

EST711

CLASE 1 - MÓDULO I (AED, ESTADÍSTICA INFERENCIAL)

CHRISTIAN ARAYA

INSTITUTO DE ESTADÍSTICA
PUCV

2021



MÓDULO I

Estadística

- **Estadística Descriptiva:** describir características de un conjunto de mediciones. Gráficos, Medidas de Resumen, Tablas.
- **Estadística Inferencial:** procedimientos para extraer conclusiones de una población en base a una muestra.

Metodología

- **1.- Planteamiento del problema:** idea, identificar problemática, objetivos y Marco Teórico.
- **2.- Diseño metodológico:** hipótesis, diseño de la investigación.
- **3.- Ejecución:** recolección de datos, análisis de datos y reporte de la investigación.

Estadística

- **Estadística Descriptiva:** describir características de un conjunto de mediciones. Gráficos, Medidas de Resumen, Tablas.
- **Estadística Inferencial:** procedimientos para extraer conclusiones de una población en base a una muestra.

Metodología

- **1.- Planteamiento del problema:** **idea, identificar problemática, objetivos y Marco Teórico.**
- **2.- Diseño metodológico:** **hipótesis**, diseño de la investigación.
- **3.- Ejecución:** recolección de datos, **análisis de datos y reporte de la investigación.**

Análisis Exploratorio de Datos

- **Datos:** colección de detalles para caracterizar algo.
 - **Entidad:** compañías presentes en una región; información sobre un lote de manufactura en un programa de muestreo de calidad.
 - **Evento:** caracterización sociodemográfica de los votantes en la última elección presidencial de Chile.
 - **Proceso:** registro realizado durante distintos procesos industriales.
- **AED:** práctica de mirar uno o más set de datos para descubrir estructuras subyacentes y entenderlas.
- Se miran gráficos y números para tratar de encontrar patrones. Nos orientamos a partir de información contextual, la imaginación, el seguimiento de algunos patrones identificados y la experiencia con set de datos similares.
- Es "exprimir" la Estadística Descriptiva.

Población vs Muestra

- **Población:** totalidad de personas, mediciones, entidades.
- Debe alinearse con el objetivo de la investigación y estar sujeta a tiempo y geografía.
- Población finita (estudiantes de pregrado PUCV en el año 2021) versus infinita (personas con diabetes en Chile en el año 2021): difícil de delimitar, de identificar.
- **Muestra:** subconjunto seleccionado para estudiar la población. Se usa un procedimiento aleatorio o no aleatorio.

Parámetro versus Estadístico

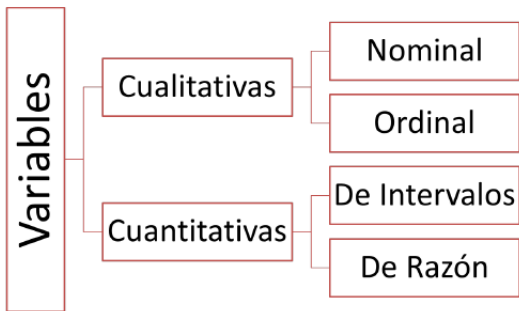
- **Parámetro:** medición numérica que describe alguna característica de la población.
 - Ejemplo: estatura promedio de los residentes de la Región Metropolitana.
- **Estadístico:** medición numérica que da cuenta de una característica de la muestra. Es una *función de la muestra*.
 - Ejemplo: encuesta CASEN (Encuesta de Caracterización Socioeconómica Nacional, Ministerio de Desarrollo Social) → ingreso promedio del hogar por decil.

Tipos de Variables (naturaleza)

- **Variables cualitativas:** atributos no numéricos.
 - Ordinales y Nominales.
- **Variables cuantitativas:** atributos numéricos.
 - Discretas y Continuas.

Escalas de Medición

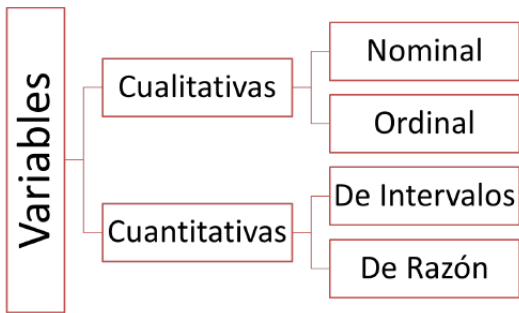
Nominal: nombres, etiquetas o categorías. Sin orden objetivo para las categorías.



Escalas de Medición

Ordinal: existe una relación ordenada entre categorías. Permite comparar, pero no se puede realizar un estudio de la magnitud de las diferencias.

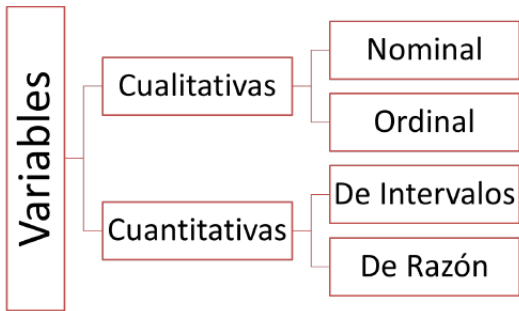
Ejemplo típico: datos recolectados en escala Likert.



Escalas de Medición

Intervalar: la diferencia entre dos valores de datos cualquiera tiene un significado. Sin embargo, los datos en este nivel no tienen un punto de partida natural desde cero. Si existe cero, es sólo una referencia, no ausencia del atributo.

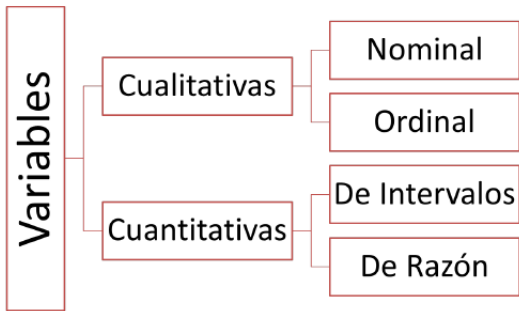
Ejemplo típico: temperatura en grados Celsius.



Escalas de Medición

De Razón: sí tiene un punto de partida o cero (donde cero es la ausencia de la característica). Para valores en este nivel, tanto las diferencias como las proporciones tienen significado.

Ejemplo típico: ingreso mensual por persona.



Gráficos

Según el tipo de variable:

Gráfico	Comentario	Tipo de Variable
Barras	Barras separadas Cantidad (frecuencia absoluta o relativa)	Nominal Ordinal Discreta*
Circular (Torta)	Relaciones entre las partes Preferir 2D	Nominal Ordinal Discreta*
Histograma	Barras pegadas Simetría en la distribución de los datos	Discreta* Continua
Polígono de Frecuencias	Segmentos lineales sobre histograma Especialmente útil para comparar grupos	Discreta* Continua
Boxplot	Medidas de posición Detección de datos atípicos 1D	Discreta* Continua

Tipos de Tablas de Frecuencias

Según el tipo de variable:

Tipo de Tabla	Tipo de Variable
No agrupada	Nominal Ordinal Discreta*
Agrupada	Discreta* Continua

Medidas de Tendencia Central

Según el tipo de variable:

Medida de Tendencia Central	Comentario	Tipo de Variable
Promedio	Sensible ante datos extremos	Variables Cuantitativas
Mediana	Centro de los datos (menor a mayor) 50 % de las observaciones Más robusta ante datos extremos	Variables Cuantitativas
Moda	Valor que más se repite (categoría de mayor frecuencia) Unimodal, bimodal, multimodal	Variables Cualitativas y Cuantitativas

Otras Medidas de Posición

Cuantiles (cuartiles, quintiles, percentiles) para variables **cuantitativas**.

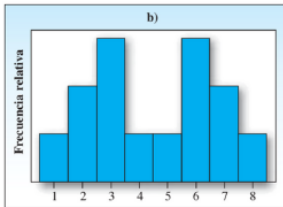
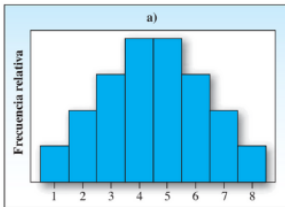
Medidas de Dispersión

Variables Cuantitativas.

Medida de Dispersión	Comentario	Tipo de Variable
Rango	Máx - Mín Fácil de calcular Sólo para set pequeños de datos	Variables Cuantitativas
Varianza	Dispersión con respecto al promedio No comparable para promedios diferentes Unidades al cuadrado	Variables Cuantitativas
Desviación Estándar	Ídem Corrige problema de la unidad	Variables Cuantitativas
Coefficiente de Variación	Corrige unidades de medidas distintas y promedios distintos Adimensional (%)	Variables Cuantitativas

Sobre el Rango

Igual rango, distintos niveles de dispersión (con respecto a una idea de centro, que responde a la intuición que tenemos de *dispersión*).



Recordatorio

A considerar:

Supuestos

Sobre distribuciones (*KS en R)

Sobre varianzas

Preferir paramétrico si se cumplen supuestos

Medias

1 o 2 poblaciones

Revisar supuestos sobre la varianza

No paramétrico

ANOVA vs KW (NP)

ANOVA:
Normalidad y
Varianzas (Levene)

Wilcoxon (NP)

Otras

Datos Categóricos:
Chi Cuadrado
Fisher Irwin

Validar que
frecuencias son
 ≥ 5

Correlación