# Argentina Programa

Estadísticas Descriptivas

# Estadísticas

Estadística Descriptiva

Tiene por objeto presentar y resumir los datos mediante cuadros, tablas y gráficos con la finalidad de describir las características del conjunto observado. Se obtienen conclusiones que no van más allá de ese conjunto.

Estadística Inferencial

Tiene por finalidad extender o generalizar conclusiones para un conjunto mayor que el de los datos observados.

# Tratamiento de datos

Los datos son la materia prima de la estadística. Cuando realizamos un estudio, o iniciamos una investigación, tenemos un conjunto de individuos. Seleccionamos uno o varios detalles o caracteres en esos individuos, que sean de interés para la investigación. Vemos cómo se manifiestan esos caracteres en cada uno de los individuos y recopilamos esa información. Esta información recopilada constituye los DATOS y su tratamiento, organización, resumen, interpretación es el trabajo estadístico.

# Tipos de datos

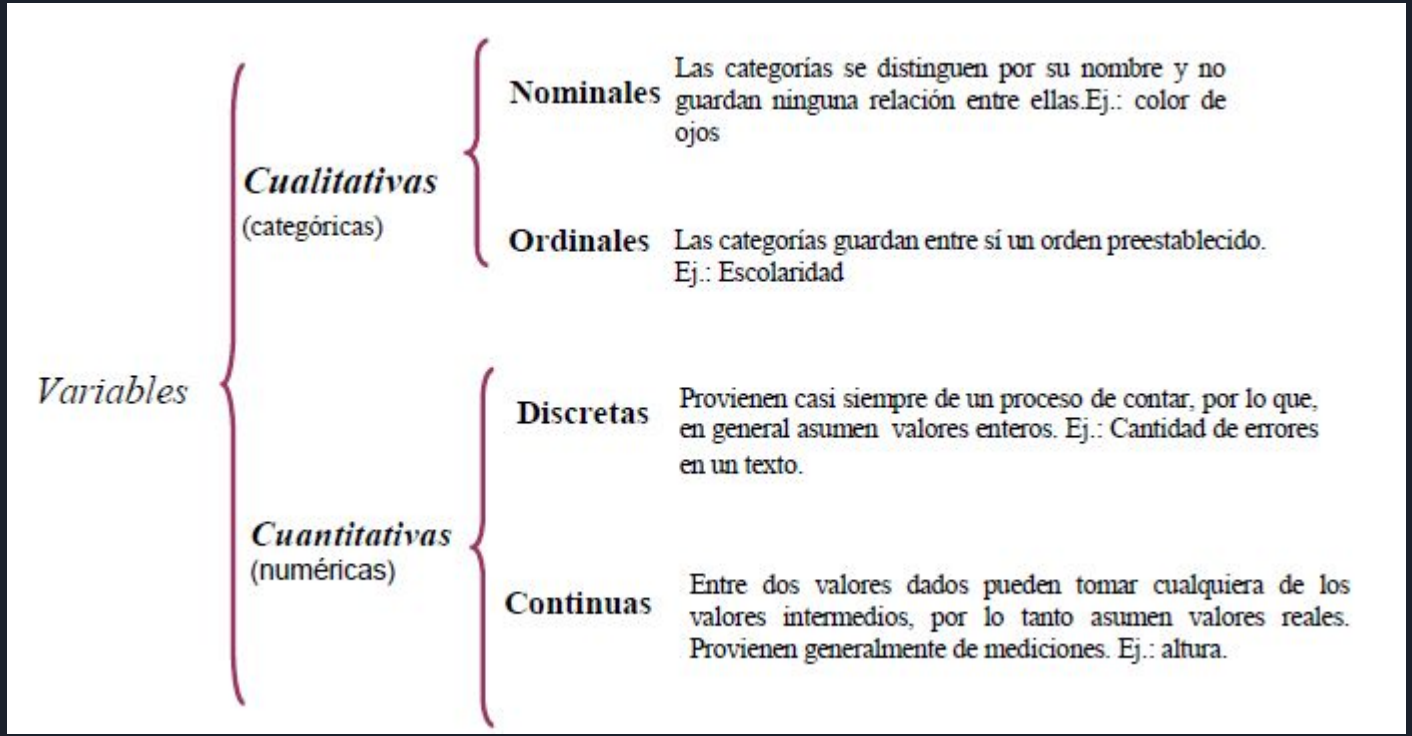
1- Datos de encuesta: la recopilación se realiza sin control de ninguno de los factores que influyen en la característica de interés.

Ejemplo: El relevamiento de datos que se hace en un censo: a cada individuo del país se le consulta sobre distintos caracteres individuales como: edad, sexo, estado civil, trabajo, ingreso, escolaridad, etc.

2- Datos experimentales: la recopilación se realiza haciendo un control sobre uno o más factores de influencia.

Ejemplo: Se quieren comparar dos métodos de enseñanza en base a los rendimientos obtenidos en los alumnos. Se eligen dos grupos de estudiantes y se les implementa un método a cada uno. Se realiza una evaluación al término de la experiencia para registrar los puntajes obtenidos.

# Tipo de variables



# ESCALAS DE MEDICIÓN

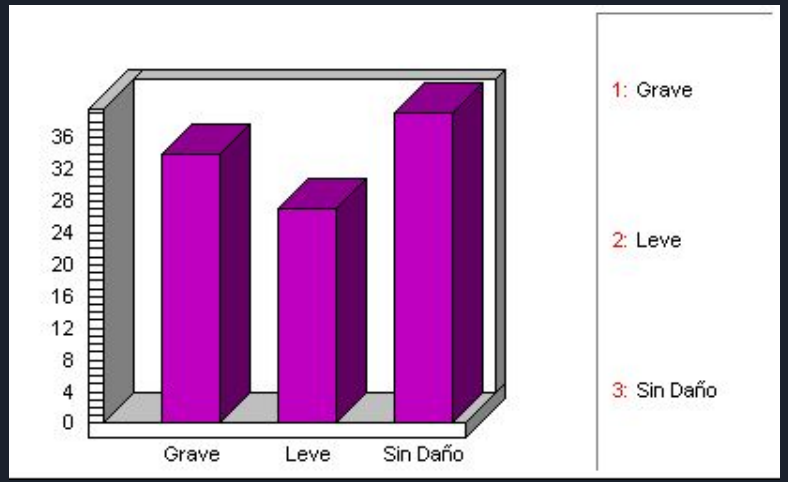
Las mediciones tienen algo así como grados de perfección, según cumplan más o menos todas las propiedades inherentes a los números. Son los niveles de medición. Estos se dividen en cuatro escalas fundamentales: nominal, ordinal de intervalos y de razón. La escala de nivel más elevado requiere normas más restrictivas, luego tiene más perfección.

# Escala nominal

Es el nivel más elemental. Divide a los objetos según sean iguales o no con respecto a una característica y se utiliza en la clasificación de atributos. Se asignan modalidades o categorías a los individuos.

# Escala ordinal

esta escala divide a los objetos en categorías iguales o no con respecto a una característica, donde las categorías están relacionadas entre sí, o sea que hay un orden que puede ser parcial o total.



# ORGANIZACIÓN DE DATOS DE VARIABLES CATEGÓRICAS

Con las variables categóricas lo más sencillo es realizar una tabla de frecuencias. Se puede representar un diagrama de tortas o de barras. En general convienen barras si la variable es ordinal, pero todo depende de la cantidad de categorías consideradas. En los ejemplos anteriores, la variables "Región" es categórica nominal y se representó un diagrama de tortas para mostrar la información. La variable "Daños neurológicos", es categórica ordinal y se representó en un diagrama de barras. Según los datos se pueden hacer barras múltiples que facilitan la comparación de distintos grupos.

# ORGANIZACIÓN DE DATOS DE VARIABLES NUMÉRICAS

**Distribuciones de frecuencia de variables discretas**

Las variables numéricas se organizan en tablas de frecuencias. La tabla resume la presencia de los datos registrados, indicando la frecuencia absoluta con que se presenta cada valor. Establecemos algunas definiciones para el conjunto de datos que se tiene:

**N** : cantidad de datos

$f_i$ : frecuencia absoluta del dato i-ésimo. Cantidad de veces que se presenta el dato en el lote.

$h_i$: frecuencia relativa del dato i-ésimo. $h_i = \frac{f_i}{N}$

$F_i$ : frecuencia absoluta acumulada hasta el dato i-ésimo. $F_i = \sum_{j=1}^{i} f_i$

$H_i$: frecuencia relativa acumulada hasta el dato i-ésimo. $H_i = \frac{F_i}{N} = \frac{\sum_{j=1}^{i} f_i}{N}$

# ORGANIZACIÓN DE DATOS DE VARIABLES NUMÉRICAS

**Distribuciones de frecuencia de variables continuas**

Si la variable es continua, no podemos ignorar que se pueden obtener valores intermedios entre dos datos dados. Por lo tanto los diagramas serán diferentes.

Cuando tenemos variables continuas, organizamos la tabla de frecuencias definiendo intervalos de clase. Esta tabla permite realizar un gráfico llamado *histograma*.

$I_i$ : intervalo iésimo - $\mathbf{I_i = [l_i ; l_{i+1})}$

$f_i$ : frecuencia absoluta del intervalo i-ésimo. Cantidad de datos en el intervalo.

$x_i'$ : punto medio del intervalo iésimo.

$h_i$: frecuencia relativa del intervalo i-ésimo. $h_i = \frac{f_i}{N}$

$F_i$ : frecuencia absoluta acumulada hasta el intervalo i-ésimo. $F_i = \sum_{j=1}^{i} f_i$

$H_i$: frecuencia relativa acumulada hasta el intervalo i-ésimo. $H_i = \frac{F_i}{N} = \frac{\sum_{j=1}^{i} f_i}{N}$

# Medidas de Posición

Cuando tenemos un conjunto de N datos, decimos que tenemos un lote y lo simbolizamos con X, la variable en estudio:


Se llama Media aritmética ( X ) de un lote de datos al promedio de los valores del lote:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

# Medidas de Posición

**a) Caso discreto:**

$$\overline{X} = \frac{X_1 f_1 + X_2 f_2 + \ldots + X_k f_k}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k} X_i f_i}{\sum_{i=1}^{k} f_i}$$

donde $k$ es el número de valores diferentes

***Ejemplo:*** Usando los datos de la Tabla 1:

| Nº de hnos. | Frecuencia |
|---|---|
| 0 | 23 |
| 1 | 37 |
| 2 | 58 |
| 3 | 57 |
| 4 | 18 |
| 5 | 10 |
| 6 | 3 |
| **Total** | 206 |

$$\overline{X} = \frac{\sum_{i=1}^{7} X_i f_i}{\sum_{i=1}^{7} f_i} =$$

$$= \frac{0\times 23 + 1\times 37 + 2\times 58 + 3\times 57 + 4\times 18 + 5\times 10 + 6\times 3}{206} =$$

$$= 2.25$$

# Medidas de Posición

**b) Caso continuo:**

Si los datos están agrupados en intervalos de clase, se ha perdido la información de los valores puntuales, por tal motivo el cálculo de la media se hace aproximado. En cada intervalo, la marca de clase, que es el punto medio del intervalo, representa cada uno de los datos. Entonces:

$$\overline{X} = \frac{f_1 x_1' + f_2 x_2' + \ldots + f_k x_k'}{f_1 + f_2 + \ldots + f_k} = \frac{\sum f_i x_i'}{\sum f_i}$$

***Ejemplo:*** En el caso de los rendimientos de 41 alumnos:

| Clases | Intervalo **I** | Frecuencia **f** | Punto medio **x´** |
|---|---|---|---|
| **1** | 10-24 | 1 | 17 |
| **2** | 24-38 | 1 | 31 |
| **3** | 38-52 | 5 | 45 |
| **4** | 52-66 | 7 | 59 |
| **5** | 66-80 | 7 | 73 |
| **6** | 80-94 | 13 | 87 |
| **7** | 94-108 | 7 | 101 |

# Medidas de posición

Mediana: Se denomina Mediana ( Me )al número real tal que a lo sumo el 50% de los datos son menores que él y a lo sumo el 50% son mayores.

Si el número de datos es impar, la mediana es el valor central. Si hubiese un número par de datos, la mediana es por convención, la media aritmética de los dos valores centrales.

Sea el lote de datos: $X_1, X_2, \ldots, X_N$. Para indicar el orden en el lote usaremos la notación:

$$X_{(1)}, X_{(2)}, \ldots, X_{(N)}$$

| N impar: | N par: |
| --- | --- |
| $Me = X_{\left(\frac{N+1}{2}\right)}$ | $Me = \frac{X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}}{2}$ |

# MEDIDAS DE DISPERSIÓN

Las medidas de dispersión complementan el análisis numérico de un lote de datos, debido a que determinan la

mayor o menor concentración de los datos. Es decir, dan una idea del alejamiento de los datos respecto a una medida de posición.

Las medidas de dispersión más comunes son el rango, el rango intercuartil, la varianza y el desvio estándar.
Rango o amplitud: El rango o amplitud es la diferencia entre el mayor y el menor valor de la variable.
Ejemplo:
Si tenemos los siguientes datos: 2, 11, 3, 7, 4, 8, 6 . El rango es 11 - 2 = 9
Rango intercuartil: Se denomina rango intercuartil es la diferencia entre el tercer cuartil y el primer cuartil.
Ejemplo:
Si tenemos los siguientes datos: 2, 11, 3, 7, 4, 8, 6. Calculamos los cuartiles: ordenamos los datos de menor a mayor: 2, 3, 4, 6, 7, 8,11. El orden de los cuartos está dado por 1.75, o sea que un dato es menos del 25% pero dos datos ya son más del 25%. Si tomamos Q1= 3, dejamos menos del 25% menores y menos del 75% mayores. En el otro extremo, análogamente Q3 = 8. Rango intercuartil = 8 3 = 5.

# MEDIDAS DE DISPERSIÓN

Varianza: La varianza se define como el promedio de los cuadrados de las desviaciones respecto de la media:

$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$. Para su cálculo suele utilizarse la siguiente fórmula equivalente: $\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2$. En el caso que los datos estén presentados en una tabla de frecuencias, la fórmula más adecuada para el cálculo de la varianza es:

$\sigma^2 = \frac{\sum_{j=1}^{m} f_j (x_j - \bar{x})^2}{N}$, donde $m$ es la cantidad de datos diferentes (si es una distribución discreta) o la cantidad de intervalos (si es una distribución en intervalos de clase, en cuyo caso se usa la marca de clase).

# Medidas de Dispersión

Desviación estándar: La desviación estándar se define como la raíz cuadrada de la varianza. Esta medida de dispersión es la más usada. En símbolos:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$$

# ANÁLISIS EXPLORATORIO DE DATOS

Describir un lote de datos X1 ,X2 , . . . , XN significa hacer referencia a la posición, dispersión, asimetría, forma, que tal lote presenta para realizar un análisis y sacar conclusiones acerca del mismo, y posibilitar, en el caso de la muestra, alguna inferencia posterior respecto de la población a la que pertenece el lote.

El Análisis Exploratorio de Datos (AED) es una herramienta ideada por Tukey, alrededor de la década del 70 y tiene la finalidad de detectar estructuras, sugerir hipótesis y facilitar un posterior Análisis Confirmatorio que se encargará de evaluar las estructuras observadas.

Tres técnicas básicas del AED son:

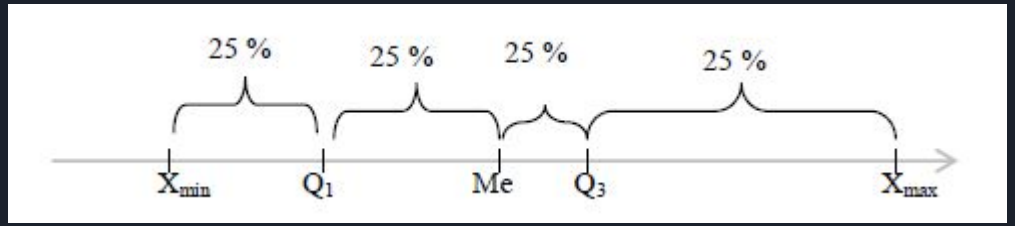
Resumen numérico

Diagrama de tallos y hojas (Stem and leaf)

Diagrama de cajas (Box-plot)

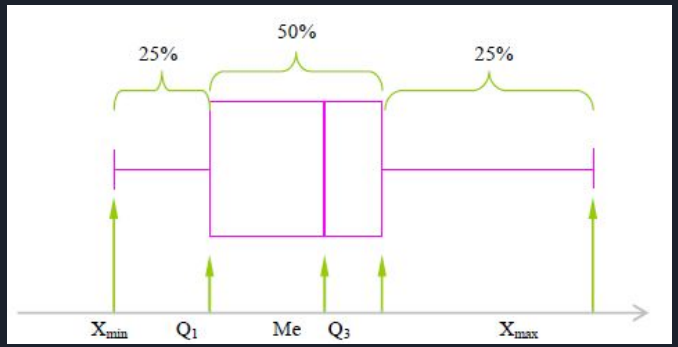
# Cuartiles

Los cuartiles son números reales que dividen la distribución de datos numéricos (ordenados de menor a mayor) en cuatro partes (los cuartos) que corresponden al 25% cada una:



# DIAGRAMA DE CAJAS

El diagrama consiste en una caja a lo largo del eje de la variable, donde se encuentra el 50% central de los datos (o sea que incluye los dos cuartos centrales), y el resto constituyen las colas de la distribución (el primer cuarto, la cola izquierda; el cuarto, la cola derecha), representadas por segmentos a los costados de la caja. La caja, por lo El diagrama consiste en una caja a lo largo del eje de la variable, donde se encuentra el 50% central de los datos (o sea que incluye los dos cuartos centrales), y el resto constituyen las colas de la distribución (el primer cuarto, la cola izquierda; el cuarto, la cola derecha), representadas por segmentos a los costados de la caja. La caja, por lo



# DIAGRAMA DE CAJAS

Si hay valores muy extremos, las colas no comienzan en los extremos sino que se destacan estos valores con una marca y la cola comienza en el dato inmediato siguiente. Se consideran los valores muy extremos con el mismo criterio tomado en el diagrama de tallos y hojas.