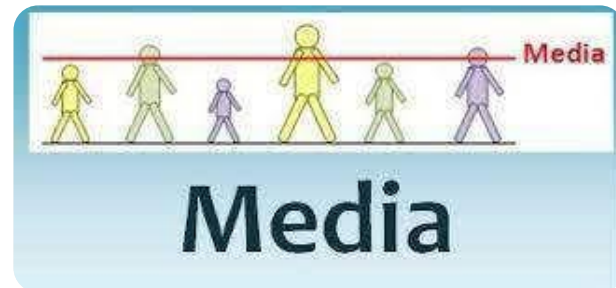
¿Cuál es la media de las edades de Andrea y sus primos?



Así, la media de las edades de Andrea y sus primos se calcula:

$$\text{Media} = \frac{3 + 5 + 6 + 8 + 9 + 9 + 9}{7} = \frac{49}{7} = 7$$

La media de edad es **7 años**.




Análisis Descriptivo

► Mediana


- Es el dato que se localiza en la mitad de un conjunto de datos ordenados, y no se ve afectada por valores extremos.
- Si el conjunto de datos es par, la mediana es el promedio de los dos números centrales.
- Si el conjunto de datos es impar, la mediana es el valor central.

Conjunto par de datos:



8 6 9 5 2 10

Ordenamos los datos de menor a mayor

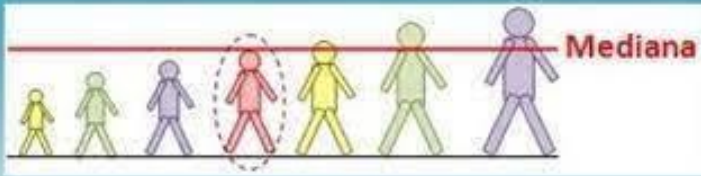


2 5 6 8 9 10

Ahora calculamos la media de los datos centrales:

$$\frac{6 + 8}{2} = \frac{14}{2} = 7$$

La mediana es 7



Mediana

Mediana

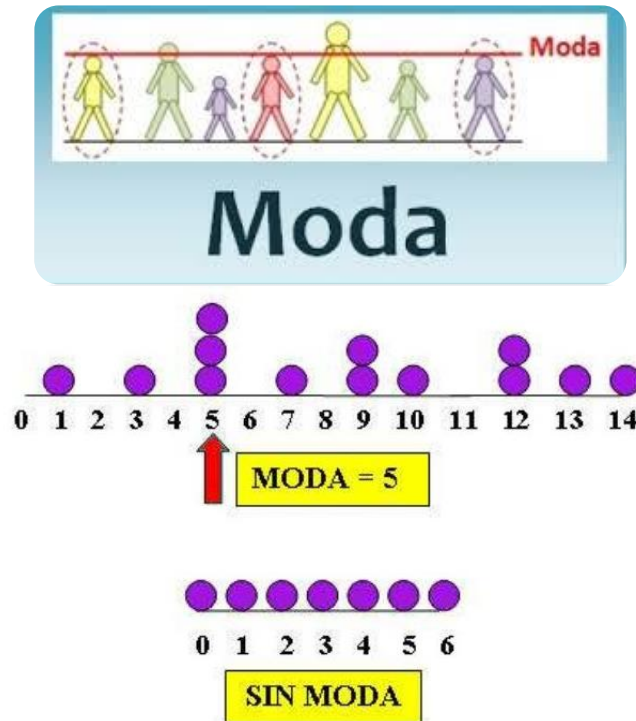
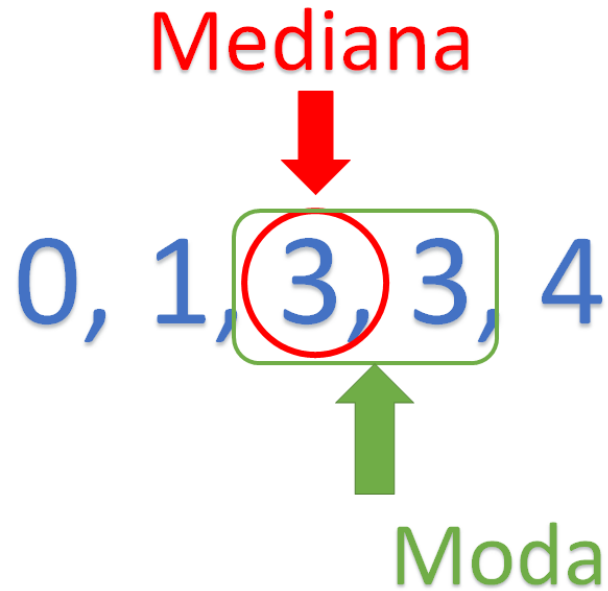
7, 7, 8, 8, 9, 9, 9, 10, 10

$\tilde{x} = 9$
 $Me = 9$

Análisis Descriptivo

► Moda

- Es el valor más repetido en un conjunto de datos, denominado como frecuencia.
- Si no existen datos repetidos, no hay moda.
- Puede existir más de una moda (2 = bimodal, 3 = trimodal, etc.)
- Se puede usar en datos cuantitativos y datos cualitativos.
- No intervienen todos los elementos. Es afectada si existe formación de intervalos.



Análisis Descriptivo

► Medidas de tendencia central

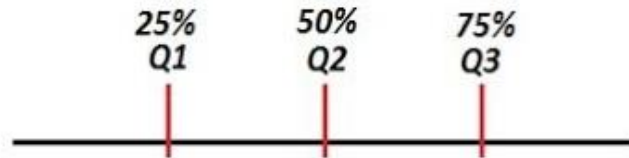
Procedimiento para calcular una **medida de tendencia central**



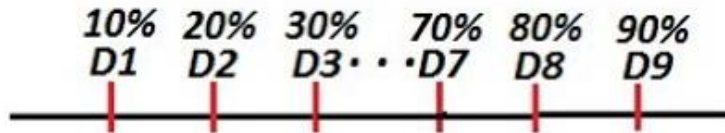
Análisis Descriptivo

► Medidas de tendencia no central (De posición)

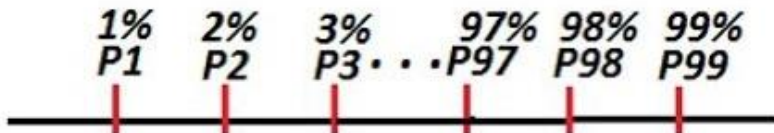
- Informan sobre los valores que no son centrales del conjunto de datos.
- Sirven para describir características de la distribución a través de una serie de valores que dividen el conjunto de datos en tramos iguales.



$\frac{k \cdot n}{4}$ **CUARTIL**



$\frac{k \cdot n}{10}$ **DECIL**

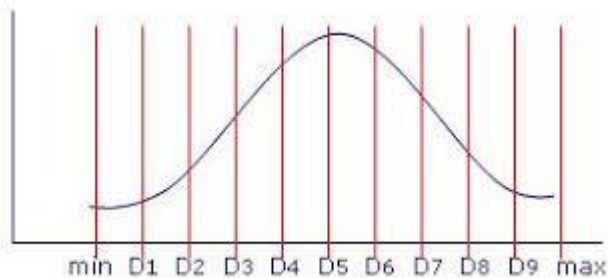
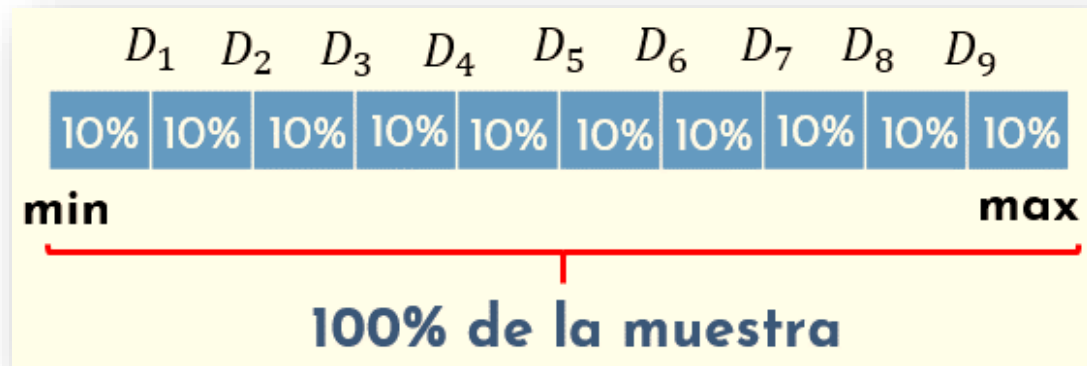


$\frac{k \cdot n}{100}$ **PERCENTIL**

Análisis Descriptivo

► Decil

- Divide el conjunto de datos ordenado en 10 partes iguales de 10%.
- Los deciles son los 9 valores (D_1 , D_2 , D_3 , D_4 , D_5 , D_6 , D_7 , D_8 , D_9) que dividen al conjunto de datos.
- D_5 coincide con la Mediana, cuando el conjunto de datos es simétrico.
- Los deciles corresponden al 10%, 20%... y al 90% de los datos.

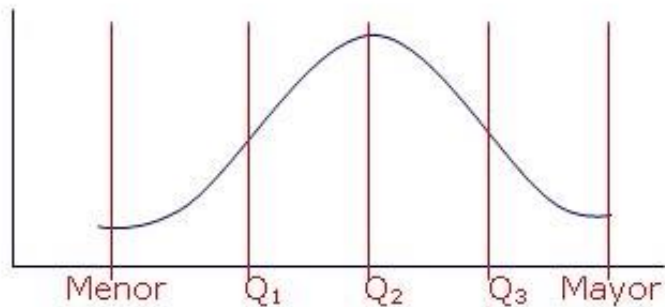
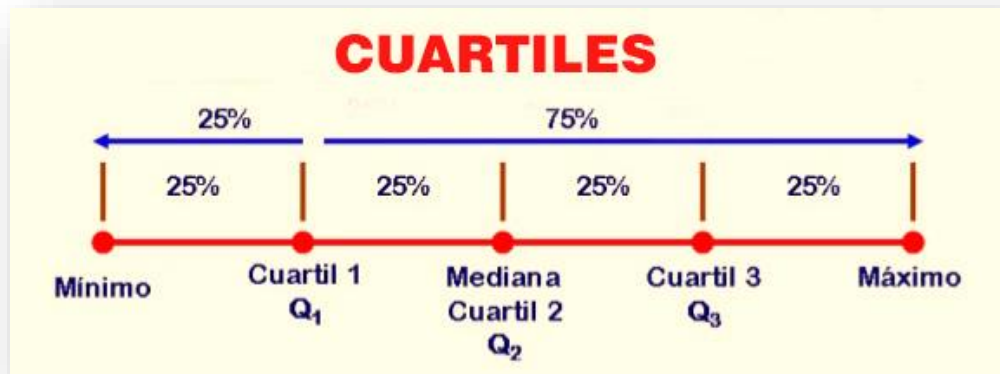


Deciles	Total Ingresos	Total Impuestos Recaudado	Tipo Impositivo Efectivo
Decil 1	2.01	0.84	41.87%
Decil 2	5.77	2.21	38.19%
Decil 3	8.73	0.46	5.23%
Decil 4	13.08	0.69	5.24%
Decil 5	18.37	1.48	8.05%
Decil 6	24.51	1.45	5.92%
Decil 7	34.07	1.96	5.75%
Decil 8	52.07	5.26	10.09%
Decil 9	94.33	6.53	6.93%
Decil 10	348.02	18.90	5.43%
Total	600.98	39.77	6.62%

Análisis Descriptivo

► Cuartil

- Divide el conjunto de datos ordenado en 4 partes iguales de 25%.
- Los cuartiles son los 3 valores (Q1, Q2, Q3) que dividen al conjunto de datos.
- Q2 coincide con la Mediana, cuando el conjunto de datos es simétrico.
- Q1 = 25% (cuartil inferior), Q2 = 50% (cuartil medio), Q3=75% (cuartil superior).



$$Q_1 = \frac{(n+1)}{4} = \frac{(22+1)}{4} = \frac{23}{4} = 5.75 = 6|$$

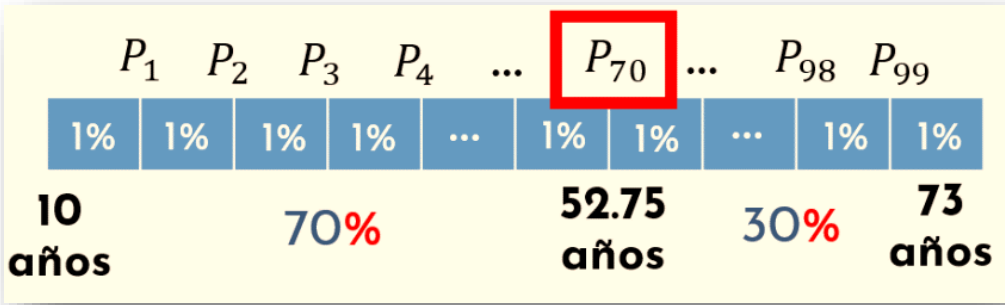
$$Q_2 = \frac{(n+1)}{2} = \frac{(22+1)}{2} = \frac{23}{2} = 11.5$$

$$Q_3 = \frac{3(n+1)}{4} = \frac{3(22+1)}{4} = \frac{69}{4} = 17.25 = 17$$

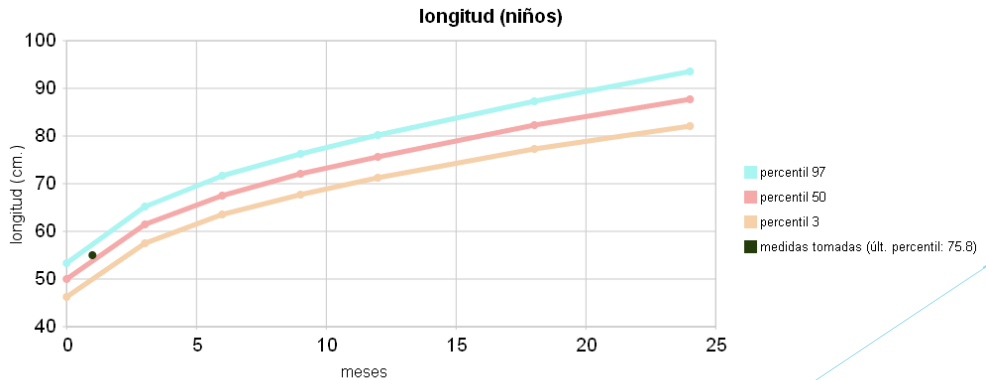
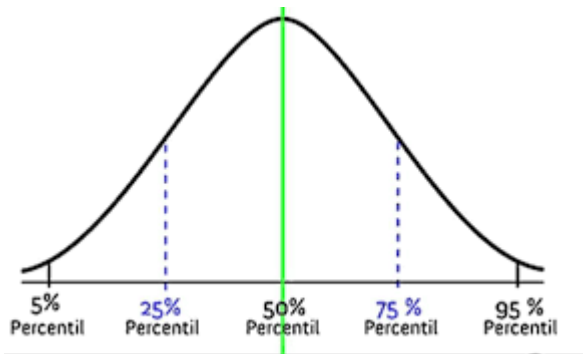
Análisis Descriptivo

► Percentil

- Divide el conjunto de datos ordenado en 100 partes iguales de 1%.
- Los percentiles son los 99 valores (P1, P2, P3, P4, P5, P6, P7, P8, P9..P99) que dividen al conjunto de datos.
- P10 => D1, P25 => Q1, P50 => Q2(Mediana), P75 => Q3, P90 => D9.



	A	B	C
1	Name	Score	Percentile
2	Daniel	66	38%
3	Garry	78	50%
4	Fancy	60	13%
5	Emy	89	75%
6	Wend	78	50%
7	Alice	65	25%
8	Zac	90	88%
9	Peter	92	100%



El niño está en el percentil 75.8 de longitud. Es decir, su longitud es algo superior a la de la mayoría de los niños de su edad.
Cálculos realizados utilizando las tablas de la [Organización Mundial de la Salud](#). Para más información, ver [preguntas frecuentes](#).

Análisis Descriptivo

► Medidas de Dispersión (De variabilidad)

- Informan sobre la variabilidad del conjunto de datos en referencia a la media.
- Sirven para indicar si el conjunto de datos de una variable son homogéneos(concentrados) o heterogéneos(dispersos) analizando su cercanía o alejamiento con respecto a la medida central.

Las Medidas de Dispersión nos indican que tanto se alejan nuestros valores de la parte central de un conjunto de datos.

RANGO

Es la diferencia entre el mayor y el menor de los datos.

DESVIACIÓN MEDIA

Es el promedio de los valores absolutos de las desviaciones respecto a su media.

VARIANZA

Es el promedio del cuadrado de las desviaciones respecto a su media.

DESVIACIÓN TÍPICA

Es la raíz cuadrada de la varianza.

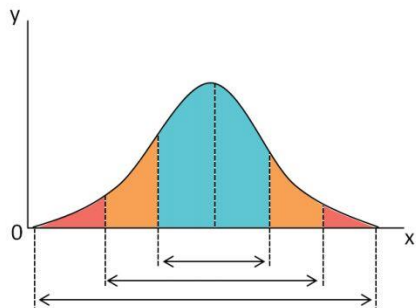
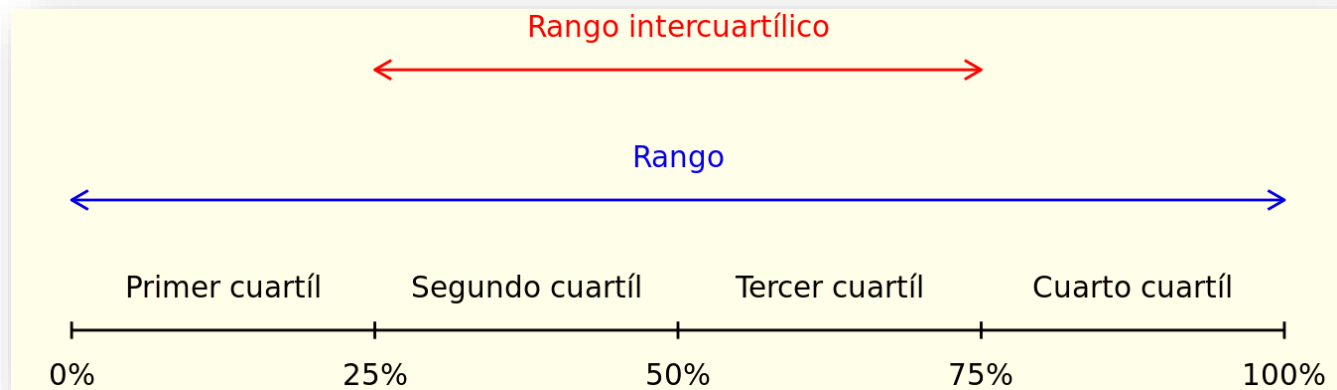
NOTA: Una desviación respecto a la media es la diferencia entre cada dato con su media o promedio.

NOTA: Una desviación respecto a la media es la diferencia entre cada dato con su media o promedio.

Análisis Descriptivo

► Rango o Recorrido

- Es la diferencia entre el mayor valor y el menor valor de un conjunto de datos.
- Suministra información solo sobre los extremos de un conjunto.
- Informa sobre la distancia entre el mínimo y el máximo valor observado.
- Se limita su utilización como parte de un análisis inicial.



- El rango intercuartilico (IQR o rango intercuartil), se produce entre el tercer y primer cuartil, mediante esta medida se elimina los valores extremos.
- $IRQ = Q3 - Q1$.

Análisis Descriptivo

► Desviación Media

- Es el promedio de los valores absolutos de las desviaciones con respecto a la media (la diferencia entre las desviaciones y el valor promedio del conjunto de datos).

Calificaciones	Desviaciones
8	$ 8 - 8.4 = 0.4$
10	$ 10 - 8.4 = 1.6$
9	$ 9 - 8.4 = 0.6$
8	$ 8 - 8.4 = 0.4$
7	$ 8 - 8.4 = 1.4$

El promedio del valor absoluto de las desviaciones.

$$DM = \frac{0.4 + 1.6 + 0.6 + 0.4 + 1.4}{5} = \frac{4.4}{5} = 0.88$$

DESVIACIÓN MEDIA

$$DM = \frac{\sum |X_i - \bar{X}|}{n}$$

- La media es 8.4.
- Se calcula todas las diferencias y se obtiene el promedio.

Análisis Descriptivo

Calcule el rango y la desviación media de los siguientes datos:

4, 6, 1, 3, 10, 7, 9 y 3

PARA EL RANGO

4, 6, 1, 3, 10, 7, 9, 3

*Tomamos el número mayor y le restamos el número menor

$$10 - 1 = 9$$

PARA LA DESVIACIÓN MEDIA

*Se obtiene el promedio de los datos:

$$\bar{x} = \frac{4 + 6 + 1 + 3 + 10 + 7 + 9 + 3}{8} = 5.3$$

*Se obtiene el promedio de los valores absolutos de las desviaciones:

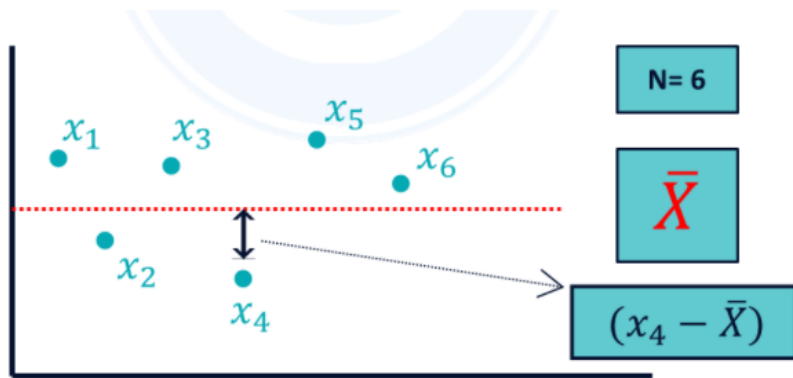
$$\begin{aligned} &|4 - 5.3| + |6 - 5.3| + |1 - 5.3| + |3 - 5.3| \\ &+ |10 - 5.3| + |7 - 5.3| + |9 - 5.3| \\ &+ |3 - 5.3| = 21 \end{aligned}$$

$$D\bar{x} = \frac{21}{8} = 2.62$$

Análisis Descriptivo

► Varianza

- Es el promedio de los cuadrados de los valores absolutos de las desviaciones con respecto a la media (**cuadrados de la diferencia entre las desviaciones y el valor promedio del conjunto de datos**).
- Es la medida de variabilidad que da cuenta del grado de homogeneidad del conjunto de datos.



$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- **Calcular la varianza** de la distribución:
- 9, 3, 8, 8, 9, 8, 9, 18

$$\bar{X} = \frac{9+3+8+8+9+8+9+18}{8} = 9$$

$$\sigma^2 = \frac{(9-9)^2 + (3-9)^2 + (8-9)^2 + (8-9)^2 + (9-9)^2 + (8-9)^2 + (9-9)^2 + (18-9)^2}{8} = 15$$

- La media es 9
- Se eleva al cuadrado cada una de las diferencias y se obtiene el promedio.

Análisis Descriptivo

► Desviación típica (estándar)

- Es la raíz cuadrada del promedio del cuadrado de los valores absolutos de las desviaciones con respecto a la media (**raíz cuadrada de los cuadrados de la diferencia entre las desviaciones y el valor promedio**), es decir, es la raíz cuadrada de la varianza.
- Es la medida del grado de dispersión de los datos con respecto al valor promedio.

Media

$$\bar{x} = \frac{2+3+6+8+11}{5} = 6$$

Desviación típica

$$\sigma = \sqrt{\frac{2^2 + 3^2 + 6^2 + 8^2 + 11^2}{5} - 6^2} = 10.8$$

Media

$$\bar{x} = \frac{12+6+7+3+15+10+18+5}{8} = \frac{76}{8} = 9.5$$

Desviación típica

$$\sigma = \sqrt{\frac{12^2 + 6^2 + 7^2 + 3^2 + 15^2 + 10^2 + 18^2 + 5^2}{8} - 9.5^2} = 23.75$$

$$\sigma = \sqrt{\frac{15^2 + 9^2 + 15^2 + 3^2 + 12^2 + 10^2 + 18^2 + 2^2}{8} - 9.5^2} = 53.12$$

1. Varianza

2. Desviación Estándar

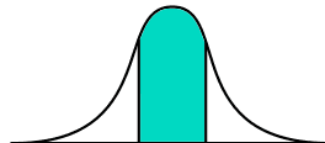
1.

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

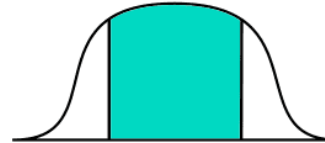
2.

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

Desviación baja



Desviación alta



Análisis Descriptivo

Calcule la varianza y la desviación típica de los siguientes datos:

4, 6, 1, 3, 10, 7, 9 y 3

PARA LA DESVIACIÓN TÍPICA

*Se saca la raíz cuadrada de la Varianza

$$\sqrt{\sigma^2} = \sqrt{8.74} = 2.95$$

PARA LA VARIANZA

*Se obtiene el promedio de los datos:

$$\bar{x} = \frac{4 + 6 + 1 + 3 + 10 + 7 + 9 + 3}{8} = 5.3$$

*Se obtiene el promedio del cuadrado de los valores absolutos de las desviaciones:

$$(4 - 5.3)^2 + (6 - 5.3)^2 + (1 - 5.3)^2 + (3 - 5.3)^2 + (10 - 5.3)^2 + (7 - 5.3)^2 + (9 - 5.3)^2 + (3 - 5.3)^2 = 69.92$$

$$\sigma^2 = \frac{69.92}{8} = 8.74$$

The background features abstract, overlapping geometric shapes in various shades of blue, ranging from light sky blue to deep navy blue. These shapes are primarily located on the right side of the frame, creating a modern, dynamic feel. The rest of the background is a solid, very light blue.

Gracias

Ing. Ricardo Velasteguí (M.B.A.)