Definiciones

Población: conjunto que representa todas las medidas de interés del investigador

- Muestra: un subconjunto de medidas tomadas de la población de interés
- Unidad experimental: el objeto a medir
- Muestra de unidades: M UE
- Población de unidades: todas las UE
- Muestra estadística: El valor de la variable característica de M UE
- Población estadística: El valor de la variable de todas las U.E

Obs: siempre es más importante la Población que la muestra porque la población es el conjunto completo de todos los elementos que se quieren estudiar, mientras que la muestra es sólo una parte de la población

Objetivo de la estadística → El objetivo es permitir inferencia. Si tomamos bien una muestra, podemos inferir sobre la población

* El procedimiento de análisis puede ser:

1) Estadística descriptiva: Utiliza métodos numéricos

1) Estadística descriptiva: utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado. (Análisis exploratorio de datos)

2) Estadística inferencial: utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer provisiones sobre un conjunto más grande de datos.

Variables → siempre tengo que pensar en la naturaleza de la variable

X: "concentración de arsénico de un pozo de agua de HLCG, LP"

- Cualitativas
  - nominal (ej: sexo, color ojos..)
  - $\boxed{\text{ordinal}}$ (ej: educación, calificación de una evaluación) → puedo definir un orden por ej: "alto, bajo, mucho, poco"

- Cuantitativas → son ⊕ importantes en la estadística experimental
  - Discretas (ej: número de hijos) → relacionadas a un conteo ($\mathbb{N} \cup \{0\}$, $\mathbb{Z}^+$, $\mathbb{Z}^{\geq 0}$)
  - Continuas (ej: altura, peso, temperatura) → $\mathbb{R}$

Escalas de medidas

→ Cualitativo → nominal
→ ordinal

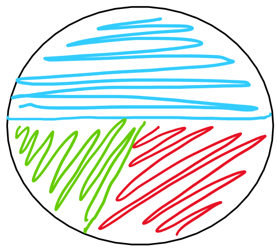
variables
- cualitativo
  - nominal
  - ordinal
- cuantitativas
  - Razón (0 absoluto) → hay ausencia de atributo
  - intervalo (0 relativo) → no hay ausencia de atributo. Ej: la T°

* Distribución de frecuencias para variables cualitativas

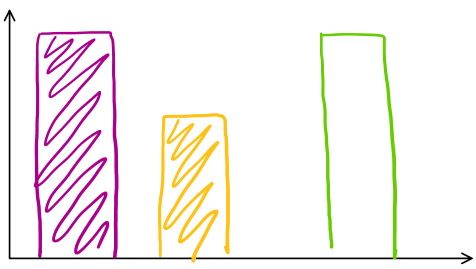
→ Las varibles de tipo cuantitativo continuas se pueden "categorizar" y transformarlas en variables de tipo cualitativo

→ se representa:

- "gráfico de sectores"



- "diagrama de barras"



← menos apropiado para variables cualitativas

* Distribución de Frecuencia para variables cuantitativas

↳ lo primero que hay que hacer es ordenar el conjunto de datos mediante el sistema de "Tallo y hojas" (no pierdo datos ⚠)

→ clases o intervalos

[100 - 200)	28	30	45
[200 - 300)	2	15	22
⋮			

*** Marca de clase:** es el punto medio entre dos valores de un intervalo.

→ Al utilizar la tabla de distribución de frecuencias se pierden datos.

Formula de Sturges

↳ una forma de determinar el número de intervalos:

- $\# intervalos = 1 + 3{,}3 . \log(n)$ — n: tamaño muestral. Por ej: los 30 pozos de estudio.

ó equivalentemente

- $\# intervalos = 1 + 3{,}3 . \ln(n)$

OBS: un cálculo nos da el número mínimo de intervalos y otro el máximo de intervalos

→ Amplitud de los intervalos (se puede redondear)

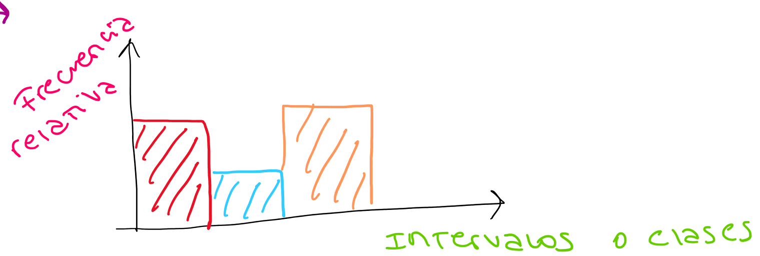
$$\Delta i = \frac{rango}{intervalos} = \frac{máx\ rango - mín\ rango}{\#\ intervalos}$$

Histograma → gráfico para variables cuantitativas continuas

→ Si los intervalos o clases tienen la misma amplitud se puede construir el histograma a partir de los pares (clases, n_i) o (clases, f_i)

→ Si los intervalos o clases tienen distinta amplitud hay que armar pares usando la densidad: $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$ (clases, densidad)

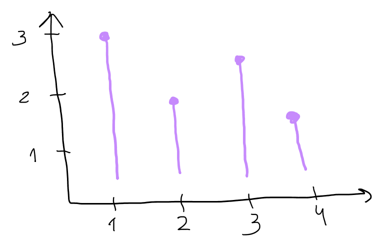
→



* Distribución de frecuencias para variables cuantitativas discretas

→ La tabla de frecuencias es similar a la tabla de frecuencias para variables cualitativas o cuantitativas continuas. La diferencia es que la variable, en este caso, toma valores enteros incluido el cero.

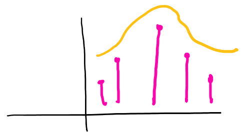
→ Se utiliza un gráfico de bastones



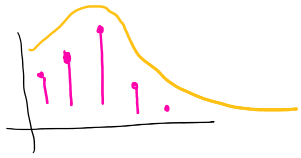
usar gráfico de barras no es apropiado. !

Si el gráfico puede ser simétrico →

El gráfico puede ser simétrico →

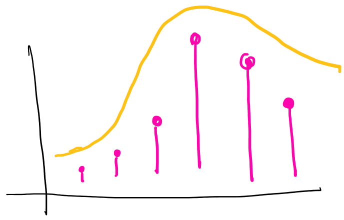


o asimétrico →



Asimétrica positiva

→



Asimétrica negativa

Medidas de posición central

↳ Están diseñadas para brindar al investigador algunos valores cuantitativos respecto de sus datos, por ejemplo saber cuál es la ubicación central en su muestra.

* Mediana → es el valor medio de un conjunto de datos

↳ Pasos: 1) ordenar los datos

2) $$\tilde{X} = \begin{cases} X\left(\frac{n+1}{2}\right) & \text{si } n \text{ es impar} \\ \dfrac{X\left(\frac{n}{2}\right) + X\left(\frac{n+1}{2}\right)}{2} & \text{si } n \text{ es par} \end{cases}$$

↳ Cuando los datos están agrupados por clases o por intervalos, la mediana es el punto medio del intervalo que acumula el 50% de los datos.

* Media (media aritmética)

Es la suma de todos los valores observados dividido por el número total de ellos

$$\overline{X} = \left(\sum_{i=1}^{k} x_i\right) \cdot \frac{1}{n}$$

Variables cuantitativas discretas

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{k} x \cdot n_x$$

Variables cuantitativas continuas

$$\overline{X}' = \frac{1}{n} \cdot \sum_{i=1}^{k} \underbrace{x_i'}_{\text{marca de clase}} \cdot n_i$$

OBS: al calcular la media utilizando la tabla (datos agrupados) obtenemos una media aproximada porque utilizamos la marca de clase que es un representante del intervalo.

* Moda

es el valor de la variable que se repite con mayor frecuencia en un conjunto de datos.

- Si todos los datos se repiten una sóla vez entonces el conjunto no tiene moda

- Si hay dos valores que se repiten igual cantidad de veces, la moda es Bimodal.

- Sólo se puede usar para variables cualitativas

- Para calcular la moda conviene ordenar los datos y ver cuál se repite más veces

- Si miramos en una tabla de distribución de frecuencias, agrupado por intervalos, se habla de clase modal, y un valor estimativo es el punto medio de esa clase modal.

Desvios → mide cuánto varían los datos del valor de la media

→ En cualquier conjunto de datos, la suma de los desvios siempre es cero, por lo tanto **no es una buena medida de dispersió**

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$$

→ • Desvio medio: $\frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$

• Desvio: $x_i - \bar{x}$

Varianza → Es una medida que expresa un desvio cuadrático medio y puede causar algunos problemas de interpretación

→ Por eso se calcula el desvio estándar:

$DE(x) = \hat{\sigma}_x = \sqrt{\frac{1}{n} \sum^{n} (x_i - \bar{x})^2}$

Por eso se calcula ...

$$D.E(x) = \hat{\sigma}_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

OBS: el desvío estándar está expresado en la misma unidad de medida que los valores del conjunto de datos.

Es una medida de dispersión o variabilidad que acompaña a la media aritmética.

$$\sigma_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$ → conjunto no agrupado

$$\tilde{\sigma}_x = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \left[(\underbrace{m_i}_{\text{marca de clase}} - \bar{x})^2 \cdot n_i\right]}$$ → conjunto agrupado

es un desvío estándar aproximado pq se utiliza la marca de clase en vez de

Coeficiente de variación → $Cv(x) = \frac{S}{x}$

S → desvío

x → media

OBS: si los datos muestran una distribución simétrica la media "$\bar{x}$" es una buena medida de posición central, y por lo tanto, el desvío estándar $\hat{\sigma}_x$ es una buena medida de dispersión.

Regla empírica → si una variable tiene un comportamiento aproximadamente simétrico, entonces verifica la regla empírica.

$(S = \tilde{\sigma})$

- $\bar{x} \pm 1 . S = (\bar{x} - S, \bar{x} + S) : \simeq 68\%$ ← deberían caer aprox. 68% de los datos

"$x \pm$ un desvío respecto ..."

- $\bar{x} \pm 1 . S = (\bar{x} - S, \bar{x} + S) \simeq 68\%$ datos

"$x \pm$ un desvío respecto de la media"

- $\bar{x} \pm 2 . S = (\bar{x} - 2S, \bar{x} + 2.S) \simeq 95\%$

- $\bar{x} \pm 3 . S = (\bar{x} - 3.S, \bar{x} + 3.S) \simeq 99\%$

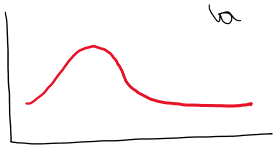
MAD → cuando los datos presentan una distribución asimétrica, la mejor medida de posición central es la mediana, por lo tanto MAd es la mejor medida de dispersión

$$MAD = Md(|x_i - Md(x)|)$$

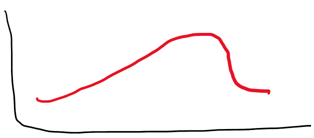
Pasos:

1) calcular los desvíos $x_i - Md(x)$
2) calcular los valores absolutos $|x_i - md(x)|$
3) ordenar los valores absolutos de los desvíos
4) Hallar la mediana de este nuevo conjunto = MAD.

Medidas de asimetría

* 

la distribución está sesgada a derecha (tiene sesgo positivo ó asimetría positiva)

* 

la distribución está sesgada a izquierda (sesgo negativo ó asimetría)

- $m_3 = \dfrac{\sum_{i=1}^{k} (x_i - \bar{x})^3}{n}$

$\boxed{m_3 = 0} \Rightarrow$ la distribución es

- Asimetría

$$a_3 = \frac{m_3}{S^3}$$

$\boxed{a_3 = 0} \Rightarrow$ la distribución es simétrica

$\boxed{a_3 > 0} \Rightarrow$ la distribución es asimétrica positiva

$\boxed{a_3 < 0} \Rightarrow$ la distribución es asimétrica negativa

medidas de curtosis

↳ • $m_4 = \dfrac{\sum_{i=1}^{k} (x_i - \bar{x})^4}{n}$

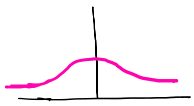
- Curtosis

$$a_4 = \frac{m_4}{S^4} - 3$$

$\boxed{a_4 = 3} \Rightarrow$ la distribución es mesocúrtica (es simétrica respecto a la media)

$\boxed{a_4 > 3} \Rightarrow$ la distribución es leptocúrtica (más puntiaguda y con colas más anchas)

$\boxed{a_4 < 3} \Rightarrow$ la distribución es platicúrtica (menos puntiaguda y con colas menos anchas)



Coeficiente de variación

↳ $CV\% = \dfrac{S}{\bar{X}} \times 100$ con $\bar{X} > 0$.

BOXPLOT → es un gráfico que sirve para visualizar el comportamiento de los datos de una muestra.

se construye con 5 medidas de posición:

* $\min(x)$ = mínimo valor del cjto. de datos
* q_1 = 1° cuartil, es la mediana de los valores que caen por debajo de la mediana. (25% de los datos)
* q_2 = Md 2° cuartil, es la mediana (50% de los datos)
* q_3 = 3° cuartil, es la mediana de los valores que caen por encima de la mediana (75% de los datos)
* $\max(x)$ = máximo valor del cjto. de datos.

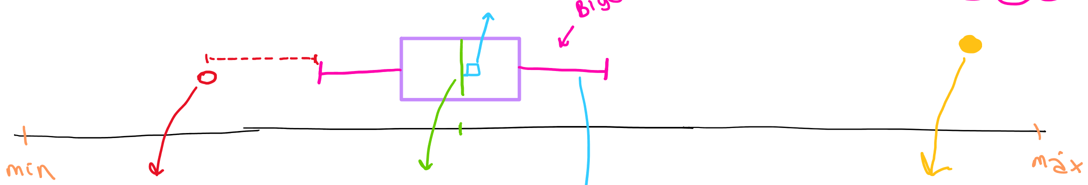
datos)

* max(x) = máximo valor del cjto. de datos.

El gráfico suele hacerse de manera horizontal, al menos de que se quiera comparar con otro gráfico y se hace en vertical.

Ej:

RIC = $q_3 - q_1$



Pto Periférico: una observación que se encuentra entre la finalización de bigote y una medida más del tamaño del bigote, se representan con una "O"

Pto extremo (outliers): es una observación que se encuentra dos RI y medios después del bigote

OBS: si los bigotes se exceden del máx o min los cortamos en esos ptos