

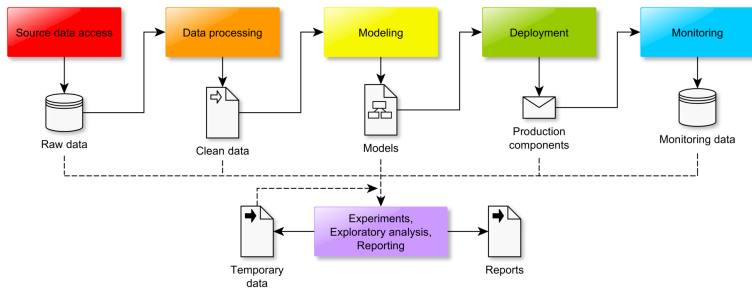
# Programación con R

Marvin J. Quispe Sedano

Junio 2021

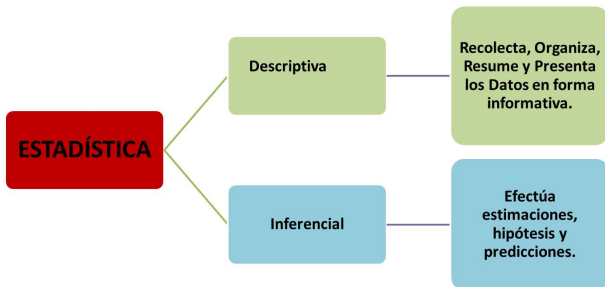
- La Estadística
- División de la estadística
- Información estadística
- Población y muestra
- Parámetro y estimador
- Clasificación de las variables
- Tablas y gráficos
- Tipos de gráficos
- Medidas estadísticas
- Datos perdidos
- Datos atípicos

- La estadística es una rama de las matemáticas aplicadas que implica la recopilación, descripción, análisis e inferencia de conclusiones a partir de datos cuantitativos.



# División de la estadística

- Las dos áreas principales de la estadística se conocen como estadística descriptiva, que describe las propiedades de los datos de muestra y población, y estadística inferencial, que utiliza esas propiedades para probar hipótesis y sacar conclusiones.



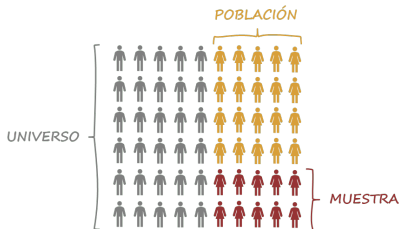
# Información estadística

- Los principales componentes de la información estadística son: **la unidad de estudio, las variables y los datos.**

Nombre	Edad	Sexo	Ingresos	
María	28	F	40.000	Unidad de datos o registro
Carlos	32	M	38.000	Observación numérica del elemento de datos <b>Ingresos</b>
Felipe	41	M	55.000	Observación no numérica, categórica, del elemento de datos <b>Sexo</b>

# Población y muestra

- La población incluye a todos los miembros de un grupo específico que se desea estudiar, con la finalidad de obtener todos los posibles resultados o mediciones que son de interés.
- La población exacta dependerá del alcance del estudio.
- Una muestra consta de algunas observaciones extraídas de la población, es decir, una parte o un subconjunto de la población.



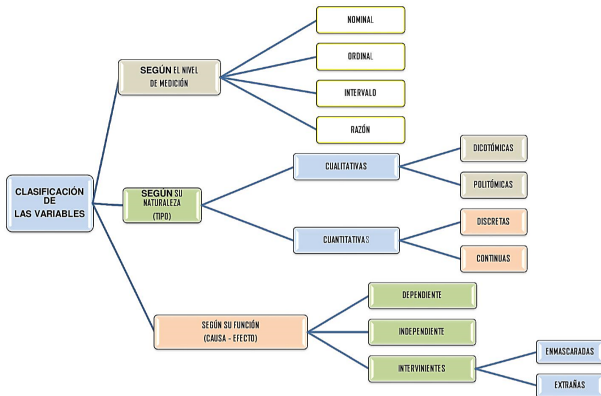
# Parámetro y estimador

- Un **parámetro** es una función definida y relacionada a una variable en específico que caracteriza a una población.
- Un **estimador** o estadístico es una función definida y relacionada a una variable en específico que caracteriza a una muestra.

MEDIDAS	POBLACION (parámetro)	MUESTRA (estadístico)
Media aritmética	$\mu$	$\bar{x}$
Varianza	$\sigma^2$	$s^2$
Desviación estándar	$\sigma$	$s$
Proporción	$\pi$	$p$
Tamaño	$N$	$n$

# Clasificación de las variables

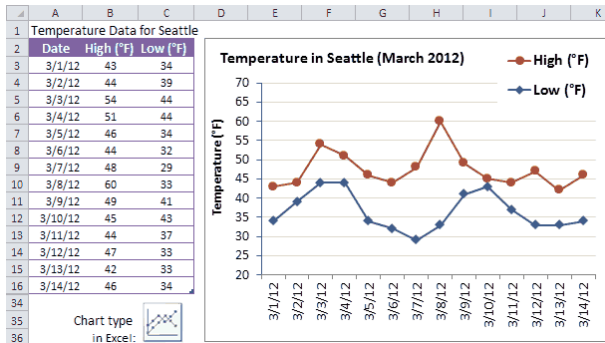
- Las variables se suelen clasificar de la siguiente manera:





# Tablas y gráficos

- Las tablas y gráficos se utilizan con frecuencia en la estadística para organizar y/o comunicar los datos de forma visual.



# Tipos de gráficos

- Las tipos de gráficos más populares son:



Pie



Bar



Column



Line



Area



Doughnut



Bubble Chart



Spider and Radar



Scatter



Comparison Chart



Stacked bar chart



Gauges

# Medidas de tendencia central

- Las medidas de tendencia central o posición son valores numéricos que pretenden resumir la distribución de los valores de una serie de observaciones, tomando como referencia principal su centro.

## MEDIDAS DE TENDENCIA CENTRAL

Las Medidas de centralización nos indican en torno a qué valor (centro) se distribuyen los datos

### MODA

Se representa por  $Mo$

Valor que tiene el mayor número de frecuencias absolutas.

Dados los números:

1,1,1,3,3,4,4,4,4,4,4,5,5,6,6,6,6,9,9

$Mo: 6$

### MEDIA

Se representa por  $\bar{X}$

Valor obtenido al sumar todos los datos y dividirlos por el número total de datos.

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{n}$$

### MEDIANA

Se representa por  $Me$

Valor que ocupa el lugar central de todos los datos, ordenados de mayor a menor.

# IMPAR  $tc = \frac{(n+1)}{2}$

# PAR  $tc1 = \left(\frac{n}{2}\right)$   $tc2 = \frac{n}{2} + 1$   $tc = \frac{tc1 + tc2}{2}$

- Las medidas de posición son valores que permiten dividir una serie de observaciones en partes porcentuales iguales.

	Definición	Fórmula
<b>Percentil</b>	Dividen la población en 100 partes	$P_1, P_2, \dots, P_{99}$
<b>Decil</b>	Se divide la población en 10 partes iguales	$D_1, D_2, \dots, D_9$
<b>Cuartil</b>	se divide la población en 4 partes	$Q_1, Q_2, Q_3$

## Medidas de dispersión

- Las medidas de dispersión son valores que nos informan sobre la distribución de una serie de observaciones.

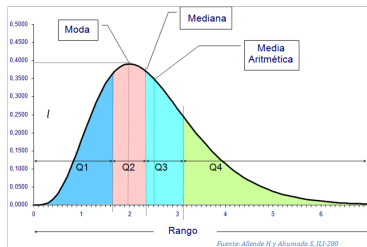
VARIANZA	DESVIACIÓN ESTÁNDAR
$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$	$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$

- $X$  → Variable sobre la que se pretenden calcular la varianza.
- $x_i$  → Observación número  $i$  de la variable  $X$ .  $i$  puede tomará valores entre 1 y  $n$ .
- $N$  → Número de observaciones.
- $\bar{x}$  → Es la media de la variable  $X$ .

RANGO ESTADÍSTICO	COEFICIENTE DE VARIACIÓN
$R = Máx_x - Mín_x$	$CV = \frac{\sigma_x}{ \bar{X} }$

- **R** → Es el rango.
- **Máx** → Es el valor máximo de la muestra o población.
- **Mín** → Es el valor mínimo de la muestra o población estadística.
- **x** → Es la variable sobre la que se pretende calcular esta medida.

- $X \rightarrow$  Variable sobre la que se pretenden calcular la varianza.
- $\sigma_x \rightarrow$  Desviación típica de la variable X.
- $|\bar{x}| \rightarrow$  Es la media de la variable X en valor absoluto con  $\bar{x} \neq 0$ .

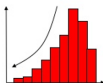


# Medidas de forma

- Las medidas de forma son valores que nos permiten identificar la forma de la distribución de una serie de observaciones.

Distribución asimétrica negativa

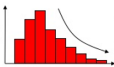
$$\bar{x} < Md < Mo$$



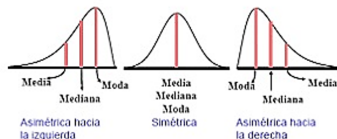
Una distribución unimodal es **asimétrica negativa** si los datos se concentran hacia los valores altos de la variable (en el lado derecho de la gráfica)

Distribución asimétrica positiva

$$\bar{x} > Md > Mo$$



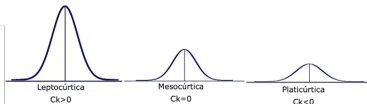
Una distribución unimodal es **asimétrica positiva** si los datos se concentran hacia los valores bajos (o pequeños) de la variable (en el lado izquierdo de la gráfica)



Cálculo: Coeficiente de Asimetría de Fisher

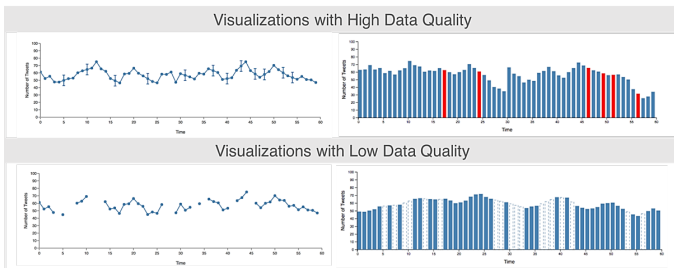
$$C_k = \frac{\sum (x_i - \bar{x})^3}{n s^3}$$

Coeficiente de curtosis y estados de apuntamiento:



# Datos perdidos: Missing

- Los datos faltantes o missing se definen como valores no disponibles, que serían útiles o significativos para el análisis, en una serie de observaciones.







# GRACIAS