

Estadística descriptiva

En la jerga de DS por *variable* vamos a entender la medición de una *característica* o *atributo*. Ejemplos de variables son: edad, peso, longitud, latitud, precio, ingreso, etc.

Clasificaciones de variables

Las variables pueden ser clasificadas por su **representación numérica**, o por cómo son **medidas** (es decir, cómo son asignados números a los atributos de acuerdo a una regla, su escala de medición). Las variables también pueden ser clasificadas de acuerdo **a cómo están asociadas unas con otras**.

Esta es una clasificación de acuerdo a su **representación numérica**. Las **variables discretas** son **contables infinitas** (pueden ser asignados números naturales $N = \{1, 2, 3, \dots\}$). Las **variables continuas** son **incontables infinitas** (pueden ser asignadas a los números reales \mathbb{R}).

Representación tabular

Columnas (variables)

	x_1	x_2	x_3	x_4
p_1	0	0	0	0

x_1 : altura

x_2 : masa corporal

x_3 : ingreso mensual

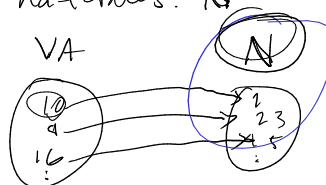
x_4 : horas hasta a comprar

VA: la medición de un atributo cuyo valor es impredecible

CONTABLE INFINITA:

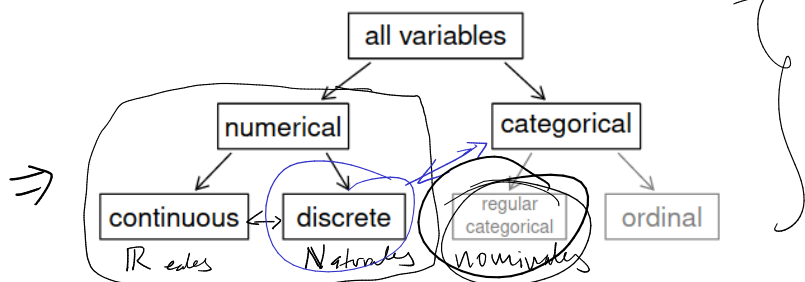
Sus elementos pueden ser puestos en correspondencia uno a uno con los enteros naturales. N

VA



INCONTABLE INFINITO:

Sus elementos no se ponen en correspondencia uno a uno con los enteros positivos.



FILA en CINÉPOLIS: Altura de personas
Formas en fila

1, 2, 3, ..., N
VA CONTABLE INF

Discrete

De acuerdo a su nivel de medición, pueden ser clasificadas como **nominales**, **ordinales**, de **intervalo** y de **razón**.

Las **variables con escala nominal** no tienen una correspondencia numérica específica. Se les asignan números solo para identificarlas. Las operaciones aritméticas de suma, resta, multiplicación, división, etc, no tienen sentido, porque el número que se les asigna no implica más o menos cantidad de su atributo. La única operación permitida es la de **conteo**, su estadística descriptiva se hace con **frecuencias y porcentajes**.

Las **variables en escala ordinal** tampoco tienen una correspondencia, siempre y cuando se sean asignados valores que preserven el orden de rango. No pueden describir grados de diferencia ni magnitud relativa entre dos observaciones con diferente orden. Por ejemplo, Alberto califica de **bueno** un libro que Luis calificó con de **malo** en una escala de {**pésimo, malo, regular, bueno, excelente**}. No podemos decir que el libro fue doblemente bueno para Alberto que para Luis. También se pueden describir con **frecuencias y porcentajes**, pero además con **mediana y rango**.

Las **variables en escala de intervalo** tienen una correspondencia precisa que debe preservar orden y magnitud. Se puede sumar y restar con valores en esta escala, pero no dividir ni multiplicar. El 0 es arbitrario (como en los grados **Celsius**) por lo que un 0 no implica ausencia del atributo, y no se puede decir que 20 °C es el doble de caliente que 10 °C. Todo intervalo en una escala de intervalo es igual que otro intervalo en la escala: 15 - 10 = 30 - 25. Sus estadísticos descriptivos son **media, desviación estándar, varianza** (y por extensión **mediana y rango**).

Las **variables en escala de razón** tienen todas las propiedades de las intervalares (conservan orden y magnitud), además de un 0 verdadero, en el que el 0 sí significa una ausencia de magnitud. Se puede sumar, restar, dividir y multiplicar en estas escalas. Pueden expresar magnitudes relativas (e.g., 4 metros es el doble de 2 metros). Sus estadísticos descriptivos, además de **media, desviación estándar, varianza, mediana y rango**, son la **media geométrica, coeficiente de variación**.

Si tienen 0 real: razón
Si no 0 real: intervalo

10°C, 20°C: tortilla es 20-10 más caliente

tortilla es 20/10 = 2 más caliente

Nominal: nombres de personas

- Sexo (biológico)

- Género

- Marcas de computadores

bueno → 1

regular → 2

...

...

Si transformamos una nominal a números, el valor es irrelevante

Ordinal

Poco Regular Mucho

1 2 3

0 1 2

-1 0 1

Orden

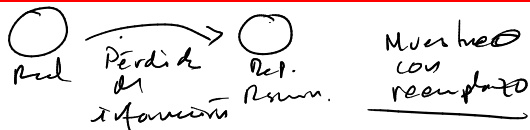
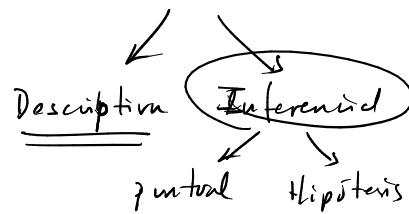
Según la representación numérica de la variable podemos usar una distribución de prob.

Exponencial

$$\Rightarrow 200/100 = 2 \text{ más alta}$$

Estadísticos descriptivos

La estadística se divide clásicamente en dos grandes ramas: la descriptiva y la inferencial (a su vez en estimación y prueba de hipótesis). La estadística descriptiva se ocupa de resumir y visualizar los datos, mientras que la inferencial se ocupa de hacer inferencias sobre la población a partir de una muestra.



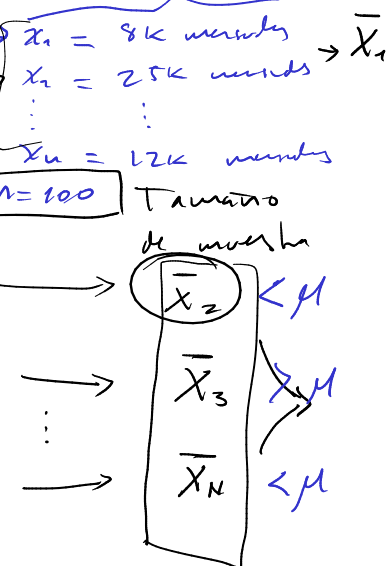
¿Por qué es necesaria la estadística?

→ La **población** es el conjunto de todos los individuos de interés en un estudio. Por individuos nos referimos a personas, animales, plantas, ítems, objetos, etc.

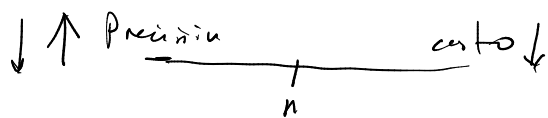
→ La **muestra** es un subconjunto de la población. La estadística descriptiva se ocupa de resumir y visualizar los datos de la muestra, mientras que la inferencial se ocupa de hacer inferencias sobre la población a partir de una muestra.

Población de μ

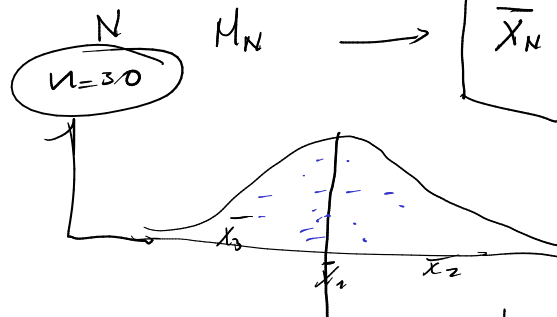
Muestra



→ Simulación



Distribución muestral de la media



Estadísticos descriptivos

De locación

Encontrar valores típicos o centrales que describan bien los datos.

los de los grandes números

la media de las medias es igual a la media poblacional

$$\bar{x}_n = \mu$$

- Media: $\bar{Y} = \sum_{i=1}^N Y_i / N$
- Mediana: $\tilde{Y} = Y_{[N+1]/2}$ si N es impar; $\tilde{Y} = (Y_{[N/2]} + Y_{[N/2+1]}) / 2$ si N es par.
- Moda: el valor más frecuente en los datos. No siempre existe, y puede haber más de una moda.

De dispersión

- Rango intercuartílico, IQR: $Q_3 - Q_1$.
- Varianza: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$
- Desviación estándar: $s = \sqrt{s^2}$,
- Desviación media absoluta: $MAD = \frac{1}{n-1} \sum_{i=1}^n |y_i - \bar{y}|$
- Rango: $\max(y) - \min(y)$

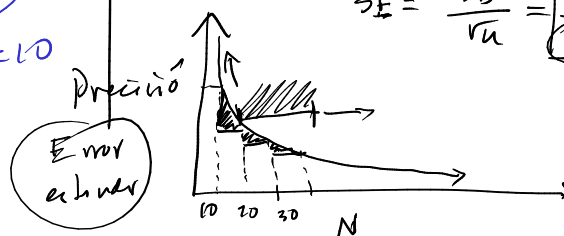
inferencia

$$\bar{x} = 11$$

$$\neq \mu = 10$$

Si tomo una muestra, ¿qué me dice sobre la población con esa muestra?

$$SE = \frac{SD}{\sqrt{n}} = \frac{1}{\sqrt{30}} * SD$$

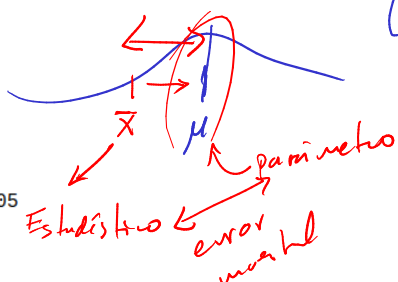


Ejercicio:

$x = (9, 10, 11, 11, 12, 14, 16, 17, 41, 61)$

Obtener: media, mediana, Q1, Q3, IQR, MAD y desviación estándar.

Media: 20.2
Mediana: 13.0
Moda: 11
Varianza: 292.17777777777775
Desviación estándar: 17.093208527885505
Desviación absoluta media: 12.32
Rango: 52



La desv. estándar de la dist. de medias se conoce como Error estándar

$$SE = \frac{desv. Est}{\sqrt{n}}$$